The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1151–1170
https://doi.org/10.1214/21-AOAS1538
© Institute of Mathematical Statistics, 2022

SPATIOTEMPORAL-TEXTUAL POINT PROCESSES FOR CRIME LINKAGE DETECTION

BY SHIXIANG ZHU[a] AND YAO XIE[b]

School of Industrial and Systems Engineering, Georgia Institute of Technology, [a]*shixiang.zhu@gatech.edu,* [b]*Yao.xie@isye.gatech.edu*

Crimes emerge out of complex interactions of human behaviors and situations. Linkages between crime incidents are highly complex. Detecting crime linkage, given a set of incidents, is a highly challenging task since we only have limited information, including text descriptions, incident times, and locations. In practice, there are very few labels. We propose a new statistical modeling framework for *spatiotemporal-textual* data and demonstrate its usage on crime linkage detection. We capture linkages of crime incidents via multivariate marked spatiotemporal Hawkes processes and treat embedding vectors of the free-text as *marks* of the incident, inspired by the notion of modus operandi (M.O.) in crime analysis. Numerical results, using real data, demonstrate the good performance of our method as well as reveals interesting patterns in the crime data: the joint modeling of space, time, and text information enhances crime linkage detection, compared with the state-of-the-art, and the learned spatial dependence from data can be useful for police operations.

1. Introduction. Spatiotemporal-textual incident data are ubiquitous in modern applications, such as social media posts, electronic health records, and crime incidents. Such incidents data typically include time, location of the incidents, and marks which include categorical or more detailed descriptions of the incidents. One essential task in analyzing such data is discovering patterns from massive incident data and identifying related incidents. Here, we focus on a particular application arising from police data analysis to identify *crime linkage* from police reports.

Crime linkage detection plays a vital role in police investigations, aiming to identify a series of incidents committed by a single perpetrator or the same criminal group. The result can help police narrow down the field of search and allocate the workforce more efficiently. Crime linkage detection is usually done by finding a similar modus operandi (M.O.), typically, using physical or other credible evidence, which are observable traces of the perpetrator, such as clothes, fingerprints, DNA, ways to enter the houses, and tools as well as witness statements (Bouhana and Johnson (2016), Woodhams, Bull and Hollin (2007)). This process is not automatic, usually laborious, and requires particular domain knowledge and crime analysis.

There is an opportunity to detect crime linkage using data: a wealth of police report data containing extensive information about crime incidents exists. An illustrative example is shown in Figure 1 which shows a series of crime incidents. Such a 911 call-for-service report records information about police incidents; when a 911 call is initiated, a unique incident ID is created. A police officer is dispatched to the scene to investigate the incident and electronically enter the incident's information into the report which contains structured and unstructured data. The structured data include time, location (street and actual longitude and latitude), and crime category. The unstructured data include narratives and free-text that records interviews with the witnesses or descriptions of the scene. Thus, the police report data is a type of *spatiotemporal* incident data with *marks*.

Received November 2020; revised August 2021.
Key words and phrases. Hawkes processes with marks, crime linkage detection, variable selection.

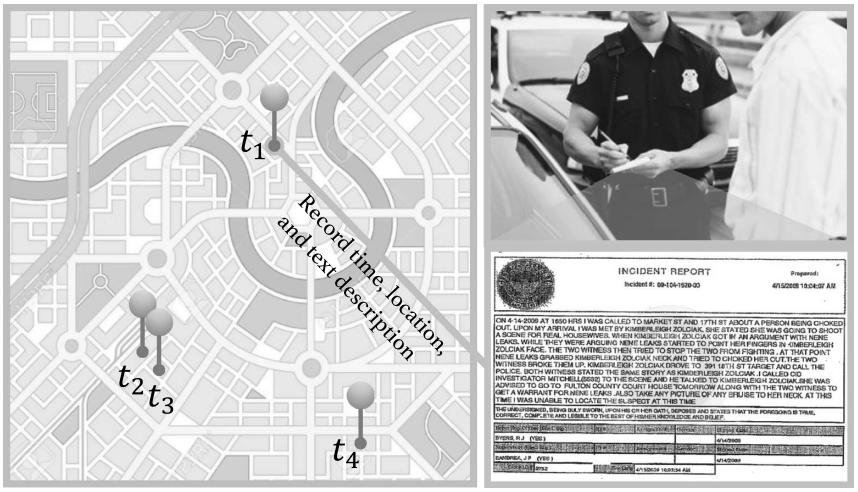


FIG. 1. *An illustrative example of a crime series consists of four crime incidents. Each crime report includes the occurrence time, location, and descriptive text that may contain important information about the incident.*

In this paper we present a framework for modeling crime incidents data, referred to as the *spatiotemporal-textual point process* (STTPP) model, based on which we can detect crime linkages without fully labeled data. Hawkes processes (Hawkes (1971)) have been used extensively to study various topics, including crime (Mohler et al. (2011)), social media (Lai et al. (2016)), and earthquake prediction (Fox, Schoenberg and Gordon (2016)). In our model each basic police patrolling geographical unit (*beat*) is regarded as a node in a network associated with a marked Hawkes process. We jointly model spatiotemporal and textual information by incorporating the text as marks of the incidents. To achieve this, we first extract the information from the free-text, using the *bag-of-words* representation, and then map the representation into embedding vectors which can be viewed as extracted *M.O.* of the incident from the free-text. The embedding is performed by the regularized Restricted Boltzmann Machine (RBM) with *keywords selection*. RBM is commonly used as a generative artificial neural network, which we adopt here to capture the joint distribution of keywords and embedding vectors. We further design a new regularization function to perform keyword selection in RBM by penalizing the total probability of the keywords being selected in the model. The keyword selection plays an important role in crime linkage detection because crime series are typically linked via a small set of keywords in the documents. Without performing keyword selection, the model can overfit the training data. Using carefully designed numerical experiments with real data, we show that our method is highly effective in detecting crime linkages, compared with other methods.

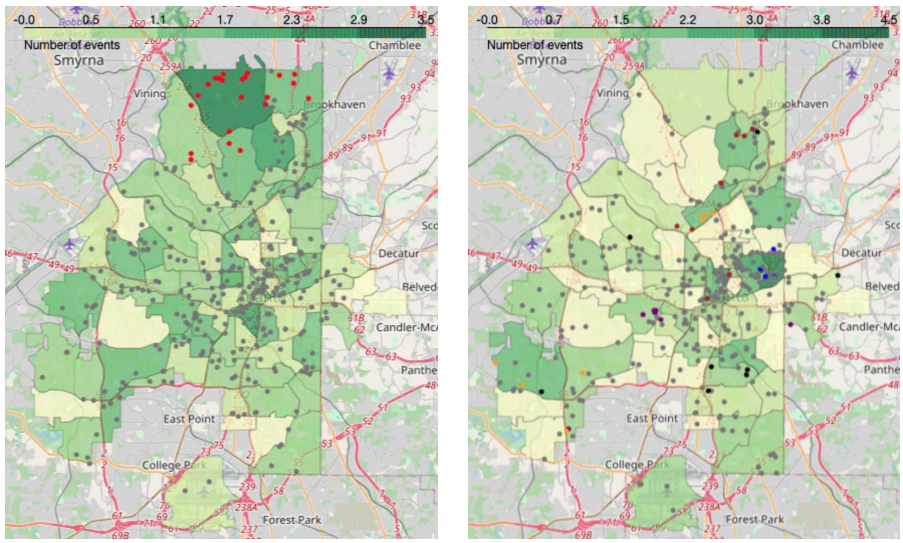
The rest of this paper is organized as follows. We start with a motivating example using real-data for crime linkage detection. Section 2 presents our proposed spatiotemporal-textual point process model as well as model estimation methods. Section 3 presents police report text analysis which will be used in the model as "marks." Finally, in Section 4 we present a study of Atlanta police data and a comparison with alternative approaches.

1.1. *Motivation with real-data example.* To motivate the connection in space, time, and police text, first, we present a motivating example. A series of residential burglaries were reported in the Buckhead, a residential neighborhood in Atlanta. From June to November 2016, 23 houses were broken into and robbed. When police arrested the perpetrator, it was found that the same person committed all these burglaries. We notice that the incidents in the same crime series tend to aggregate in time and space. This series of burglaries consisting of

Call time: *Oct 12th, 2016, 09:40:00.000* **Location**: [redacted] *Rd NW, Atlanta, GA 30327* **Description**: *I, Ofc.* [redacted] *assigned to* [redacted] *in Vehicle No.* [redacted] *was dispatched to* [redacted] *regarding a residential burglary. I spoke with the victim, Mrs.* [redacted] *and she advised me that she left her home around 0940 hrs. this morning and when she returned home around 1220 hrs. she discovered her side door to her garage kicked in and the door to the house was open. Mrs.* [redacted] *went on to say that she had left the door to the house unlocked and did not activate the alarm. Mrs.* [redacted] *advised me that her serving for 12, sterling silver flatware was taken along with the 2 drawers that they were in. Mrs.* [redacted] *advised me that the flatware is engraved with "GFG", she also stated that the knives had "Mother of Pearl handles". They also stole several (10 or more) sterling silver serving set pieces and they were also engraved with "GFG" or "D". she is not sure of the value at this time. Mrs.* [redacted] *also advised me that her bedroom was ransacked and that several pieces of jewelry are missing. The pieces were diamond and gold. Mrs.* [redacted] *advised me that she will have to take an inventory to see what was stolen and the value. Mrs.* [redacted] *stated that one of the pieces of jewelry that was stolen was a gold and diamond necklace (Infinity Necklace valued at 8, 000.00 dollars). Mrs.* [redacted] *advised me that the missing jewelry estimate is around 50, 000 dollars also advised me that after she takes inventory she will make a list and forward it to our CID. I dusted the home for fingerprints with negative results. No cameras or witnesses. I notified Investigator* [redacted] *of the burglary.*

FIG. 2. *A real police report of the residential burglary in Buckhead, Atlanta. Sensitive information has been covered. Note that a set of keywords (in red) is highly correlated with the* M.O. *of this crime.*

23 cases occur within four months, and 12 of them are within just three weeks. Figure 3(a) shows the spatial locations of the 23 burglaries which are clustered within four neighboring beats in a relatively small area. A similar phenomenon can also be observed in another robbery series, shown in Figure 3(b). Besides the locations and times of the incidents, detailed descriptions have been recorded as text (entered by the police officer who investigated the case). From these police reports a clear pattern was identified; the affected houses had their bedrooms ransacked, drawers pulled out, and valuable jewelry stolen. An example of a desensitized police report (with sensitive information masked) on a residential burglary is shown in Figure 2. Upon examining the document, we notice that the keywords (marked in



(a) Burglary (b) Robbery

FIG. 3. *Spatial distribution of*: (*a*) *burglary cases and* (*b*) *robbery cases in Atlanta from early* 2016 *to the end of* 2017. *The red dots in* (*a*) *represent an identified series of burglaries committed by the same perpetrator*; *the orange*, *blue*, *brown*, *black*, *and purple dots in* (*b*) *represent five identified series of robberies*; *the gray dots represent unidentified burglaries or robberies. The color map indicates the average number of burglaries or robberies reported in each beat. We observe that the linked incidents in the same crime series tend to occur within a few* (*two to four*) *neighboring beats in a relatively small area of the city.*

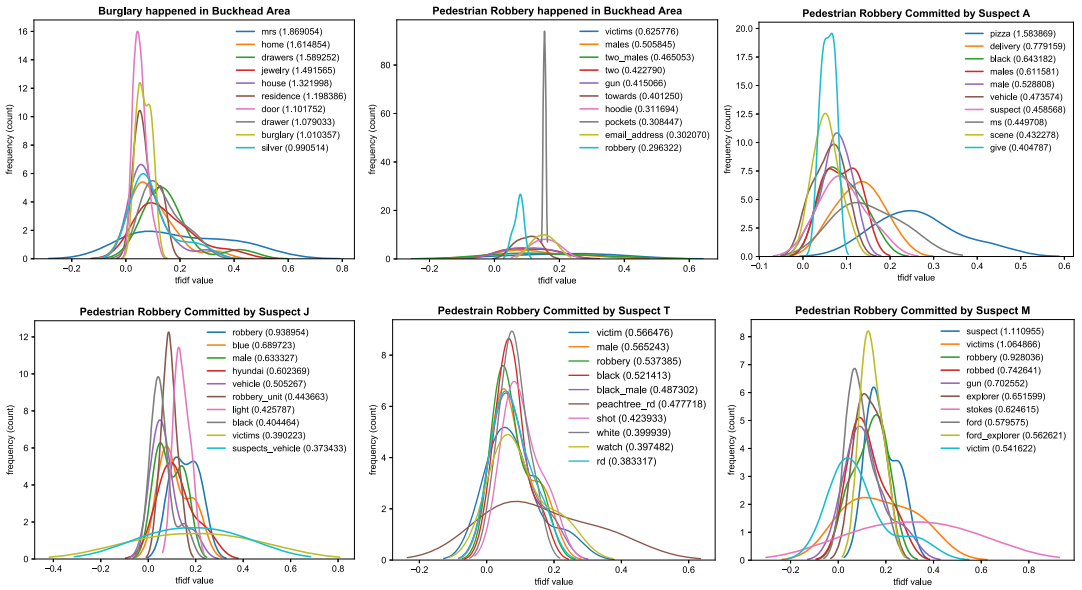


FIG. 4. *Distributions of the top* 10 *high-frequency keywords for six crime series identified by the Atlanta Police Department. Note that the cooccurring high-frequency keywords for different crime series are different, likely to be related to different* M.O. *of these crime series.*

red) such as *silver*, *bedroom*, *ransacked*, *forced entry*, *bedroom*, *jewelry*, *drawers* frequently appear in the police reports, as shown in Figure 4. This provides vital clues in detecting crime linkages from many unsolved cases which is related to the *M.O.* of this crime series. When examining other crime series, we find that a different set of high-frequency keywords occur for different series. Motivated by this, we aim to develop an algorithm that, when combining with time, location information, and the cooccurrence of keywords from police reports, can automatically capture these related incidents and help police investigators to identify M.O.

1.2. *Related work.* According to Porter (2016), there are three main types of approaches to detect crime series, which are *pairwise case linkage*, *reactive linkage*, and *crime series clustering*, respectively: (1) *Pairwise linkage detection* (Cocx and Kosters (2006), Lin and Brown (2006), Nath (2006)) aims to identify whether a pair of crimes were committed by the same offender or criminal group, where each pair is usually considered separately. Such works include Cocx and Kosters (2006), Lin and Brown (2006), which evaluate the similarity between cases according to the weights determined by experts, and Nath (2006) which learns the similarity from data by considering all incidents jointly. However, they do not consider the common *M.O.* of a crime series. (2) *Reactive linkage* (Porter (2016), Woodhams, Bull and Hollin (2007)) is similar to *pairwise case linkage*, which starts with a seed of one or more crimes, and discovers one crime at a time for a crime series. (3) *Crime series clustering* (Adderley (2004), Adderley and Musgrove (2003), Dahbur and Muscarello (2003), Ma, Chen and Huang (2010), Porter (2016), Wang et al. (2015)) discovers all clusters of crime incidents simultaneously; however, this approach requires labels which is infeasible in practice.

Recently, there has been much work on modeling discrete incident data using point processes. In particular, Hawkes processes (also known as self-exciting point processes) have been widely used for studying human dynamics (Fox, Schoenberg and Gordon (2016), Lai et al. (2016), Mohler et al. (2011)) which are characterized by the mutual "triggering" effect among incidents. Such models can capture inhomogeneous interincident times and causal (spatial and temporal) effect. Our method extends the vanilla version of multivariate Hawkes

processes by considering the unstructured text information as marks of the incidents. Unlike the stochastic declustering methods developed by Veen and Schoenberg (2008), Zhuang, Ogata and Vere-Jones (2002), Zhuang, Ogata and Vere-Jones (2004), which require an exact parametric form of triggering function, we derive an expectation maximization (EM) algorithm (Dempster, Laird and Rubin (1977), McLachlan and Krishnan (2008)) to estimate our model.

There are works studying general correlations between incidents that are not necessarily crime incidents by performing incident embeddings and evaluating their similarities in the embedding space. Such works include Zhang et al. (2017), which uses tweet token as context while capturing correlation between time, location, and keywords (tokens) in the same tweet, Du et al. (2016), Hong et al. (2017), which are based on Recurrent Neural Networks (RNN) and treat category as marks of incidents, and Quinn et al. (2011) which infers the causal relationship for neural data. Our problem's particular challenge is that our data involves complex marks which are the police report's text description.

The *spatiotemporal-textual* incident data has also been consider in Andrade, Rocha-Junior and Costa (2017), Liu, Jian and Lu (2010), Wang et al. (2012). However, these existing works do not use Hawkes process models or consider crime linkage detection. Another related work (Kuang, Brantingham and Bertozzi (2017)) uses the topic model for text, and it solves the problem of crime category classification which is very different in nature from crime linkage detection.

Existing works on regularized RBMs consider various forms of regularization on the weights of RBMs to yield sparse outputs (Halkias, Paris and Glotin (2013), Keyvanrad and Homayounpour (2017), Luo et al. (2011), Shen et al. (2019)) or sparse feature (Ranzato, Boureau and LeCun (2007), Ranzato et al. (2006)). However, they mainly focus on producing sparsity in the neural network connection which is different from the probabilistic sparsity required for the keywords selection in our scenario. Our paper extends our prior preliminary work (Zhu and Xie (2018), Zhu and Xie (2019)) which only considers text information.

2. Model. Consider a sequence of n *spatiotemporal-textual* incidents, where each observation is a tuple consisting of time, location, and text,

$$(t_1, s_1, \boldsymbol{x}_1), (t_2, s_2, \boldsymbol{x}_2), \ldots, (t_n, s_n, \boldsymbol{x}_n). \tag{1}$$

For the ith incident, $t_i \in [0, T]$ denotes time, T is the time horizon, and $t_i < t_{i+1}$; $s_i \in \mathcal{S} \subset \mathbb{R}^2$ denotes the spatial location of the ith incident that consists of the latitude and longitude of the incident; $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iq}]^\intercal \in \mathbb{R}^q$ corresponds to the fixed-length *bag-of-words* representation (Harris (1954)) with q keywords.

Here, x_{ij} is the *TF-IDF* (term-frequency-inverse-document-frequency) value (Gomaa and Fahmy (2013)) of the jth keyword, and q is the total number of the keywords that appeared in the corpus (the collection of all text documents). We will design a mapping function $\varphi : \mathbb{R}^q \to \{0, 1\}^m$ (further discussed in Section 3) to project the *bag-of-words* representation into the m-bit binary embeddings $\boldsymbol{h} \in \{0, 1\}^m$: $\boldsymbol{h} = \varphi(\boldsymbol{x})$.

2.1. *Spatiotemporal-textual process.* We model the *spatiotemporal-textual* incident data using multivariate marked Hawkes processes (Daley and Vere-Jones (2003)). A widely accepted hypothesis is that crime incidents exhibit the so-called triggering effect: once a crime incident occurs, this will increase the chance of similar incidents in the nearby neighborhood and the near future (Mohler (2014), Mohler et al. (2011), Park et al. (2021)). This effect motivates the adoption of the self-exciting Hawkes processes. Let $\mathcal{H}_t$ denote the σ-algebra generated by all historical incidents before time t. Therefore, the conditional intensity function of the Hawkes process (Rasmussen (2011)) defines the probability density that an incident

occurs at the location s, at time t, and with text $\boldsymbol{h}$, conditioning on the history $\mathcal{H}_t$ of incidents happens before time t,

$$\lambda(s,t,\boldsymbol{h}|\mathcal{H}_t) = \lim_{\Delta s,\Delta \boldsymbol{h},\Delta t\to 0} \frac{\mathbb{E}[N(B(s,\Delta s)\times B(\boldsymbol{h},\Delta \boldsymbol{h})\times [t,t+\Delta t))|\mathcal{H}_t]}{|B(s,\Delta s)||B(\boldsymbol{h},\Delta \boldsymbol{h})|\Delta t}, \tag{2}$$

where $N(C)$ is the counting measure, defined as the number of incidents that occur in the set $C \subseteq [0,T]\times \mathcal{S}\times \mathbb{R}^m$ and $|B(v,\Delta v)|$ denotes the Lebesgue measure of a ball centered at v with the radius Δv. Hawkes process is a *self-exciting* point process with the conditional intensity being influenced by the past incidents positively,

$$\lambda(s,t,\boldsymbol{h}|\mathcal{H}_t) = \mu(s) + \sum_{j:t_j<t} g(s,s_j,t,t_j,\boldsymbol{h},\boldsymbol{h}_j), \tag{3}$$

where $\mu(s)$ is the base intensity. Below, we omit $\mathcal{H}_t$ from the notation while remembering it is a conditional intensity.

We assume the triggering kernel function to be separable in space, time, and marks, as commonly assumed in the point process literature (see the review in Reinhart (2018)),

$$g(s,s_j,t,t_j,\boldsymbol{h},\boldsymbol{h}_i) = \upsilon(s,s_j)\nu(t,t_j)\kappa(\boldsymbol{h},\boldsymbol{h}_j) \geq 0,$$

where υ, ν, κ are three kernel functions for two incidents in time, location, and mark spaces, respectively. We would like to remark that the separable kernel will lead to a more computationally efficient procedure since the evaluation of the likelihood function requires integrating the intensity function over the whole space, which is high dimensional, as explained later in Section 2.2. Moreover, separable kernels in our case lead to interpretable results. To achieve this goal, we made the following choices for the kernels in our paper out of many possible forms: (1) For the temporal kernel ν we assume a commonly used exponential function $\nu(t,t_j) = \beta\exp\{-\beta(t-t_j)\}$, where $t > t_j$ and the parameter $\beta > 0$ captures the decay rate of the influence, since linked crime incidents usually aggregate in time; note that the kernel integrates to one over t. (2) Since police departments operate by *beats* (geographical units for police patrolling), we discretize the location into d disjoint units (according to beats) and replace the location s by the beat index $k \in \{1,2,\ldots,d\}$. After discretization, the spatial function $\upsilon(s,s_j)$ is represented by the coefficient α_{s,s_j}, the influence strength of beat s_j to beat s. If $\alpha_{u,v} = 0$, then beat v has no influence to beat u. Note that the spatial influence can be directional, that is, $\alpha_{u,v} \neq \alpha_{v,u}$. Define the coefficient matrix $A = \{\alpha_{u,v}\} \in \mathbb{R}^{d\times d}$, $\alpha_{u,v} \geq 0$. (3) For text we choose the inner product between text embeddings as kernel function. Since *textual similarity* is commonly measured by the normalized inner product between embeddings, we use this as our kernel function $\kappa(\boldsymbol{h},\boldsymbol{h}_j) = \boldsymbol{h}^\top\boldsymbol{h}_j/m := \tilde{\boldsymbol{h}}^\top\tilde{\boldsymbol{h}}_j$, where $\tilde{\boldsymbol{h}}$ and $\tilde{\boldsymbol{h}}_j$ denote normalized embeddings.

The rationale for choosing the form of the intensity function as above is trifold: (1) The influence of incidents is causal: the current incident only depends on past incidents, and their influence decays over time. (2) The spatial coefficient measures the correlation between two discrete locations, and the correlation may not decay over the distance since, in our context, crime incidents are not necessarily linked to what happens in the nearby neighborhood (e.g., criminals may travel). This approach allows us to capture more complex spatial influence. (3) We assume two incidents with higher textual similarity are more likely to be linked since text similarity can imply similar (*M.O.*). Moreover, we choose to use the inner product of embedding because it is the most common measure in natural language processing. Also, in our setting this leads to the closed-form likelihood function.

Following the above modeling assumptions, the conditional intensity of the kth dimension can be written as

$$\lambda_t^k(\boldsymbol{h}) = \mu_k + \sum_{j:t_j<t} \alpha_{k,s_j}\beta e^{-\beta(t-t_j)}\tilde{\boldsymbol{h}}^\top\tilde{\boldsymbol{h}}_j, \quad \forall t,k, \tag{4}$$

where μ_k is a base intensity for beat k which can be related to the ambient crime rate of the beat. Note that we use an unconventional approach to model marks by including them as a part of the kernel, which can model the triggering effect of the incidents with similar $\boldsymbol{h}$, as motivated by the fact that crime incidents with similar *M.O.* tend to trigger each other. In contrast, the conventional marked point process usually assumes the intensity function to be factorized into two terms—the conditional intensity function $\lambda_g(s,t)$ of the space and time (independent of marks) and the conditional distribution $f(x|k,t)$ of the mark given time and location: $\lambda(k,t,\boldsymbol{h}) = \lambda_g(k,t) f(\boldsymbol{h}|k,t)$.

2.2. *Model estimation and likelihood function.* In this section we explain the estimation procedure of our model. First, we choose to estimate the base intensity $\{\mu_k\}$ by the average number of incidents that occurred in that beat which can be viewed as a nonparametric estimation procedure; similar nonparametric estimation for the base intensity, using kernel estimation, has been considered in (Mohler (2014)). Our approach is computationally more efficient compared to the traditional stochastic declustering algorithm. We also validate that the performance of our approach is comparable to that of the stochastic declustering in Section 5 of the Supplementary Material (Zhu and Xie (2022)).

The spatial coefficient A are estimated based on the likelihood function. The log-likelihood function for n incidents in $[0,T]$ can be derived, based on the conditional intensity (4) (detailed derivation in Section 1 of the Supplementary Material, Zhu and Xie (2022)),

$$\begin{aligned}\ell(A) = {} & \sum_{i=1}^{n} \log\left(\mu_{s_i} + \sum_{j=1}^{i-1} \alpha_{s_i,s_j} \beta e^{-\beta(t_i - t_j)} \tilde{\boldsymbol{h}}_i^{\top} \tilde{\boldsymbol{h}}_j\right) \\ & - \sum_{k=1}^{d} \mu_k |\Omega| T - \sum_{j=1}^{n} \sum_{k=1}^{d} \sum_{\boldsymbol{h}\in\Omega} \alpha_{k,s_j} \big(1 - e^{-\beta(T-t_j)}\big) \tilde{\boldsymbol{h}}^{\top} \tilde{\boldsymbol{h}}_j.\end{aligned} \tag{5}$$

Given fixed β, the spatial coefficient can be estimated by maximum likelihood $\widehat{A} := \arg\max_A \ell(A)$; the related optimization problem can be solved efficiently by an expectation-maximization (EM) algorithm in Section 2.3. It can be shown that $\ell(A)$ is concave (Simma and Jordan (2010)), and, hence, there is a unique global maximizer.

We treat the influence parameter $\beta > 0$ as a tuning parameter that is estimated separately (discussed in Section 4.4). We take this approach, because if we treat both A and β as unknown, the corresponding maximum likelihood problem is nonconvex, and we cannot easily find a global optimal solution.

2.3. *Spatial coefficients and crime linkage estimation by EM algorithm.* In this section we discuss the model estimation procedure, based on the expectation-maximization (EM) algorithm, and also explain how to evaluate the likelihood that two crime incidents are linked by introducing a set of auxiliary variables. The EM algorithm is derived following the similar strategy as in Reinhart (2018). Introduce a set of auxiliary variables $\{p_{ij}\}$ satisfying

$$\forall i, \quad \sum_{j=1}^{i} p_{ij} = 1, \quad p_{ij} \geq 0,$$

where $\{p_{ij}\}, \forall i, j : i > j$ can be interpreted as the probability that the ith incident is triggered by the jth incident (i.e., there is a linkage between ith and jth incidents) and $\{p_{ii}\}, \forall i$ can be interpreted as the probability that the ith incident is generated due to background process. These auxiliary variables will enable us to derive a computationally efficient EM algorithm, and they can also be used for crime linkage detection: given a incident of interest i, we can

use $p_{ij} \in [0, 1]$ to measure the likelihood that j is triggered by i, as explained below. As shown in Section 2 of the Supplementary Material (Zhu and Xie (2022)), we can obtain a lower bound to the likelihood function (5), using Jensen's inequality,

$$\begin{aligned}\ell(A) \geq & \sum_{i=1}^{n}\left(p_{ii}\log(\mu_{s_i}) + \sum_{j=1}^{i-1} p_{ij}\log\bigl(\alpha_{s_i,s_j}\beta e^{-\beta(t_i-t_j)}\tilde{\boldsymbol{h}}_i^{\mathsf{T}}\tilde{\boldsymbol{h}}_j\bigr) - \sum_{j=1}^{i} p_{ij}\log p_{ij}\right) \\ & - \sum_{k=1}^{d}\mu_k|\Omega|T - \sum_{k=1}^{d}\sum_{j=1}^{n}\sum_{\boldsymbol{h}\in\Omega}\alpha_{k,s_j}\bigl(1-e^{-\beta(T-t_j)}\bigr)\tilde{\boldsymbol{h}}^{\mathsf{T}}\tilde{\boldsymbol{h}}_j.\end{aligned} \tag{6}$$

From the form of the lower bound, it can be seen that the $\{p_{ij}\}$ can be interpreted as the probability that one incident triggers another. When maximizing the lower bound, the optimizers can be expressed in closed forms, as shown in Section 3 of the Supplementary Material (Zhu and Xie (2022)). Thus, the EM algorithm involves the following iteration: in each iteration r,

$$p_{ii}^{(r)} = \frac{\mu_{s_i}}{\mu_{s_i} + \sum_{l=1}^{i-1}\alpha_{s_i,s_l}^{(r)}\beta e^{-\beta(t_i-t_l)}\tilde{\boldsymbol{h}}_i^{\mathsf{T}}\tilde{\boldsymbol{h}}_l}, \tag{7a}$$

$$p_{ij}^{(r)} = \frac{\alpha_{s_i,s_j}^{(r)}\beta e^{-\beta(t_i-t_j)}\tilde{\boldsymbol{h}}_i^{\mathsf{T}}\tilde{\boldsymbol{h}}_j}{\mu_{s_i} + \sum_{l=1}^{i-1}\alpha_{s_i,s_l}^{(r)}\beta e^{-\beta(t_i-t_l)}\tilde{\boldsymbol{h}}_i^{\mathsf{T}}\tilde{\boldsymbol{h}}_l}, \quad j < i, \tag{7b}$$

$$\alpha_{u,v}^{(r+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\mathbb{I}\{s_i = u, s_j = v\}p_{ij}}{\sum_{j=1}^{n}\mathbb{I}\{s_j = v\}(1-e^{-\beta(T-t_j)})\sum_{\boldsymbol{h}\in\Omega}\tilde{\boldsymbol{h}}^{\mathsf{T}}\tilde{\boldsymbol{h}}_j}. \tag{7c}$$

Given the precomputed text embeddings, the EM algorithm can be performed efficiently. Moreover, when implementing the algorithm, it appears that we need to sum $\boldsymbol{h} \in \Omega$; in fact, we can simplify the computation while achieving good performance. Rather than naively enumerating all possible embeddings, which will result in summing over 2^m terms, we first perform text embedding (discussed in the next section) and examine the actual support of the learned embeddings $\boldsymbol{h}$. Then, we define Ω as the union of the observed embeddings from training data. The resulted $|\Omega| \ll 2^m$. For instance, for our real-data corpus with 10,056 documents, the embedding uses $m = 1000$, and a naive enumeration will require to sum 2^{1000} terms. Using our simplification, the size of the set $|\Omega|$ is 1743. Based on the result of the EM algorithm, finding the most related incidents of the ith incident can be done by selecting incidents with the largest p_{ij}.

3. Text embedding with keyword selection. In this section we present our text embedding method with keywords selection. Recall that we treat the text embedding as marks for the spatiotemporal Hawkes process. The reason for considering text embedding $\boldsymbol{h}$ for marks instead of the raw bag-of-words representation $\boldsymbol{x}$ is two-fold: (1) The EM algorithm (equations (7a)–(7c)), if we were to use $\boldsymbol{x}$ (a 7039 dimensional real-valued vector), would require to integrate $\boldsymbol{x}$ over the entire space Ω; however, high-dimensional numerical integral is intractable, in general. Using the embedding representation, in the EM algorithm we need to sum over $\boldsymbol{h}$ (a 1000 dimensional binary vector), which is a much simpler task than integration over $\boldsymbol{x}$. (2) The police report text is unstructured and contains much noisy information; text embedding can be viewed as a certain "denoising" procedure. Moreover, the important information about the crime incidents is contained in the frequency of certain keywords that appeared in the entire corpus, as illustrated in Figure 2; embedding can capture the inherent latent connection from the raw text. Note that we perform the text embedding and model estimation for the Hawkes processes separately because jointly learning the two is mathematically intractable.

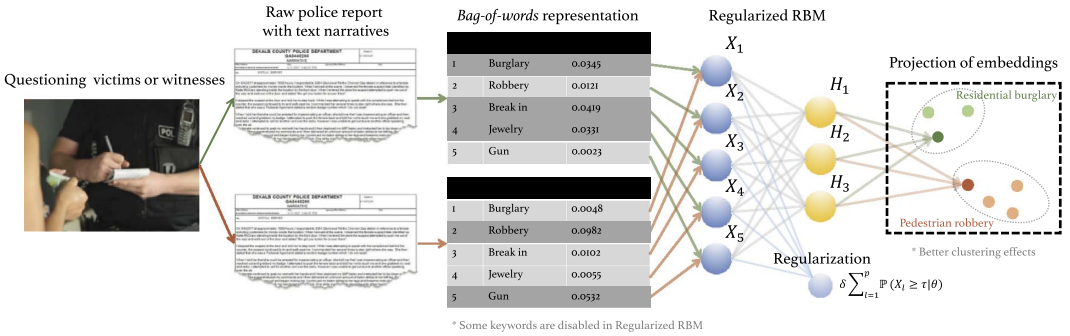


FIG. 5. *Structure of the regularized RBM for text embedding.*

Text embedding is a commonly used technique in natural language processing (Mikolov et al. (2013)), where each document is viewed as a combination of a set of keywords. The idea is to map words with similar semantic meanings to be closer to each other in the embedding space. Here, we represent each police report as a feature vector using the *bag-of-words* representation. To perform embedding, we use the Restricted Boltzmann Machine (RBM) (Fischer and Igel (2012)) which characterizes the joint distribution of keywords and unknown latent variables (the embeddings). Moreover, we introduce a new regularization function for keyword selection which is important for detecting crime linkages. Recall the example in Figure 4, where documents in different crime series tend to have a different distribution of high-frequency keywords. Thus, these cooccurrent keywords in each crime series tend to be highly related to the *M.O.* of the crime series. Since such high keywords defining *M.O.* are only a small portion of the entire vocabulary, it motivates us to perform keyword selection in embedding based on a proper regularization. Thus, we penalize the total probability that the keywords are "active" in the model.

3.1. *Restricted Boltzmann Machine* (*RBM*). An RBM is a probability graphical model, which can also be viewed as a two-layer neural network. As shown in Figure 5, the RBM consists of two types of units, the so-called visible and hidden units. In our context, the visible units correspond to keywords, and the hidden units are also called the embeddings. Assume that the visible layer has q units, denoted by a vector $\boldsymbol{X} = [X_1, X_2, \ldots, X_q]^\intercal \in \mathbb{R}^q$, and the hidden layer has m units, denoted by $\boldsymbol{H} = [H_1, H_2, \ldots, H_m]^\intercal \in \{0, 1\}^m$. Parameters of the network are $\theta = \{\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{c}\}$ which include weights $\boldsymbol{w} = \{w_{lj}\} \in \mathbb{R}^{q \times m}$, visible bias $\boldsymbol{b} = \{b_l\} \in \mathbb{R}^q$, and hidden bias $\boldsymbol{c} = \{c_j\} \in \mathbb{R}^m$. Given training data, the parameters of RBM are tuned to maximize the likelihood of the data under the model.

Consider the Gaussian–Bernoulli RBM, where the joint distribution of visible and hidden units are specified by

$$p(\boldsymbol{X}, \boldsymbol{H}|\theta) = \frac{1}{Z} \exp\{-E_\theta(\boldsymbol{X}, \boldsymbol{H})\},$$

and the partition function $Z = \sum_{\boldsymbol{X}, \boldsymbol{H}} e^{-E_\theta(\boldsymbol{X}, \boldsymbol{H})}$ is a normalization constant. The energy function is given by

$$E_\theta(\boldsymbol{X}, \boldsymbol{H}) = \sum_{l=1}^{q} \frac{(X_l - b_l)^2}{2\sigma^2} - \sum_{j=1}^{m} c_j H_j - \sum_{l=1}^{q} \sum_{j=1}^{m} \frac{X_l}{\sigma} H_j w_{lj}, \tag{8}$$

where σ^2 is the variance of the Gaussian noise for visible variables. It can be shown that the visible and hidden variables are independent upon conditioning on the hidden and the visible variables, respectively: $p(\boldsymbol{X}|\boldsymbol{H}; \theta) = \prod_{l=1}^{q} p(X_l|\boldsymbol{H}; \theta)$, and $p(\boldsymbol{H}|\boldsymbol{X}; \theta) =$

$\prod_{j=1}^{m} p(H_j|\boldsymbol{X};\theta)$. This property simplifies our derivation. Let $\mathcal{N}(x;\mu,\sigma^2)$ denote the probability density function of normal random variable with mean μ and variance σ^2. The conditional probabilities for the ith keyword and the jth entry in the embedding are given by

$$p(X_l|\boldsymbol{H};\theta) = \mathcal{N}\Big(X_l; b_l + \sigma \sum\nolimits_{j=1}^{m} w_{lj} H_j, \sigma^2\Big), \tag{9a}$$

$$p(H_j = 1|\boldsymbol{X};\theta) = \operatorname{sigm}\Big(c_j + \sum\nolimits_{l=1}^{q} \frac{X_l}{\sigma} w_{ij}\Big), \tag{9b}$$

where the sigmoid function is defined as $\operatorname{sigm}(x) = 1/(1+e^{-x})$. The marginal distribution of keywords is given by

$$p(\boldsymbol{X}|\theta) = \sum\nolimits_{\boldsymbol{H}} p(\boldsymbol{X},\boldsymbol{H}|\theta) = \frac{1}{Z} \sum\nolimits_{\boldsymbol{H}} \exp\{-E_\theta(\boldsymbol{X},\boldsymbol{H})\},$$

which is also known as the Gibbs distribution.

3.2. *Regularized RBM.* Consider independent and identically distributed (i.i.d.) training data $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$. The RBM's log-likelihood function is given by $\mathcal{L}(\theta) = \sum_{i=1}^{n} \log p(\boldsymbol{x}_i|\theta)$. We now present a regularized maximum likelihood estimate of the RBM model,

$$\min\nolimits_\theta \Big\{-\mathcal{L}(\theta) + \delta \sum\nolimits_{l=1}^{q} \mathbb{P}(X_l \geq \tau|\theta)\Big\}, \tag{10}$$

where $\delta > 0$ is the regularization parameter and the threshold τ a hyperparameter that controls the threshold for keywords selection. We assume that keywords with TF-IDF value less than 10^{-2} can be ignored and set $\tau = 10^{-2}$ in experiments. The regularization function penalizes the total probability of visible units (keywords) being "selected" by the model, and, thus, it encourages "stochastic sparse" activation patterns and only selects a subset of keywords. Here, we do not directly use the standard ℓ_1-norm type of regularizer on the weights because we want the output to be sparse in a probabilistic sense.

Computing the likelihood of a Markov random field or its gradient is, usually, computationally intensive. Thus, we employ sampling-based methods to approximate the likelihood function and its gradient (Hinton (2012)) and perform stochastic gradient descent (Lan (2020)). In each iteration we optimize one variable while fixing other variables, and gradients are evaluated from samples.

A benefit of our regularization function is that its gradient can be derived in closed form. Below, let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and cumulative distribution function of the standard normal random variable, respectively. Let $\langle\cdot\rangle_P$ denote the expectation with respect to a distribution P. We can write the regularization term as

$$\mathbb{P}(X_l \geq \tau|\theta) = \big\langle \mathbb{P}(X_l \geq \tau|\boldsymbol{H};\theta)\big\rangle_{p(\boldsymbol{H})} = 1 - \left\langle \Phi\Big(\frac{\tau - b_l - \sigma \sum_{j=1}^{m} w_{lj} H_j}{\sigma}\Big)\right\rangle_{p(\boldsymbol{H})}. \tag{11}$$

Let

$$\tau_l' = \tau - b_l - \sigma \sum_{j=1}^{m} H_j w_{lj}, \quad l = 1,\dots,n.$$

The detailed derivation of gradients is shown in Section 4 of the Supplementary Material (Zhu and Xie (2022)). This leads to a simple procedure for performing stochastic gradient descent in the parameters of the RBM model,

$$\Delta w_{lj} = \langle X_l H_j\rangle_{p(\boldsymbol{H}|\boldsymbol{X})\tilde{p}(\boldsymbol{X})} - \langle X_l H_j\rangle_{p(\boldsymbol{X},\boldsymbol{H})} - \delta\Big\langle \frac{H_j \phi(\tau_l')}{1-\Phi(\tau_l')}\Big\rangle_{p(\boldsymbol{H}|\boldsymbol{X})\tilde{p}(\boldsymbol{X})},$$

$$\Delta b_l = X_l - \langle X_l \rangle_{p(\boldsymbol{X})} - \frac{\delta}{2\sigma^2} \left\langle \frac{\phi(\tau_l')}{1 - \Phi(\tau_l')} \right\rangle_{p(\boldsymbol{H}|\boldsymbol{X})\tilde{p}(\boldsymbol{X})},$$

$$\Delta c_j = p(H_j = 1|\boldsymbol{X}) - \big\langle p(H_j = 1|\boldsymbol{X}) \big\rangle_{p(\boldsymbol{X})},$$

where $\tilde{p}(\boldsymbol{X})$ denotes the empirical distribution. To evaluate the gradient, we adopt the k-step contrastive divergence (CD-k) algorithm (Hinton (2002)).

4. Real-data study. Using our methods, we now study a large-scale police dataset and demonstrate their competitive performance for linkage detection. All the code in this article can be found in the Supplementary Material (Zhu and Xie (2022)).

4.1. *Dataset*. We study a data set of 10,056 crime incidents recorded by the Atlanta Police Department from early 2016 to the end of 2017. Each incident is associated with a 911 call, with information including crime category, time and location of the incident, and comprehensive text descriptions entered by the police officer. Two crime incidents are said to be linked if they are in the same crime series. We only have a handful of identified crime series by police, consisting of six crime series and a total of 56 incidents in these series. Thus, this also shows the importance of an unsupervised learning approach. Here, the labels for the crime series are not used for fitting the model but are only used for validation. The data are available in the Supplementary Material (Zhu and Xie (2022)).

We preprocess the raw data as follows: (1) *Discretize the continuous geolocation of the crime incidents according to beats*. We associate each crime incident using the policing beat index. Atlanta is divided into 80 disjoint beats. (2) *Normalize the call time*. We regard the call time (when the dispatch center receives the 911 call) as each crime incident's initial time. For ease of calculation, we then normalize this time to the range of [0, 1], that is, the time horizon corresponds to $T = 1$. (3) *Initialize base intensities*. We estimate the base intensity by estimating the average number of incidents in each beat and within the time horizon. (4) *Construct bag-of-words representations for text documents*. We normalize the text to lowercases so that, for example, the distinction between "The" and "the" are ignored; we also remove stop-words, independent punctuation, low term-frequency (TF) terms, and the terms that appeared in most documents (high document frequency terms). Then, we compute the *bag-of-words* vector for each police report using 7039 keywords. For the document's feature vector each entry corresponds to the *TF-IDF* value of a keyword or bigram (a sequence of two adjacent words from a string of text) that appears in the corpus. (5) We study two particular categories of crime, burglary and robbery, as the same category cases may define similar *M.O.* There are 349 *burglary* crimes (23 of them are labeled) and 333 *robbery* crimes (23 of them are labeled), respectively. We also compare with groups with 305 *mixed* types of crimes (56 of them are labeled).

4.2. *Text embedding results*. First, we examine the text embedding and keywords selection results. Recall that regularized RBM aims to select the most relevant keywords for clustering in the embedding space. The number of keywords selected in the model depends on the regularization parameter value which is typically chosen by cross-validation. Since we have very few labeled crime series data, we will take the following approach for cross-validation. Consider two types of police reports, including "burglary, robbery", and group the rest of the category as "mixed" type. Based on this, we examine the embedding space to map the police reports to the same category as much as possible (use this approach as weak supervision). Specifically, we use the average *silhouette score* (Rousseeuw (1987)) as a metric which measures the average distance of a point to its cluster (cohesion) relative to other clusters (separation). The silhouette ranges from -1 to 1 (the higher, the better), and

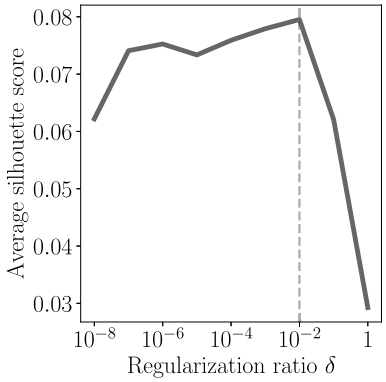


FIG. 6. *Optimal choice of regularization parameter δ for regularized RBM by five-fold cross-validation. The vertical axis corresponds to average silhouette scores for clustering "burglary, robbery", and "mixed" types of incidents in the embedding space by regularized RBM. The model attains its best performance at $\delta = 10^{-2}$, which we use in the subsequent experiments.*

we select the value of the regularizing parameter δ to maximize the average silhouette score. As shown in Figure 6, based on *five*-fold cross-validation, the optimal choice of $\delta = 10^{-2}$.

The selected keywords (with *TF-IDF* value $> 10^{-2}$) can be observed in the *reconstructed corpus*, which is generated by performing Gibbs sampling according to the conditional probability (9a), given the text embeddings. Note that the selected subset of keywords in the reconstructed corpus is small, 280 out of 7038 keywords, as desired. The selected keywords are supposed to be the most important for embedding. Thus, we examine the selected keywords in Figure 7, together with high *TF-IDF* keywords. These keywords correspond to common descriptors in crime reports. The combination of keywords with high TF-IDF values may portray a certain aspect of the incident and hence help to identify crime linkages. For example, a combination of high TF-IDF keywords, including *home*, *door*, *window*, *stolen*, may be associated with a burglary incident, whereas a combination of *midnight*, *Toyota Corolla*, *pistol* may be associated with an armed robbery incident. Some keywords are strong indicators of crime type, for example, *marijuana*, *vandalism*, whereas some keywords may reveal the stage of the investigation of the related crime series, for example, *wasn't sure*, *suspect*, *identified*, *arrestee*, *police custody*, *Miranda*. In Figure 8 we also show that the embeddings generated by the regularized RBM can better cluster incidents for the identified crime series. This confirms that our proposed regularized RBM is more effective in clustering police reports than the vanilla RBM by removing irrelevant keywords.

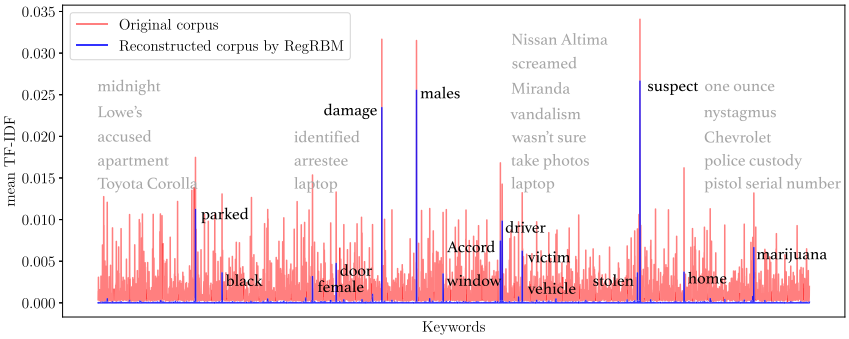
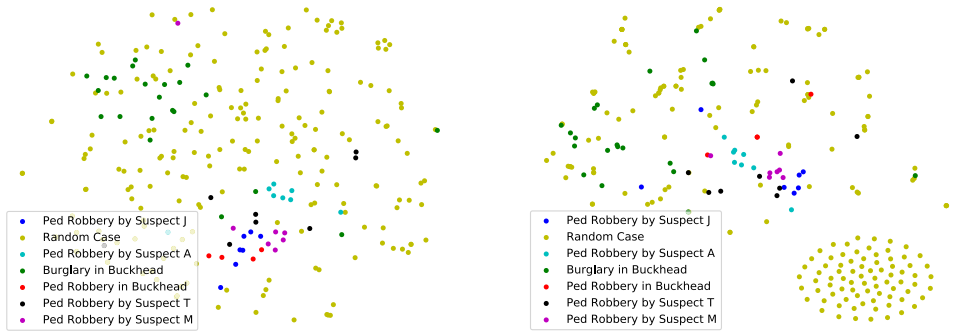


FIG. 7. *The average TF-IDF value of keywords in the original corpus (shown in red) and the selected keywords (shown in blue). Crime analysts in the Atlanta Police Department validate that the selected keywords by our model, such as stolen, home, marijuana, play a significant role in defining the* M.O. *and linking crime incidents.*



(a) vanilla RBM ($\delta = 0$) (b) regularized RBM with $\delta = 10^{-2}$

FIG. 8. *Comparison of embeddings generated by vanilla RBM and our proposed regularized RBM. Each dot represents a crime incident's embedding, projected on a two-dimensional space via t-SNE (Van der Maaten and Hinton (2008)); we show* 56 *labeled and* 500 *random crime incidents in both figures.*

4.3. *Evaluation metrics and procedure.* We adopt standard performance metrics, including precision, recall, and F_1 score. This choice is because linkage detection can be viewed as a binary classification problem, where we aim to identify if there is a linkage between two arbitrary crime incidents in the data. Define the set of all truly related incident pairs as U, the set of positive incident pairs retrieved by our method as V. Then, precision P and recall R are defined as

$$P = |U \cap V|/|V|, \qquad R = |U \cap V|/|U|,$$

where $|\cdot|$ is the number of elements in the set. The F_1 score combines the *precision* and *recall*: $F_1 = 2PR/(P + R)$ and the higher F_1 score the better. Since numbers of positive and negative pairs in real data are highly unbalanced, we do not use the ROC curve (true positive rate vs. false-positive rate) in our setting.

The evaluation procedure is as follows. The six identified crime series include 56 incidents with more than 1000 detected crime linkages between them which can be used as "true labels" for training and validation purposes. We first shuffle the labeled data set randomly and split the crime series into k groups ($k = 5$), where each group has, at least, one identified crime series. For each unique group we take the group as a hold-out or test data set and take the remaining groups as a training data set. Then, we fit a model, using all of the unlabeled data, to find the optimal A, and optimal β is chosen for each location by a grid search in $[10^{-8}, 10^{0}]$, using the labeled training data. Lastly, we evaluate the fitted model with optimal A, μ, β on all the labeled test sets and obtain the average F_1 score. Given all possible incident pairs in a group of crime incidents, we retrieve the top N pairs with the highest linkage likelihood p_{ij}, $\forall i, j$ values returned our algorithm. If two crime incidents of a retrieved pair were indeed in the same crime series, then it is a success. Otherwise, the pair is unlinked, and it is a misdetection. In our data, burglary has 55,278 pairs in total, and 97 are linked; robbery has 60,726 pairs in total, and 231 are linked; mixed has 46,360 pairs in total, 328 of them are linked.

4.4. *Choice of temporal coefficient β by cross-validation.* We show that an optimal choice of the parameter β for crime linkage detection exists, which achieves the best bias-variance tradeoff. Specifically, we perform the k-fold ($k = 5$) cross-validation with a grid search to find the optimal β using the labeled data. The tradeoff can be intuitively explained, as in (7b); when β are too small, the long-range temporal dependence is not captured, and

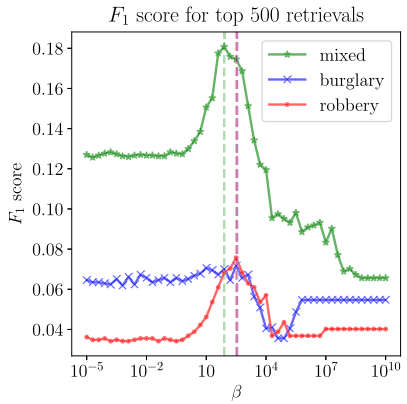


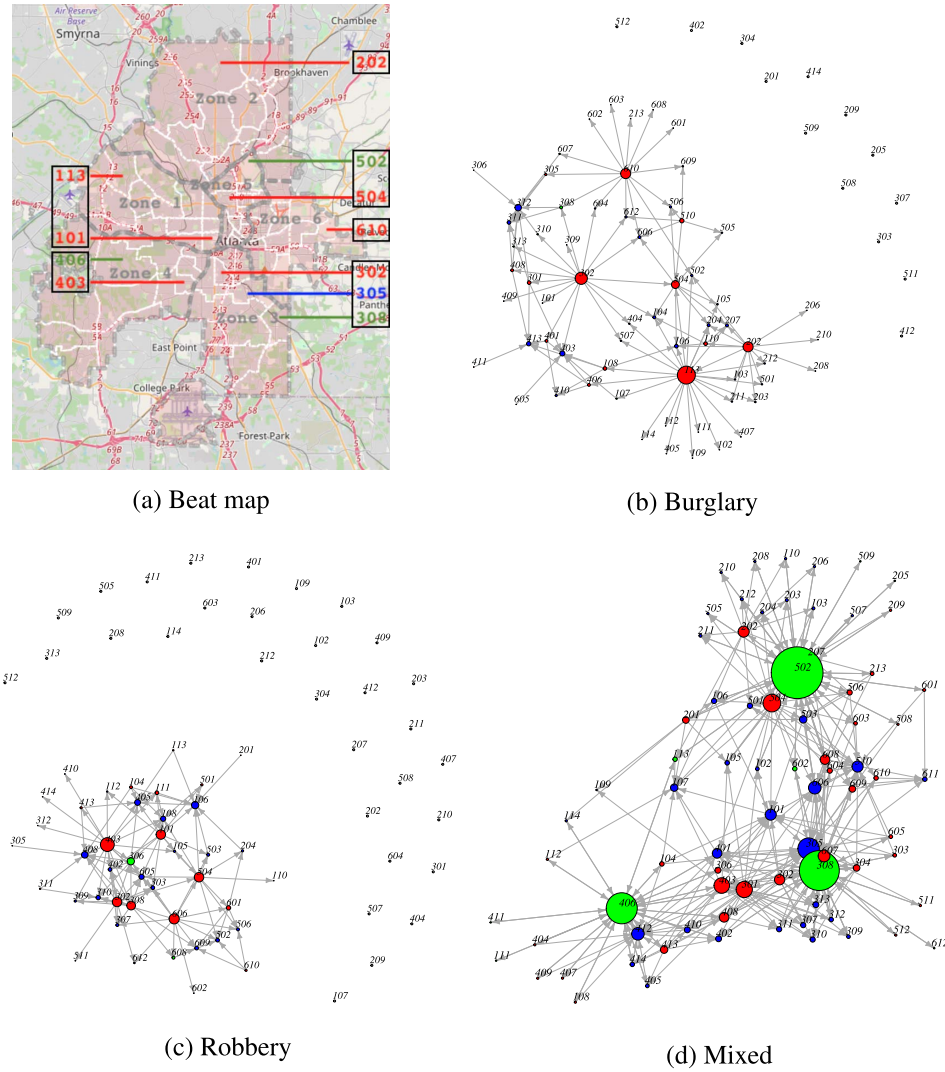
FIG. 9. *Choice of the optimal β by five-fold cross-validation*: *F_1 scores over the temporal correlation parameter β when retrieving the top $N = 500$ incident pairs with the highest correlated probabilities p_{ij}; results are obtained by repeating over* 50 *random experiments*.

when β are too large, the process may not forget the history. Moreover, in our model the influence kernel jointly captures spatiotemporal and text influence. When β is too large, the temporal influence may dominate the contribution of textual correlation. An appropriately set temporal coefficient β can improve the performance of our method. A real-data example is shown in Figure 9, where the vertical dash lines in the figures indicate where the model attains its best performance regarding the F_1 score (defined in Section 4.6). We test `STTPP+RegRBM` using $N = 500$ pairs of arbitrarily retrieved results (including both linked and unlinked cases). We note that, in the experiments, $\beta \approx 10^2$ leads to the best performance. In practice, when there is a handful of labeled data indicating crime series (like the dataset we have here), we can use this small amount of training data to preselect an optimal β used in our model.

4.5. *Estimated spatial coefficients $\alpha_{u,v}$ and interpretations*. The EM algorithm can estimate the spatial coefficient matrix A, capturing the directional influence between locations. The magnitude of the coefficient captures how large the influence is. In our experiments we randomly initialize the coefficients in $(0, 1)$.

Now, we visualize the estimated A in Figure 10, treating it as the adjacency matrix of a weighted and directed graph, where nodes represent beats and edges represent the spatial influences. We threshold the estimated spatial coefficients and only keep edges if $a_{ij} > 0.5$. For burglary and robbery (Figure 10(b), (c)), some beats are isolated and have no connection with any other beats. A few beats, indicated by large red dots in the graph, have a dominating influence on their surrounding beats. An interesting observation is that the linked incidents in the same crime series usually occurred within a few (two to four) neighboring beats, indicating that the perpetrators usually confine their criminal activities in a relatively small area of the city with which they might be familiar. The situation becomes more complicated when we consider all types of cases (more than 160 categories of crimes) altogether, as shown in Figure 10(d). Also, some of the beats, such as 113, 202, 302, and 610, are the crime hotspots which are very likely to trigger subsequent crime incidents. Increasing patrols in these regions may help to curb future crime and enhance the safety of the city.

4.6. *Comparison with alternative methods*. We compare our method, referred to as the `STTPP+RegRBM`, to the vanilla RBM without regularization, referred to as the (`STTPP+RBM`), as well as other alternative methods. We may consider crime linkage detection as an information retrieval task: given a new incident, we would like to find a number of the most relevant incidents from historical data. We compare with related alternative



(a) Beat map

(b) Burglary

(c) Robbery

(d) Mixed

FIG. 10. (*a*) *Police beat map of Atlanta*: *regions with white boundary are beats*; *the ID of beats with significant influence learned by our model has been highlighted*. (*b-d*) *Directed graphs corresponding to the estimated* $\widehat{A}$ *for three scenarios*, *where each node represents a beat and each edge corresponds to* $\alpha_{u,v} > 0.5$. *The red nodes are the beats with larger outdegree than indegree*, *indicating that the incidents in these beats are more likely to trigger subsequent incidents in the connected beats*; *the blue nodes are the opposite*; *the green nodes indicate the beats have equal indegree and outdegree*, *and the grey nodes are isolated. The dots' size indicate the indegree and outdegree differences*. *We notice that certain beats are more influential than others*.

methods, including latent semantic analysis (performed by singular value decomposition) and latent Dirichlet allocation (LDA) (commonly used in natural language processing). We also compare with Autoencoder (AE), a neural network-based embedding technique. However, these alternative methods learn embeddings for documents in feature space without considering spatiotemporal information. Hence, we extend AE and SVD by including the spatiotemporal information as part of their inputs which are concatenations of the bag-of-words feature vector and incident time and location (latitude and longitude); each dimension of the input is normalized to [0, 1]. We refer to these two methods as AE+ST and SVD+ST, respectively. As a sanity check, we also consider the random-pick strategy as one of the baselines which randomly selects the subset of variables to be incorporated in the model. To compare with embedding methods, we compute their inner products in the embedding space for each

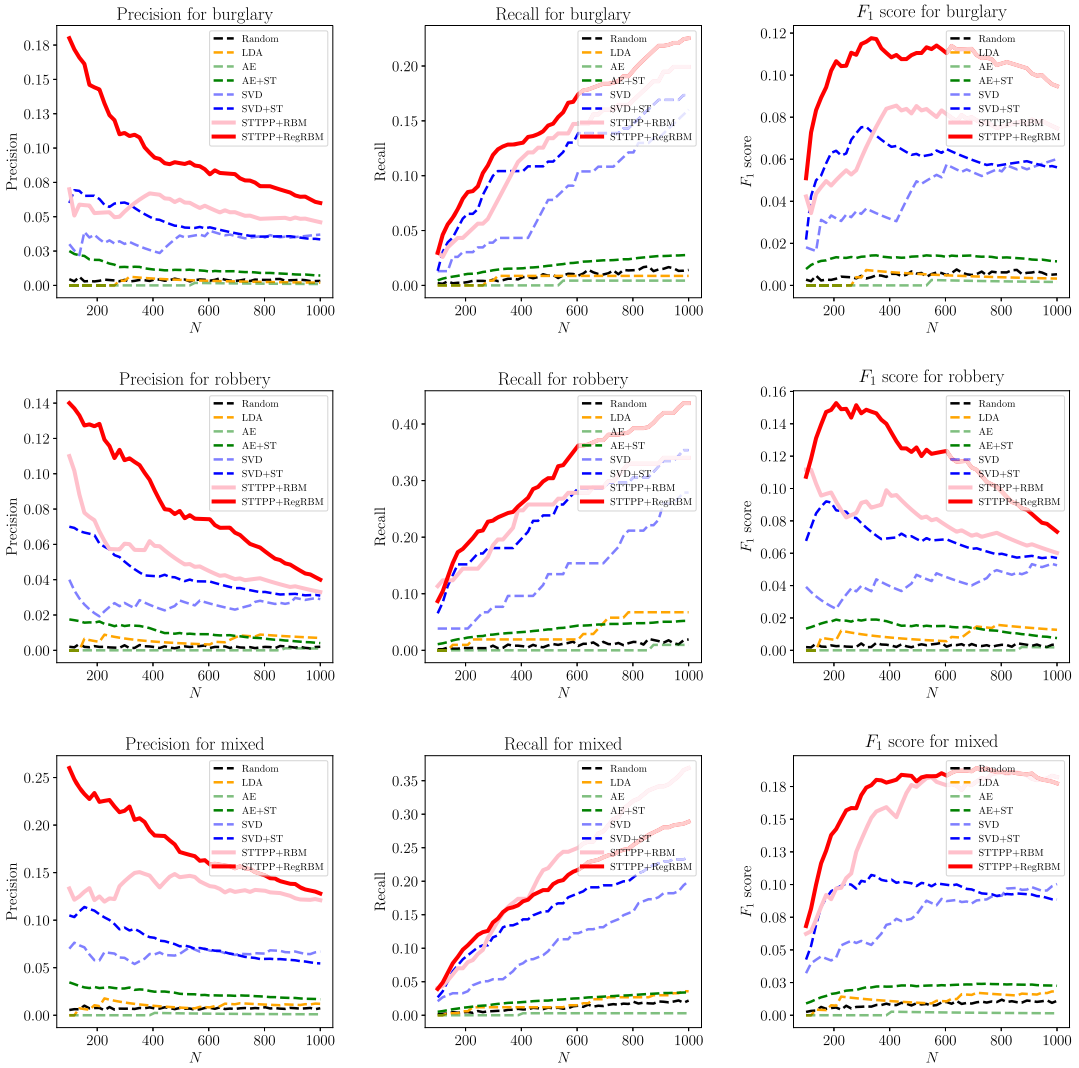


FIG. 11. *Comparison between our method* STTPP+RegRBM *and baselines with respect to precision, recall, and* F_1 *scores for crime linkage detection. We consider a different number of retrievals on data group burglary (the first row), robbery (the second row), and mixed (the third row). The vertical dash lines indicate the location of their best performance.*

pair of incidents as a similarity score. Based on this, we find the most similar pairs as the retrieval results.

As shown in Figure 11, our STTPP methods achieve a much higher F_1 score than other methods. The proposed method attains its best performance at $N = 314$, $N = 207$, and $N = 389$ on burglary, robbery, and mixed datasets. This indicates that properly incorporating spatiotemporal information will drastically improve the accuracy of linkage detection. In particular, our proposed STTPP+RegRBM greatly outperforms STTPP+RBM (without keyword selection) on single-category crimes, including robbery and burglary. This may be because the *M.O.*s are distributed around a small set of keywords in these cases, and keywords' selection plays a critical role in the feature extraction.

5. Discussions. This paper introduced a new framework for modeling police incidents as *spatiotemporal-textual* incidents and demonstrated its usage for crime linkage detection.

We addressed the challenge of a lack of labeled related incidents (and thus, we face an unsupervised learning problem). We developed a model based on the multivariate marked Hawkes processes, with marks being the incident textual descriptions' embeddings. We incorporated the textual similarity as a component in the point process intensity function while still enjoying computationally tractable likelihood function. We also developed a new embedding technique with keywords' selection for text modeling using a regularized Restricted Boltzmann Machine (RBM). The proposed method was validated using a real dataset and compared with alternative methods. We want to remark that the proposed algorithm has been adopted by the Atlanta Police Department and implemented in their AWARE system in 2018. Although we focus on police report analysis in this paper, our model can be used for other types of *spatiotemporal-textual* incident data, such as social media data and electronic health records.

As the linked crime series are hand-labeled by crime analysts, there are likely cases linking to the existing crime series; however, they are not identified yet. We could not consider these in the performance evaluation (due to the lack of the ground truth). It could happen that the algorithm actually detects linkages that are unknown but indeed exist. Nevertheless, the algorithm is shown to be effective, based on known labeled cases, and provides a potentially helpful tool for crime analysts to cast their attention to unknown but potential links.

We believe our approach based on embedding enjoys certain robustness. The text of the police reports is "noisy," in a highly unstructured form with variable length, and are entered by different police officers with personal styles. The embedding is based on a transformation of the raw text of a police report into a fixed-length bag-of-words vector and then mapped using the regularized RBM which captures the intrinsic similarities between text documents rather than apparent keywords. The other aspects of robustness for crime data analysis, such as missing data, are indeed very interesting and can be a topic of further research.

Finally, we want to remark that the crime linkage found by performing analysis based on the Hawkes processes model can be treated as a form of Granger causality (Xu, Farajtabar and Zha (2016)) which is a weaker causal inference for time series and does not consider confounding factors. For instance, reports from the same officer may be similar and reflect the officer's opinion. We believe that using embedding rather than the raw text bag-of-words features will alleviate such an issue to a certain extent since it captures the underlying similarities of crime incidents, as demonstrated in our real-data examples. Further causal inference analysis for crime linkage detection can be a future research direction. Another future direction is to extend the pairwise evaluation for crime linkage detection to consider higher-order interactions among incidents.

Funding. We thank Atlanta Police Foundation Award, National Science Foundation (NSF) DMS-1830210, CCF-1650913, and CMMI-1917624 for support.

SUPPLEMENTARY MATERIAL

Proofs and other technical results (DOI: 10.1214/21-AOAS1538SUPPA; .pdf). The online Appendix contains proofs of some technical results discussed in the text.

Python code (DOI: 10.1214/21-AOAS1538SUPPB; .zip). The folder "sttpp" contains the Python code of the proposed spatiotemporal-textual point process and its EM algorithm. The folder "rrbm" contains the Python code of the proposed regularized RBM model and the Python code for text preprocessing.

Desensitized police data (DOI: 10.1214/21-AOAS1538SUPPC; .zip). The data files include more than ten thousand 911-call records, where each record contains its exact time,

location, and the corresponding bag-of-words vector. The police text corpora can be read using *nltk.corpus* package.[1]

REFERENCES

ADDERLEY, R. (2004). The use of data mining techniques in operational crime fighting. In *Intelligence and Security Informatics* 418–425. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-25952-7_32

ADDERLEY, R. and MUSGROVE, P. (2003). Modus operandi modelling of group offending: A data-mining case study. *Int. J. Police Sci. Manag.* **5** 265–276. https://doi.org/10.1350/ijps.5.4.265.24933

ANDRADE, D. C., ROCHA-JUNIOR, J. A. B. and COSTA, D. G. (2017). Efficient processing of spatio-temporal-textual queries. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, *WebMedia'17*. 165–172. ACM, New York. https://doi.org/10.1145/3126858.3126877

BOUHANA, N. and JOHNSON, S. D. (2016). Consistency and specificity in burglars who commit prolific residential burglary: Testing the core assumptions underpinning behavioural crime linkage. *Legal Criminol. Psychol.* **21** 77–94. https://doi.org/10.1111/lcrp.12050

COCX, T. K. and KOSTERS, W. A. (2006). A distance measure for determining similarity between criminal investigations. In *Industrial Conference on Data Mining* 511–525. Springer, Berlin.

DAHBUR, K. and MUSCARELLO, T. (2003). Classification system for serial criminal patterns. *Artif. Intell. Law* **11** 251–269. https://doi.org/10.1023/B:ARTI.0000045994.96685.21

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DU, N., DAI, H., TRIVEDI, R., UPADHYAY, U., GOMEZ-RODRIGUEZ, M. and SONG, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD'16*. 1555–1564. ACM, New York. https://doi.org/10.1145/2939672.2939875

FISCHER, A. and IGEL, C. (2012). An introduction to restricted Boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Lecture Notes in Computer Science* **7441**. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-33275-3_2

FOX, E. W., SCHOENBERG, F. P. and GORDON, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Ann. Appl. Stat.* **10** 1725–1756. MR3553242 https://doi.org/10.1214/16-AOAS957

GOMAA, W. H. and FAHMY, A. A. (2013). Article: A survey of text similarity approaches. *Int. J. Comput. Appl.* **68** 13–18.

HALKIAS, X., PARIS, S. and GLOTIN, H. (2013). Sparse penalty in deep belief networks: Using the mixed norm constraint.

HARRIS, Z. S. (1954). Distributional structure. *Word* **10** 146–162. https://doi.org/10.1080/00437956.1954.11659520

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410 https://doi.org/10.1093/biomet/58.1.83

HINTON, G. E. (2002). Training products of experts by minimizing constractive divergence. *Neural Comput.* **14** 1771–1800. https://doi.org/10.1162/089976602760128018

HINTON, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade* 599–619. Springer, Berlin.

HONG, S., WU, M., LI, H. and WU, Z. (2017). Event2vec: Learning representations of events on temporal sequences. In *Web and Big Data* 33–47. Springer, Berlin.

KEYVANRAD, M. A. and HOMAYOUNPOUR, M. M. (2017). Effective sparsity control in deep belief networks using normal regularization term. *Knowl. Inf. Syst.* **53** 533–550. https://doi.org/10.1007/s10115-017-1049-x

KUANG, D., BRANTINGHAM, P. J. and BERTOZZI, A. L. (2017). Crime topic modeling. *Crime Science* **6** 12. https://doi.org/10.1186/s40163-017-0074-0

LAI, E. L., MOYER, D., YUAN, B., FOX, E., HUNTER, B., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2016). Topic time series analysis of microblogs. *IMA J. Appl. Math.* **81** 409–431. MR3564661 https://doi.org/10.1093/imamat/hxw025

LAN, G. (2020). *First-Order and Stochastic Optimization Methods for Machine Learning. Springer Series in the Data Sciences*. Springer, Cham. MR4219819 https://doi.org/10.1007/978-3-030-39568-1

[1] https://www.nltk.org/api/nltk.corpus.html

LIN, S. and BROWN, D. E. (2006). An outlier-based data association method for linking criminal incidents. *Legal Criminol. Psychol.* **41** 604–615. https://doi.org/10.1016/j.dss.2004.06.005

LIU, X., JIAN, C. and LU, C.-T. (2010). A spatio-temporal-textual crime search engine. In *Proceedings of the* 18*th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, *GIS'*10. 528–529. ACM, New York. https://doi.org/10.1145/1869790.1869881

LUO, H., SHEN, R., NIU, C. and ULLRICH, C. (2011). Sparse group restricted Boltzmann machines. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, *AAAI'*11. 429–434. AAAI Press, Menlo Park.

MA, L., CHEN, Y. and HUANG, H. (2010). AK-modes: A weighted clustering algorithm for finding similar case subsets. In 2010 *IEEE International Conference on Intelligent Systems and Knowledge Engineering* 218–223. https://doi.org/10.1109/ISKE.2010.5680876

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The eM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2392878 https://doi.org/10.1002/9780470191613

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. and DEAN, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the* 26*th International Conference on Neural Information Processing Systems—Volume* 2, *NIPS'*13. 3111–3119. Curran Associates, Red Hook.

MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546

NATH, S. V. (2006). Crime pattern detection using data mining. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, IEEE Press, New York. https://doi.org/10.1109/WI-IATW.2006.55

PARK, J., SCHOENBERG, F. P., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2021). Investigating clustering and violence interruption in gang-related violent crime data using spatial–temporal point processes with covariates. *J. Amer. Statist. Assoc.* 1–14.

PORTER, M. D. (2016). A statistical approach to crime linkage. *Amer. Statist.* **70** 152–165. MR3511045 https://doi.org/10.1080/00031305.2015.1123185

QUINN, C. J., COLEMAN, T. P., KIYAVASH, N. and HATSOPOULOS, N. G. (2011). Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.* **30** 17–44. MR2774388 https://doi.org/10.1007/s10827-010-0247-2

RANZATO, M. A., BOUREAU, Y.-L. and LECUN, Y. (2007). Sparse feature learning for deep belief networks. In *Proceedings of the* 20*th International Conference on Neural Information Processing Systems*, *NIPS'*07. 1185–1192. Curran Associates, Red Hook.

RANZATO, M., POULTNEY, C., CHOPRA, S. and LECUN, Y. (2006). Efficient learning of sparse representations with an energy-based model. In *Proceedings of the* 19*th International Conference on Neural Information Processing Systems*, *NIPS'*06. 1137–1144. MIT Press, Cambridge.

RASMUSSEN, J. G. (2011). Temporal point processes: The conditional intensity function.

REINHART, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* **33** 299–318. MR3843374 https://doi.org/10.1214/17-STS629

ROUSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

SHEN, T., JIANG, J., LIN, W., GE, J., WU, P., ZHOU, Y., ZUO, C., WANG, J., YAN, Z. et al. (2019). Use of overlapping group LASSO sparse deep belief network to discriminate Parkinson's disease and normal control. *Front. Neurosci.* **13** 396. https://doi.org/10.3389/fnins.2019.00396

SIMMA, A. and JORDAN, M. I. (2010). Modeling events with cascades of Poisson processes. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, *UAI'*10. 546–555. AUAI Press, Catalina Island, CA.

VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.

VEEN, A. and SCHOENBERG, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.* **103** 614–624. MR2523998 https://doi.org/10.1198/016214508000000148

WANG, B., DONG, H., BOEDIHARDJO, A. P., LU, C.-T., YU, H., CHEN, I.-R. and DAI, J. (2012). An integrated framework for spatio-temporal-textual search and mining. In *Proceedings of the* 20*th International Conference on Advances in Geographic Information Systems* 570–573.

WANG, T., RUDIN, C., WAGNER, D. and SEVIERI, R. (2015). Finding patterns with a rotten core: Data mining for crime series with cores. *Big Data* **3** 3–21. https://doi.org/10.1089/big.2014.0021

WOODHAMS, J., BULL, R. and HOLLIN, C. R. (2007). *Case Linkage*. Springer, Berlin. https://doi.org/10.1007/978-1-60327-146-2_6

XU, H., FARAJTABAR, M. and ZHA, H. (2016). Learning granger causality for Hawkes processes. In *Proceedings of the* 33*rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 1717–1726. PMLR, New York, New York, USA.

ZHANG, C., LIU, L., LEI, D., YUAN, Q., ZHUANG, H., HANRATTY, T. and HAN, J. (2017). TrioVecEvent: Embedding-based online local event detection in geo-tagged tweet streams. In *ACM SIGKDD International Conference* 595–604. https://doi.org/10.1145/3097983.3098027

ZHU, S. and XIE, Y. (2018). Crime incidents embedding using restricted Boltzmann machines. In 2018 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*) 2376–2380. https://doi.org/10.1109/ICASSP.2018.8461621

ZHU, S. and XIE, Y. (2019). Crime event embedding with unsupervised feature selection. In *ICASSP* 2019—2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*) 3922–3926. https://doi.org/10.1109/ICASSP.2019.8682285

ZHU, S. and XIE, Y. (2022). Supplement to "SpatioTemporal-Textual Point Processes for Crime Linkage Detection." https://doi.org/10.1214/21-AOAS1538SUPPA, https://doi.org/10.1214/21-AOAS1538SUPPB, https://doi.org/10.1214/21-AOAS1538SUPPC

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2002). Stochastic declustering of space-time earthquake occurrences. *J. Amer. Statist. Assoc.* **97** 369–380. MR1941459 https://doi.org/10.1198/016214502760046925

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *J. Geophys. Res.*, *Solid Earth* **109**.