# TEMPORAL EVENT MODELING OF SOCIAL HARM WITH HIGH DIMENSIONAL AND LATENT COVARIATES

by

**Xueying Liu**

**A Dissertation**

*Submitted to the Faculty of Purdue University*

*In Partial Fulfillment of the Requirements for the degree of*

**Doctor of Philosophy**


Department of Computer and Information Science

Indianapolis, Indiana

August 2022

# THE PURDUE UNIVERSITY GRADUATE SCHOOL
# STATEMENT OF COMMITTEE APPROVAL

**Dr. George Mohler, Co-Chair**

Department of Computer and Information Science

**Dr. Shiaofen Fang, Co-Chair**

Department of Computer and Information Science

**Dr. Murat Dundar**

Department of Computer and Information Science

**Dr. Mohammad Hasan**

Department of Computer and Information Science

**Dr. Honglang Wang**

Department of Mathematical Sciences

**Approved by:**

Dr. Shiaofen Fang

To my family

# ACKNOWLEDGMENTS

Words cannot express my gratitude to my professor and co-chair of my committee Dr. George Mohler for his invaluable patience and continuous support. I also could not have undertaken this journey without my exam committee Dr. Dundar, Dr. Fang, Dr. Hasan, and Dr. Wang, who generously provided knowledge and expertise.

I had the pleasure of working with my colleagues, and co-authors, especially Dr. Carter, Dr. Ray, Dr. Carson, and Dr. Xiao who have contributed on my research projects. I appreciated their inspirations, and insights on the projects. Thanks should also go to the graduate office for their help, prompt response, and moral support.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS

$\mathbb{P}(\cdot)$ Poisson distribution

$\mathbb{R}$ Set of real numbers

# ABBREVIATIONS

*i.i.d.* independent and identically distributed

# ABSTRACT

The counting process is the fundamental of many real-world problems with event data. Poisson process, used as the background intensity of Hawkes process, is the most commonly used point process. The Hawkes process, a self-exciting point process fits to temporal event data, spatial-temporal event data, and event data with covariates. We study the Hawkes process that fits to heterogeneous drug overdose data via a novel semi-parametric approach. The counting process is also related to survival data based on the fact that they both study the occurrences of events over time. We fit a Cox model to temporal event data with a large corpus that is processed into high dimensional covariates. We study the significant features that influence the intensity of events.

# 1. INTRODUCTION

A sequence of events is a collection of events that come one after another in a particular order. A temporal event sequence is a series of timestamps and covariates associated with each event ordered ascending in time. Each timestamp denotes the time when the event occurs, while the covariates are features indicating spatial information, type, and ID of the event.

A temporal event sequences could be Electronic Health Records (EHRs) which consist of patients' medical history, diagnoses, medications, etc. at the time of visits [1]–[3]. It could also be a sequence of time and longitude/ latitude coordinates indicating when and where incidents occur, for instance, earthquakes [4], [5], railway accidents [6] and gunshot violence [7], [8]. Modeling temporal event sequences has also seen an explosion of interest resulting in the study of social networks [9] and sequence prediction [10], intervention in social harm and infectious diseases [11], [12], and biological studies [13].

Counting process is a mathematical term that describes the action of counting the number of events as they occur in order [14]. A counting process is a non-negative, integer-valued, increasing stochastic process [15] that arises in many real-world scenarios, and it is extremely useful in statistical analysis of event sequence data.

Poisson processes is one of the examples of counting processes when the number of events fall within a region of finite size. This counting process is a random variable with a Poisson distribution [16]. A Hawkes process is a self-exciting point process whose background rate follows the intensity of a Poisson process [17], [18]. We will discuss Poisson processes and Hawkes processes in more details in Chapter 2.

In [19], renewal counting processes and their application in insurance were studied. Cox regression model for counting processes are proposed in 1972 [20] and thoroughly studied in [21] for censored survival data. More details of Cox models will be provided in Chapter 2.

In this dissertation, we discuss two models that are related to temporal event sequences modeling: (1) Hawkes point process models; (2) Cox proportional hazard models.

## 1.1 Latent Covariates Estimation for Heterogeneous Data

Heterogeneous data integration has emerged as an important issue as a huge amount of information becomes available in different formats and from various data sources [22], [23]. With the development of genomics technologies, there is an increasing needs in interpretation of heterogeneous gene expression data [24]. Similar problem also arises from wireless technology, where a mobile device needs to connect to different data sources and wireless networks [25].

Various methods have been proposed to handle heterogeneous data issues [26]. Early approaches to handling heterogeneous data include MIMIC (Mixtured indicators and multiple causes) model [27]. Later on there are researches that focus on embedding heterogeneous data into a latent space based on their co-occurrence [28], or clustering heterogeneous data into latent groups by similarity [29], [30]. A more recent approach is dealing with the heterogeneous data using neural networks [31]. Another approach is to model the high-dimensional heterogeneous data in a mixed model system [32], [33].

In Chapter 3, we consider the modeling of two datasets of space-time drug and opioid overdose events in Indianapolis. The first dataset consists of emergency medical calls for service (EMS) events. These events are non-fatal overdoses and include a date, time and location, but no information on the cause of the overdose. The second dataset consists of overdose death events (including location) and are accompanied by a toxicology report that screens for substances present or absent in the overdose event. We develop a marked point process model for the heterogeneous dataset that uses non-negative matrix factorization to reduce the dimension of the toxicology reports to several categories. We then use an Expectation-Maximization algorithm to jointly estimate model parameters of a Hawkes process and simultaneously infer the missing overdose category for the nonfatal overdose EMS data.

Criminology and public health disciplines have leveraged spatio-temporal event modeling in attempts to predict social harm for effective interventions [34]–[36]. Fifty percent of crime has been shown to concentrate within just 5 percent of an urban geography [37]. Geographic concentrations of drug-related emergency medical calls for service [38], drug activity [39],

and opioid overdose deaths mirror those of crime [40]. In particular, over half of opioid overdose deaths in Indianapolis occur in less than 5% of the city [40].

Patterns of repeat and near-repeat crime in space and time further suggest that not only does crime concentrate in place but that such events are an artifact of a contagion effect resulting from an initiating criminal event [41]. Similar observations have also explained the diffusion of homicide events [42]. Experiments of predictive policing models using spatio-temporal Hawkes and self-exciting point processes demonstrates that such empirical realities can be harnessed to direct police resources to reduce crime [43]. Thus, the inter-dependence and chronological occurrence of event types in crime and public health lend promise to how to best predict other social harm events, such as opioid overdoses.

We show that the point process defined on the integrated, heterogeneous data outperforms point processes that use only homogeneous coroner data. We also investigate the extent to which overdoses are contagious, as a function of the type of overdose, while controlling for exogenous fluctuations in the background rate that might also contribute to clustering. We find that opioid overdose deaths exhibit significant excitation, with branching ratio ranging from .72 to .98.

### 1.2 Survival Modeling on Transition of Suicidal Ideation

In 2019, approximately 47,500 deaths in the U.S. were attributed to suicide by the Center for Disease Control [44]. Given that suicide can be preventable by early intervention, recent data mining research has focused on the analysis of social media text, content and networks to identify suicide ideation and to better understand social media user risk, trajectories, interactions, and potential interventions.

One line of recent research focuses on detecting suicide ideation in online user content on sites such as Twitter and Reddit [45]–[47]. Other research has focused on modeling data from text messages [48] and surveys [49], [50]. While some studies utilized text based features input into classical machine learning models, more recently deep learning has been used to detect suicide ideation in text data [51]–[53]. A comprehensive survey on machine

learning for suicide detection can be found in [54], and [55] provides a survey on mining social networks to improve suicide prevention.

Reddit in particular has been the focus of recent data mining research on suicide, as several subreddits such as 'r/suicidewatch' provide forums for individuals thinking about suicide, drug addiction, and/or depression and who may be seeking help from others online. In [56], the authors analyzed discourse patterns of posts and comments on four Reddit online communities including r/depression, r/suicidewatch, r/anxiety and r/bipolar. In [57], detection methods were developed for suicide ideation in text on r/suicidewatch and related subreddits and in [58], the authors showed how to improve detection on r/suicidewatch by combining graph and language models. Other work has focused on determining the impact of the COVID-19 pandemic on suicide ideation on Reddit [59], creating an automated question answering system for suicide risk assessment using posts and comments extracted from r/suicidewatch [60], and predicting the degree of suicide risk on r/suicidewatch and related subreddits [61].

While a great deal of work has focused on detecting suicide ideation in online posts, there has been limited research on the temporal dynamics of users and suicide ideation. For example, a user who has suicidal thoughts may post on social media, at which point another may be able to intervene and provide mental health support. However, it is possible that earlier posts may have contained early indicators that could also have been points for interventions. In this work our goal is to better understand these earlier events through time-to-event survival analysis of transitions from other subreddit forums to r/suicidewatch.

In Figure 4.1, we show three example post sequences from Reddit that illustrate the type of dynamics we would like to model in the present paper. The first user posts on r/LongDistance several times, indicating that they feel sad and are having relationship problems due to long distance, the user then posts on r/teenagers a few times expressing their confusion and then later post on r/suicidewatch. Our goal is to identify which subreddits have a higher association with users transitioning to posting on r/suicidewatch, which text based features are associated with such transitions, and the time between posts from other forums and the first post on r/suicidewatch. We note that temporal dynamics of suicide ideation on Reddit were considered in [62], however the authors analyzed day of week and hour of day trends

in the times of posts rather than analyzing the inter-event time dynamics of transitions to r/suicidewatch.

### 1.3 Contribution

We summarize the major contributions of this dissertation:

- Latent covariates estimation: We combine heterogeneous data that contains spatial-temporal information and high dimensional marks to model the overall contagion pattern of events. We introduce a semi-supervising algorithm that learns the missing marks in the data. We compare our model to baseline models on both synthetic and real-world datasets to validate its superiority for event type prediction and model parameters recovery.

- Time-to-event modeling: We collect and process data from social network that contains temporal information of events as well as corresponding text and user interaction. We learn users' suicidal ideation transition from a convolution of suicidal thought analysis model, sentence embedding model, and Cox proportional hazard model. We find statistically significant features from high dimensional text, indicators for source and suicide risk.

### 1.4 Orgnization

This dissertation is organized as follows. In Chapter 2, we provide brief background of topics covered in later chapters, such as Hawkes models and Cox models and how they both connect to counting process. We also provide an overview on social harm events, including opioid overdose deaths and suicidal ideation. In Chapter 3, we discuss a semi-supervised estimation approach towards handling heterogeneous data and propose a Hawkes process model in recovering latent covariates of heterogeneous drug overdose datasets. In Chapter 4, we provide details on the data we collected from Reddit (r/suicidewatch and connected subreddits).We present our results of time-to-event modeling of transitions to r/suicidewatch, including the important features that indicate transitions.

# 2. BACKGROUND AND RELATED WORKS

In this chapter, we discuss background material and some related literature on: point process and its variations, including Poisson process, and Hawkes process. We also review the Cox proportional hazard model as a survival analysis model that relates to counting process. Topics such as heterogeneous data, social harm, and natural language processing models are also covered.

## 2.1 Point Process

**Definition 2.1** (Counting process [14]). *For a given time $t$, let $N(t)$ be the number of events that have occurred up to, and including $t$. Then $N(t)$ is a counting process.*

The process jumps up one unit each time an event is observed. This definition lends itself to a point process, a random collection of points falling in some space [63]. In many applications, temporal point processes are used to describe data that are contained within a finite set of time points [64].

One of the many well-known examples of a point process is the homogeneous (stationary) Poisson process, where the jumps occur randomly and independently of each other, with a constant rate [14].

**Proposition 2.1.1** (Khinchin's Existence Theorem). *For a homogeneous point process, the limit*

$$\lambda = \lim_{h \downarrow 0} \frac{\Pr\{N(0,h] > 0\}}{h} \tag{2.1}$$

*exists, though it may be infinite.*

Proof to Proposition 2.1.1 is provided in [65].

The parameter $\lambda$ is the intensity or rate of a point process. When $\lambda$ is finite, we can rewrite Equation (2.1) as

$$\begin{aligned} \Pr\{N(t, t+\Delta t] > 0\} &= \Pr\{\text{there is at least one point in } (t, t+\Delta t]\} \\ &= \lambda \Delta t + o(\Delta t) \ \ (h \to 0) \end{aligned} \tag{2.2}$$

### 2.1.1 Poisson process

#### Homogeneous Poisson process

**Definition 2.2** (Homogeneous Poisson process [66], [67]). *A homogeneous Poisson process $N(t)$ is characterized by a rate parameter, $\lambda$, when it satisfies two conditions:*

1. *For any interval $(t, t+\Delta t]$, $\Delta N(t, t+\Delta t] \sim \text{Pois}(\lambda \cdot \Delta t)$;*
2. *For any non-overlapping intervals $(t, t+\Delta t]$ and $(s, s+\Delta s]$, $\Delta N((t, t+\Delta t])$ and $\Delta N((s, s+\Delta s])$ are independent.*

#### Inhomogeneous Poisson process

**Definition 2.3** (Non-homogeneous Poisson process [66]). *An inhomogeneous Poisson process $N(t)$ with rate function $\lambda(t)$ is a point process with two conditions:*

1. *For any interval $(t_s, t_e]$, $\Delta N(t_s, t_e] \sim \text{Pois}\left(\int_{t_s}^{t_e} \lambda(t)dt\right)$;*
2. *For any non-overlapping intervals $(t, t+\Delta t]$ and $(s, s+\Delta s]$, $\Delta N((t, t+\Delta t])$ and $\Delta N((s, s+\Delta s])$ are independent.*

Let $f(t_{n+1}|\mathcal{H}_{t_n})$ be the `conditional density function` of the next event occurred at time $t_{n+1}$ given the history of previous events.$\mathcal{H}_{t_n}$ denotes the point process history up to time $t_n$. Then the `conditional intensity function` is defined [68] as

$$\lambda^*(t) = \frac{f(t|\mathcal{H}_{t_n})}{1 - F(t|\mathcal{H}_{t_n})}, \tag{2.3}$$

where $F(t|\mathcal{H}_{t_n})$ is the corresponding cumulative distribution function for any $t > t_n$.

The expression in (2.3) can be interpreted as the mean number of events in a region conditioned on the past [68], i.e., $\mathbf{E}[N([t, t+\Delta t])|\mathcal{H}_{t\_}]$.

For a Poisson process, $\lambda^*(t) = \lambda(t)$ since the conditional intensity function is independent of the past.

The general likelihood function of a Poisson process takes the form [65]

$$\begin{aligned} L_{(o,T]}(N; t_1, \cdots, t_N) &= \mathrm{e}^{-\Lambda(0,T]} \prod_{\mathrm{i}=1}^{N} \lambda(t_\mathrm{i}) \\ &= \exp\left( - \int_0^T \lambda(t) dt + \int_0^T \log\lambda(t) N(\mathrm{d}t) \right), \end{aligned} \tag{2.4}$$

where $\Lambda(a_\mathrm{i}, b_\mathrm{i}] = \int_{a_\mathrm{i}}^{b_\mathrm{i}} \lambda(x)\mathrm{d}x$.

### 2.1.2 Hawkes process

A point process $N(t)$ is `self-exciting` if $\mathrm{cov}\{N(s,t), N(t,u)\} > 0$ for $s < t < u$ [69]. In words, occurrence of points in such a point processes excites the process in the sense that the chance of a subsequent occurrence is increased for some period of time after the previous event [17]. These point processes are widely used in seismology [70], [71], neural science [72], [73], epidemiology [12], [74], [75], crime [34], [76], finance [77], [78] and more.

A `univariate` Hawkes process can be simply temporal [79]

$$\begin{aligned} \lambda(t|\mathcal{H}_t) &= \mu + \int_0^t g(t-u)\mathrm{d}N(u) \\ &= \mu + \sum_{\mathrm{i}:t_\mathrm{i}<t} g(t - t_\mathrm{i}), \end{aligned} \tag{2.5}$$

where $\mu$ is a constant background rate of events and $g(\cdot)$ is the triggering function which determinces the form of self-excitation. Sequence $\{t_1, t_2, \cdots, t_n\}$ denotes the observed times of events. The background process is a Poisson process with rate $\mu$, and $g(\cdot)$ determines the intensities of offspring processes of triggered events.

A spatio-temporal Hawkes process is an extension of Equation (2.5)

$$\lambda(s,t|\mathcal{H}_t) = \mu(s) + \sum_{\mathrm{i}:t_\mathrm{i}<t} g(s - s_\mathrm{i}, t - t_\mathrm{i}), \tag{2.6}$$

where $\mu(\cdot)$ is a function of space that models spatial clustering, and sequence $\{s_1, s_2, \cdots, s_n\}$ denotes the sequence of locations of observed occurrences. Form of $g(\cdot)$ is often taken to be separable in space and time for simplicity, while the choice of form depends on the actual

application. For example, in analyzing aftershock activities, a modified Omori formula (a decay law) [80] is used [81].

$$g(t-t_{\mathrm{i}}, x-x_{\mathrm{i}}, y-y_{\mathrm{i}}, m_{\mathrm{i}}) = \frac{K_0 \mathrm{e}^{a(m_{\mathrm{i}}-M_0)}}{(t-t_{\mathrm{i}}+c)^{(1+w)}((x-x_{\mathrm{i}})^2+(y-y_{\mathrm{i}})^2+d)^{(1+\rho)}}, \tag{2.7}$$

where the process history $\mathcal{H}_t = \{(t_{\mathrm{i}}, x_{\mathrm{i}}, y_{\mathrm{i}}, m_{\mathrm{i}}) : t_{\mathrm{i}} < t\}$ consists of earthquake epicenters $(x_{\mathrm{i}}, y_{\mathrm{i}})$, earthquake magnitudes $m_{\mathrm{i}}$ at time $t_{\mathrm{i}}$. $K_0$ is a normalizing constant related to the expected number of direct aftershocks triggered by earthquake i; $c, d, a, w$, and $\rho$ are normalizing parameters and $M_0$ is the lowest magnitude of earthquake considered.

In many other applications, researchers often model the contagion between events as a convolution of a spatial Gaussian mixture model and a temporal Hawkes process with an exponential kernel [82], [83]

$$g(t-t_{\mathrm{i}}, x-x_{\mathrm{i}}, y-y_{\mathrm{i}}) = K_0\, w \exp(-w(t-t_{\mathrm{i}})) \times \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-x_{\mathrm{i}})^2+(y-y_{\mathrm{i}})^2}{2\sigma^2}\right), \tag{2.8}$$

where we let $K_0$ denote the expected number of events that are triggered in the point process ; $w$ denote the parameter that controls how fast the rate $\lambda$ goes back to its baseline level $\mu$ after an occurrence, and $\sigma$ denote the standard deviation in the Gaussian triggering kernel. The value of $\sigma$ determines the extent to which the triggering effect spreads in space [83].

The likelihood of Hawkes processes (2.5) follows Equation (2.4)

$$L(\Theta) = \left[\prod_{\mathrm{i}=1}^{N} \lambda(t_{\mathrm{i}})\right] \exp\left(-\int_0^T \lambda(t)\mathrm{d}t\right), \tag{2.9}$$

where $\Theta$ is the parameter vector. Hence the log-likelihood is [65]

$$l(\Theta) = \sum_{\mathrm{i}=1}^{N} \log(\lambda(t_{\mathrm{i}})) - \int_0^T \lambda(t)\mathrm{d}t. \tag{2.10}$$

Log-likelihood of a spatio-temporal Hawkes process (2.6) follows similarly:

$$l(\Theta; X) = \sum_{\mathrm{i}=1}^{N} \log(\lambda(s_{\mathrm{i}}, t_{\mathrm{i}})) - \int_0^T \int_X \lambda(s,t) \mathrm{d}s\mathrm{d}t, \tag{2.11}$$

where $X$ is the spatial domain of occurrences.

### Multivariate Hawkes Process

For a $p$-dimensional multivariate Hawkes process, the conditional intensity function of the i-th node takes the form

$$\lambda_{\mathrm{i}}(t) = \mu_{\mathrm{i}} + \sum_{\mathrm{j}=1}^{p} \sum_{\mathrm{i}:t_{\mathrm{i}}<t} g_{\mathrm{ij}}(s - s_{\mathrm{i}}, t - t_{\mathrm{i}}) \mathrm{d}N_{\mathrm{j}}(s), \tag{2.12}$$

where $\mu_{\mathrm{i}}$ is the background rate for process (node) i, and $g_{\mathrm{ij}}(\cdot)$ models the intensities of triggering effect between events from node i and any other node j.

Log-likelihood of the multivariate Hawkes process (2.12) is given by [83]

$$l(\Theta; X) = \sum_{k=1}^{N} \log(\lambda_{\mathrm{i}_k}(t_k)) - \sum_{\mathrm{i}=1}^{p} \int_0^T \int_X \lambda_{\mathrm{i}}(s,t) \mathrm{d}s\mathrm{d}t, \tag{2.13}$$

where $\mathrm{i}_k$ is the point process that event $k$ is associated with.

### 2.1.3 Estimation

Hawkes process models are mostly fit using maximum likelihood [79], when the log-likelihood in Equations (2.11) and (2.13) are compulationally tractable to evaluate. Alternatively, [4], [84] showed the likelihood can be maximized with the expectation-maximization (EM) algorithm [85]. We introduce a latent variable $l_{\mathrm{i}}$ for each event i, which indicates where the event comes from. If the event is from the background, $l_{\mathrm{i}} = 0$; otherwise it is triggered

by a previous event j, and $l_{\mathrm{i}} = \mathrm{j}$. Then for a spatio-temporal Hawkes process in (2.11) the complete-data log-likelihood is given by

$$\begin{aligned} l_c(\Theta; X) = &\sum_{\mathrm{i}=1}^{N} \mathbb{I}(l_{\mathrm{i}} = 0) \log(\mu(s_{\mathrm{i}})) \\ &+ \sum_{\mathrm{i,j}}^{N} \mathbb{I}(l_{\mathrm{i}} = \mathrm{j}) \log(g(s_{\mathrm{i}} - s_{\mathrm{j}}, t_{\mathrm{i}} - t_{\mathrm{j}})) \\ &- \int_0^T \int_X \lambda(s,t) \mathrm{d}s\mathrm{d}t, \end{aligned} \tag{2.14}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

In the E-step, we estimate the probabilities based on the current parameters values $\hat{\Theta}$

$$\Pr(l_{\mathrm{i}} = \mathrm{j}) = \frac{g(s_{\mathrm{i}} - s_{\mathrm{j}}, t_{\mathrm{i}} - t_{\mathrm{j}})}{\lambda(s_{\mathrm{i}}, t_{\mathrm{i}})},\ t_{\mathrm{j}} < t_{\mathrm{i}} \tag{2.15}$$

$$\Pr(l_{\mathrm{i}} = 0) = \frac{\mu(s_{\mathrm{i}})}{\lambda(s_{\mathrm{i}}, t_{\mathrm{i}})}. \tag{2.16}$$

In the M-step, we update the model parameters $\Theta$ by maximizing the expectation of $l_c(\Theta)$ in (2.14)

$$\begin{aligned} \mathbb{E}[l_c(\Theta)] = &\sum_{\mathrm{i}=1}^{N} \Pr(l_{\mathrm{i}} = 0) \log(\mu(s_{\mathrm{i}})) \\ &+ \sum_{\mathrm{i,j}}^{N} \Pr(l_{\mathrm{i}} = \mathrm{j}) \log(g(s_{\mathrm{i}} - s_{\mathrm{j}}, t_{\mathrm{i}} - t_{\mathrm{j}})) \\ &- \int_0^T \int_X \lambda(s,t) \mathrm{d}s\mathrm{d}t. \end{aligned} \tag{2.17}$$

We utilize this approach in Chapter 3.

### 2.1.4 Survival Analysis

In this subsection, we provide an overview on survival data, and show the key ideas used in the analysis of survival data using the counting process approach.

### Censoring

Survival time (also known as failure time) is the time to an event of interest [86]. The analysis of group data is defined as survival analysis [87].

These analyses are difficult and complicated when individuals may not be observed to experience the event before the end of study [88], or we may lose touch with them during the study [87]. For an individual who is observed without failure (event) by the end of study and has a failure time, such incomplete observation of the failure time is called `censoring`.

### Counting Process

The survival function $S(t)$ [14], defined as $S(t) = P(T \geq t)$, gives the probability that the failure occurs later than an observed time $t$. The cumulative distribution function (CDF) of the survival time gives the cumulative probability for a given $t$:

$$F(t) = P(T < t) = 1 - S(t)$$

The hazard function $h(t)$ is defined as the probability that an event will occur in the time interval $[t, t + \Delta t)$ given that the event has not occurred before time $t$ [14]:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{\frac{d(1-S(t))}{dt}}{S(t)} = -\frac{d}{dt}\left[\log S(t)\right], \tag{2.18}$$

where $f(t)$ is the probability density function (PDF) of the survival time.

Consider $N$ independent non-negative random variables $T_1, T_2, \cdots, T_N$ corresponding to survival times of $N$ individuals and let $h_{\mathrm{i}}(t)$ denote the hazard rate at $T_{\mathrm{i}}$. We may obtain counting processes $N^c_{\mathrm{i}}(t) = \mathbb{I}\{T_{\mathrm{i}} \geq t\}$ from the survival times. These individual counting processes can add together to become an aggregated process $N^c(t)$ [14]:

$$N^c(t) = \sum_{\mathrm{i}=1}^{N} N^c_{\mathrm{i}}(t). \tag{2.19}$$

This process counts the occurrences of failures by time $t$.

It follows from intensity function (2.3) that $N_{\mathrm{i}}^{c}(t)$ have intensity processes [14]:

$$\lambda_{\mathrm{i}}^{c}(t) = h_{\mathrm{i}}(t)\mathbb{I}\{T_{\mathrm{i}} > t\},\ \forall \mathrm{i} = 1, \cdots, N, \tag{2.20}$$

and aggregated intensity process [14]:

$$\lambda^{c}(t) = h(t)Y^{c}(t), \tag{2.21}$$

where $Y^{c}(t) = \sum_{\mathrm{i}=1}^{N} \mathbb{I}\{T_{\mathrm{i}} \geq t\}$ is the number of individuals when event could occur before time $t$. Here we assume $h_{\mathrm{i}}(t) = h(t)$, which means the survival times are *i.i.d.* with hazard rate $h(t)$.

We extend this analysis to a study with censored survival time $\widetilde{T}_{\mathrm{i}}$. We introduce an indicator variable $D_{\mathrm{i}}$ [14] where

$$\begin{aligned} &D_{\mathrm{i}} = 1 \text{ when } \widetilde{T}_{\mathrm{i}} = T_{\mathrm{i}},\ \text{i.e., uncensored;} \\ &D_{\mathrm{i}} = 0 \text{ when } \widetilde{T}_{\mathrm{i}} < T_{\mathrm{i}},\ \text{i.e., censored.} \end{aligned} \tag{2.22}$$

Under the independent censoring assumption [14]:

$$\mathrm{P}(t \leq \widetilde{T}_{\mathrm{i}} < t + \Delta t, D_{\mathrm{i}} = 1 | \widetilde{T}_{\mathrm{i}} \geq t, \text{past}) = \mathrm{P}(t \leq T_{\mathrm{i}}, t + \Delta t | T_{\mathrm{i}} \geq t), \tag{2.23}$$

the counting processes

$$N_{\mathrm{i}}(t) = \mathbb{I}\{\widetilde{T}_{\mathrm{i}} \geq t, D_{\mathrm{i}} = 1\},\ \forall \mathrm{i} = 1, \cdots, N \tag{2.24}$$

denote the number of observed events.

Moreover, the independent censoring assumption provides the intensity process for $N_{\mathrm{i}}(t)$:

$$\lambda_{\mathrm{i}}(t) = h_{\mathrm{i}}(t)Y_{\mathrm{i}}(t), \tag{2.25}$$

where $Y_{\mathrm{i}}(t) = \mathbb{I}\{\widetilde{T}_{\mathrm{i}} \geq t\}$ indicates if individual i is at risk before time $t$.

Similarly as in the uncensored scenario, we are interested in the aggregated counting process [14]:

$$N(t) = \sum_{\mathrm{i}=1}^{N} N_{\mathrm{i}}(t) = \sum_{\mathrm{i}=1}^{N} \mathbb{I}\{\widetilde{T}_{\mathrm{i}} \geq t, D_{\mathrm{i}} = 1\}. \tag{2.26}$$

Equation (2.21) remains valid for process (2.26).

### The Cox Model

We are interested in the effect of covariates in survival studies, hence Cox proportional hazard model was brought up in [20]. This semi-parametric model is similar to multivariate regression model which enables the difference between survival times to be tested while taking other factors into consideration [86].

The Cox model [20] with $p$ covariates takes the form

$$h(t|\boldsymbol{x}_{\mathrm{i}}) = h_0(t) \exp(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_{\mathrm{i}}(t)), \tag{2.27}$$

where $\boldsymbol{x}_{\mathrm{i}}(t) = (x_{\mathrm{i}1}(t), \cdots, x_{\mathrm{i}p}(t))^T$ is a vector of covariates for individual i, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is the coefficients vector of regression.

For a counting process $N_{\mathrm{i}}(t)$, the intensity process derived from (2.25) is:

$$\lambda_{\mathrm{i}}(t) = Y_{\mathrm{i}}(t)\, h_0(t) \exp(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_{\mathrm{i}}(t)). \tag{2.28}$$

Then for an aggregated counting process (2.26), the intensity process is

$$\lambda(t) = \sum_{\mathrm{i}=1}^{n} \lambda_{\mathrm{i}}(t) = \sum_{\mathrm{i}=1}^{n} Y_{\mathrm{i}}(t)\, h_0(t) \exp(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_{\mathrm{i}}(t)). \tag{2.29}$$

The `conditional probability` of an occurrence for individual i at time $t$, given the past and knowing that an occurrence is observed at the time is [14]:

$$\boldsymbol{\pi}(\mathrm{i}|t) = \frac{\lambda_{\mathrm{i}}(t)}{\lambda(t)} = \frac{Y_{\mathrm{i}}(t) \exp(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_{\mathrm{i}}(t))}{\sum\limits_{\mathrm{i}=1}^{n} Y_{\mathrm{i}}(t) \exp(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_{\mathrm{i}}(t))}. \tag{2.30}$$

The partial likelihood for $\boldsymbol{\beta}$ is computed by multiplying the conditional intensities in (2.30) over all observations [14]:

$$L(\boldsymbol{\beta}) = \prod_{T_\mathrm{j}} \frac{\exp(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_{\mathrm{i_j}}(T_\mathrm{j}))}{\sum_{l \in R_\mathrm{j}} \exp(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_l(T_\mathrm{j}))}, \tag{2.31}$$

where $\mathrm{i_j}$ is the index of individual with an event occurs at $T_\mathrm{j}$, and $R_\mathrm{j} = \{l | Y_l(T_\mathrm{j}) = 1\}$ is the `risk set` at $T_\mathrm{j}$, i.e., the set of individuals at risk before $T_\mathrm{j}$ and have not been censored [14].

We will revisit the Cox proportional hazard model in Chapter 4.

### Kaplan-Meier and Nelson-Aalen Estimators

The Kaplan-Meier (KM) estimator [89] is defined to estimate probability of surviving in a given length of time. Formally put, the probability of an individual who has not experienced the event of interest at time $T_\mathrm{j}$, $S(T_\mathrm{j})$, is calculated from $S(T_{\mathrm{j}-1})$ [14], [90]. The KM estimator is [14]:

$$\widehat{S}(T_\mathrm{j}) = \widehat{S}(T_{\mathrm{j}-1}) \left(1 - \frac{d_\mathrm{j}}{n_\mathrm{j}}\right) = \prod_{T_\mathrm{i} \le T_\mathrm{j}} \left(1 - \frac{d_\mathrm{i}}{n_\mathrm{i}}\right) = \prod_{T_\mathrm{i} \le T_\mathrm{j}} \left(1 - \frac{1}{Y(T_\mathrm{i})}\right), \tag{2.32}$$

where $d_\mathrm{j}$ is the number of events at $T_\mathrm{j}$, and $n_\mathrm{j}$ is the number of individuals at risk at $T_\mathrm{j}$. $Y(T_\mathrm{i}) = \sum_{l=1}^{N} Y_l(T_\mathrm{i})$ is the number of individuals at risk before $T_\mathrm{i}$. The Kaplan-Meier survival curve is a plot of the Kaplan-Meier survival probability against time, which provides a summary of the data, such as median survival time [90].

We will use KM plots several times in Chapter 4 to interpret the model.

The Nelson-Aalen (NA) estimator [91]–[93] is a non-parametric estimator that is derived from the hazard rate $h(t)$. The cumulative hazard rate

$$H(t) = \int_0^t h(s)\mathrm{d}s \tag{2.33}$$

is estimated by the NA estimator [14]

$$\widehat{H}(t) = \sum_{T_\mathrm{j} \le t} \frac{d_\mathrm{j}}{n_\mathrm{j}} = \sum_{T_\mathrm{j} \le t} \frac{1}{Y(T_\mathrm{j})}. \tag{2.34}$$

An alternative estimator of the survival function $S(t)$ is obtained from Equations (2.18) and (2.33)

$$\widehat{S}(t) = \exp(-\widehat{H}(t)). \tag{2.35}$$

It has been shown that KM estimator is related to the NA estimator in the way that the survival function is related to the cumulative hazard rate [14]. The KM estimator of the survival function $S(t)$ is

$$\widehat{S}(t) = \prod_{T_j \leq t} \{1 - \Delta\widehat{H}(T_j)\}, \tag{2.36}$$

where $\Delta\widehat{H}(T_j) = 1/Y(T_j)$.

## 2.2 Social Harm

Social harm [94] is defined as socially constructed flows which cause damage and chaos to the structures and processes of human activities. As stated in [95], social harm is the "negative collective impacts associated with an illegal or disorderly act, or social control intervention". The concept of social harm has become more and more important when studying the problem of delivery of safety for individuals, their families, and the communities.

Suicidal behavior is a leading cause of death and disability worldwide. Fortunately, recent developments in suicide theory and research have shown that suicide can be preventable by early intervention [96].

In this section, we provide an overview on why social harms such as drug overdose need our attention and key developments in suicide research.

### 2.2.1 Drug Overdose

A drug overdose is taking excessive dose of a drug obtained either legally via a prescription, or illegally. An overdose may result in serious, harmful symptoms or death. According to CDC [97], "there were an estimated 100,306 drug overdose deaths in the United States during 12-month period ending in April 2021, an increase of 28.5% from the 78,056 deaths during the same period the year before".

Opioids are a leading cause in these deaths and these trends are characterized by three distinct time periods [98]. In the 1990s overdose deaths were driven by prescription opioid-related deaths [99], whereas reduced availability of prescriptions led to an increase of heroin-related deaths beginning in the 2010s [99]–[101]. Illicit fentanyl, a synthetic opioid 50 to 100 times more potent than morphine [102], has become a major cause of opioid-related deaths since around 2013. It is estimated that in 2016 around half of opioid-related deaths contained fentanyl [103], and fentanyl mixed into heroin and cocaine is likely contributing to many of these overdose deaths [104], [105].

In the estimated overdose death cases from CDC [106], number of opioids-related overdose deaths increased to 75,673 in the 12-month period ending in April 2021, a 35% increase from the year before. Based on provisional data released by CDC [97], number of overdose deaths from synthetic opioids (primarily fentanyl), psychostimulants (such as methamphetamine), cocaine, as well as natural and semi-synthetic opioids all increased.

Drug overdose adds adverse effects on the family system and individual members. These effects include emotional burden, economic burden, and relationship stress [107]. Drug overdose impacts the social functioning of individuals and also creates a burden for society. This social harm contributes to family and community, health, education, crime, and employment issues [108].

### 2.2.2 Suicidal Ideation

Suicide is a leading cause of death in the United States, with 45,979 deaths in 2020, according to CDC [109]. The number of individuals with suicidal ideation is even higher. Suicide affects all ages, especially among youth and young adults [110].

Some groups are at higher risk for suicide. These groups include veterans, sexual and gender minorities, victims of violence (e.g., childhood abuse, bullying, or sexual violence), and so much more [111].

In recent research of suicidal theory, one key development is to distinguish the development of suicidal ideation and the evolution from ideation to suicide attempts [112]. Suicidal ideation is defined as developing the idea of, or planning suicide; and suicide attempt is a

non-fatal, harmful behavior with an intent to die [112]. In Chapter 4, we conduct a study using Reddit data to explore the transition to suicidal ideation.

# 3. ESTIMATION OF HAWKES PROCESSES WITH HETEROGENEOUS DATA

A version of this chapter was previously published by Annals of Applied Statistics. Liu, X., Carter, J., Ray, B., Mohler, G. Point process modeling of drug overdoses with heterogeneous and missing data. Annals of Applied Statistics, 2021, Vol. 15, No. 1, 1–14 https://doi.org/10.1214/20-AOAS1384.

## 3.1 Introduction

Over 500,000 drug overdose deaths have occurred in the United States since 2000 and over 70,000 of these deaths occurred in 2017 [113]. Opioids are a leading cause in these deaths and these trends are characterized by three distinct time periods [98]. In the 1990s overdose deaths were driven by prescription opioid-related deaths [99], whereas reduced availability of prescriptions led to an increase of heroin-related deaths beginning in the 2010s [99]–[101]. Illicit fentanyl, a synthetic opioid 50 to 100 times more potent than morphine [102], has become a major cause of opioid-related deaths since around 2013. It is estimated that in 2016 around half of opioid-related deaths contained fentanyl [103], and fentanyl mixed into heroin and cocaine is likely contributing to many of these overdose deaths [104], [105].

Criminology and public health disciplines have leveraged spatio-temporal event modeling in attempts to predict social harm for effective interventions [34]–[36]. Fifty percent of crime has been shown to concentrate within just 5 percent of an urban geography [37]. Geographic concentrations of drug-related emergency medical calls for service [38], drug activity [39], and opioid overdose deaths mirror those of crime [40]. In particular, over half of opioid overdose deaths in Indianapolis occur in less than 5% of the city [40].

Patterns of repeat and near-repeat crime in space and time further suggest that not only does crime concentrate in place but that such events are an artifact of a contagion effect resulting from an initiating criminal event [41]. Similar observations have also explained the diffusion of homicide events [42]. Experiments of predictive policing models using spatio-temporal Hawkes and self-exciting point processes demonstrates that such empirical realities

can be harnessed to direct police resources to reduce crime [43]. Thus, the inter-dependence and chronological occurrence of event types in crime and public health lend promise to how to best predict other social harm events, such as opioid overdoses.

In this work we consider the modeling of two datasets of space-time drug and opioid overdose events in Indianapolis. The first dataset consists of emergency medical calls for service (EMS) events. These events are non-fatal overdoses and include a date, time and location, but no information on the cause of the overdose. The second dataset consists of overdose death events (including location) and are accompanied by a toxicology report that screens for substances present or absent in the overdose event. We develop a marked point process model for the heterogeneous dataset that uses non-negative matrix factorization to reduce the dimension of the toxicology reports to several categories. We then use an Expectation-Maximization algorithm to jointly estimate model parameters of a Hawkes process and simultaneously infer the missing overdose category for the nonfatal overdose EMS data.

We show that the point process defined on the integrated, heterogeneous data outperforms point processes that use only homogeneous coroner data. We also investigate the extent to which overdoses are contagious, as a function of the type of overdose, while controlling for exogenous fluctuations in the background rate that might also contribute to clustering. We find that opioid overdose deaths exhibit significant excitation, with branching ratio ranging from .72 to .98.

The outline of the paper is as follows. In Section II, we provide an overview of our modeling framework. In Section III, we run several experiments on synthetic data to validate the model and also on Indianapolis drug overdose data to demonstrate model accuracy on the application. We discuss several policy implications and directions for future research in Section IV.

## 3.2 Methods

### 3.2.1 Self-exciting point processes

In this work we consider a self-exciting point process of the form, [84]:

$$\lambda(x, y, t) = \mu_0 \nu(t) u(x, y) + \sum_{\mathrm{i}: t_\mathrm{i} < t} g(x - x_\mathrm{i}, y - y_\mathrm{i}, t - t_\mathrm{i}), \tag{3.1}$$

where $g(x, y, t)$ is a triggering kernel modeling the extent to which risk following an event increases and spreads in space and time. The background Poisson process modeling spontaneous events is assumed separable in space and time, where $u(x, y)$ models spatial variation in the background rate and $\nu(t)$ may reflect temporal variation arising from time of day, weather, seasonality, etc. The point process may be viewed as a branching process (or superposition of Poisson processes), where the background Poisson process with intensity $\mu_0 \nu(t) u(x, y)$ yields the first generation and then each event $(x_\mathrm{i}, y_\mathrm{i}, t_\mathrm{i})$ triggers a new generation according to the Poisson process $g(x - x_\mathrm{i}, y - y_\mathrm{i}, t - t_\mathrm{i})$.

We allow for self-excitation in the model to capture spatio-temporal clustering of overdoses present in the data. For example, a particular supply of heroin may contain an unusually high amount of fentanyl, leading to a cluster of overdoses in a neighborhood where the drug is sold and within a short time period.

Model 3.1 can be estimated via an Expectation-Maximization algorithm [4], [114], leveraging the branching process representation of the model. Let $L$ be a matrix where $l_{\mathrm{ij}} = 1$ if event i is triggered by event j in the branching process and $l_{\mathrm{ii}} = 1$ if event i is a spontaneous event from the background process. Then the complete data log-likelihood is given by,

$$\begin{aligned} &\sum_\mathrm{i} l_{\mathrm{ii}} \log(\mu_0 \nu(t_\mathrm{i}) u(x_\mathrm{i}, y_\mathrm{i})) - \int \mu_0 \nu(t) u(x, y, ) dx dy dt \\ &\quad + \sum_{\mathrm{ij}} l_{\mathrm{ij}} \log(g(x_\mathrm{i} - x_\mathrm{j}, y_\mathrm{i} - y_\mathrm{j}, t_\mathrm{i} - t_\mathrm{j})) \\ &\quad - \sum_\mathrm{j} \int g(x - x_\mathrm{j}, y - y_\mathrm{j}, t - t_\mathrm{j}) dx dy dt. \end{aligned} \tag{3.2}$$

Thus estimation decouples into two density estimation problems, one for the background intensity and one for the triggering kernel. Because the complete data is not observed, we introduce a matrix $P$ with entries $p_{ij}$ representing the probability that event i is triggered by event j.

Given an initial guess $P_0$ of matrix $P$, a non-parametric density estimation procedure can be used to estimate $u$ and $v$ from $\{t_k, x_k, y_k, p_{kk}\}_{k=1}^{N}$, providing estimates $u_0, v_0$ in the maximization step of the algorithm.

More specifically, we estimate $u$ and $v$ using leave-one-out kernel density estimation,

$$\begin{aligned} v(t_i) &= \frac{1}{N_b} \sum_{i \neq j} \frac{p_{jj}}{2\pi {b_1}^2} \exp\left\{ -\frac{(t_i - t_j)^2}{2{b_1}^2} \right\}, \\ u(x_i, y_i) &= \frac{1}{N_b} \sum_{i \neq j} \frac{p_{jj}}{2\pi {b_2}^2} \exp\left\{ -\frac{(x_i - x_j)^2 + (y_i - y_j)^2}{2{b_2}^2} \right\}, \end{aligned} \tag{3.3}$$

where $N_b = \sum_i p_{ii}$ is the estimated number of background events and $b_1$, $b_2$ are the kernel bandwidths that can be estimated via cross-validation or based on nearest neighbor distances. Because $u$ and $v$ are chosen to integrate to 1, we then have the ML estimate $\hat{\mu}_0 = N_b$

We assume the triggering kernel is given by a separable function that is exponential in time (Figure 3.1) with parameter $\omega$ and Gaussian in space with parameter $\sigma$ [82],

$$g(x, y, t) = K_0 \left( w \cdot \exp\{-wt\} \right) \cdot \frac{1}{2\pi\sigma^2} \cdot \exp\left\{ -\frac{1}{2\sigma^2}(x^2 + y^2) \right\}. \tag{3.4}$$

We then obtain an estimate for the parameters using weighted sample averages from the data $\{t_i - t_j, x_i - x_j, y_i - y_j, p_{ij}\}_{t_i > t_j}$,

$$\begin{aligned} \hat{K}_0 &= \sum_{t_i > t_j} p_{ij} \Big/ \sum_{i,j} p_{ij}, \\ \hat{w} &= \sum_{t_i > t_j} p_{ij} \Big/ \sum_{t_i > t_j} p_{ij} \cdot (t_i - t_j), \\ \hat{\sigma} &= \sqrt{\sum_{t_i > t_j} p_{ij} \cdot [(x_i - x_j)^2 + (y_i - y_j)^2] \Big/ 2 \cdot \sum_{t_i > t_j} p_{ij}} \end{aligned} \tag{3.5}$$

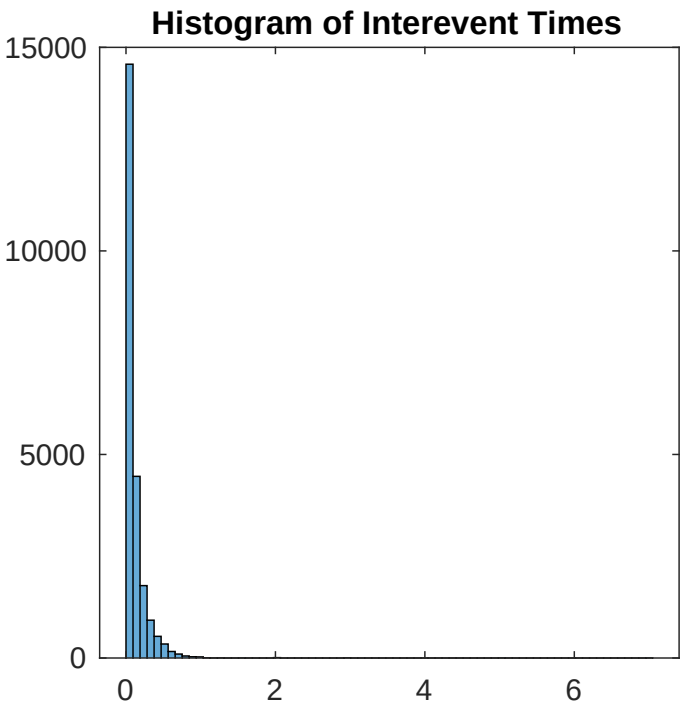



**Figure 3.1.** Histogram of inter-event times of real data, suggests that time triggering function is exponential.

In the estimation step, we estimate the probability that event i is a background event via the formula,

$$p_{\mathrm{ii}} = \frac{\mu_0 u(x_{\mathrm{i}}, y_{\mathrm{i}}) v(t_{\mathrm{i}})}{\lambda(x_{\mathrm{i}}, y_{\mathrm{i}}, t_{\mathrm{i}})}, \tag{3.6}$$

and the probability that event i is triggered by event j as,

$$p_{\mathrm{ij}} = \frac{g(x_{\mathrm{i}} - x_{\mathrm{j}}, y_{\mathrm{i}} - y_{\mathrm{j}}, t_{\mathrm{i}} - t_{\mathrm{j}})}{\lambda(x_{\mathrm{i}}, y_{\mathrm{i}}, t_{\mathrm{i}})}, \tag{3.7}$$

[115]. We then iterate for $n = 1, ..., N_{em}$ between the expectation and maximization steps until a convergence criteria is met:

1. Estimate $u_n, v_n$, and $g_n$ using (3.3) and (3.5).

2. Update $P_n$ from $u_n, v_n$, and $g_n$ using (3.6) and (3.7).

### 3.2.2 Modeling with heterogeneous event data

In this work we assume that we are given two datasets $A$ and $B$, though our modeling framework extends more generally to three or more. Event dataset $A$ contains low dimensional, unmarked space-time events, whereas dataset $B$ contains space-time events with high-dimensional marks. In our application, drug overdoses that do not result in death comprise dataset $A$, whereas those overdoses that do result in death are accompanied by a high-dimensional mark, namely the toxicology screen conducted by the coroner. Event dataset $B$ therefore contains a much smaller number of events compared to $A$.

Next we use non-negative matrix factorization (NMF) [116] to reduce the dimension of the high-dimensional mark of dataet $B$ into an indicator for $K$ groups. Each toxicology report consists of an indicator (presence or absence) for each one of 133 drugs the test screens. These reports then are input into a overdose-drug matrix analogous to a document-term matrix in text analysis using NMF. We then use NMF to factor overdose-drug matrices into the product of two non-negative matrices, one of them representing the relationship between drugs and topic clusters and the other one representing the relationship between topic clusters and specific overdose events in the latent topic space. The second matrix yields the cluster membership of each event (the cluster is the argmax of the column corresponding to each event).

### 3.2.3 Estimation of a marked point process with missing data

Merging dataset $A$ and $B$, we now have marked event data $(x_{\mathrm{i}}, y_{\mathrm{i}}, t_{\mathrm{i}}, k_{\mathrm{i}})$ where the mark $k_{\mathrm{i}}$ is one of $k = 1, ..., K$ clusters and is unknown for event data coming from $A$ but is known for event data from $B$.

Model (3.1) can be extended by adding in the group labels

$$\lambda^k(x, y, t) = \mu_0^k u^k(x, y) v^k(t) + \sum_{\substack{\mathrm{i}: t_{\mathrm{i}} < t \\ k_{\mathrm{i}} = k}} g^k(x - x_{\mathrm{i}}, y - y_{\mathrm{i}}, t - t_{\mathrm{i}}), \tag{3.8}$$

where $g^k$ is modelled as follows:

$$\begin{aligned} g^k(x,y,t) =& K_0^k \left(w^k \cdot \exp\left\{-w^k t\right\}\right) \cdot \\ & \frac{1}{2\pi{\sigma^k}^2} \cdot \exp\left\{-\frac{1}{2{\sigma^k}^2}\left[x^2+y^2\right]\right\}, \end{aligned} \tag{3.9}$$

Here we assume each cluster $k$ has its own parameters $(\omega^k, \mu_0^k, \sigma^k, K_0^k)$.

We then extend the branching structure matrix $P$ to a set of $K$ matrices, $P^k$, with initial guess $P_0^k$ and entries:

$$p_{\mathrm{ij}}^k = \begin{cases} \frac{1}{K}, & \text{if i = j and event i from } A \\ 1, & \text{if i = j, event i from } B \text{ and belongs to group } k \\ 0, & \text{otherwise} \end{cases}$$

Then $P^k$ can be updated similarly for each cluster $k = 1, \cdots, K$:

$$p_{\mathrm{ii}}^k = \frac{u^k(x_{\mathrm{i}}, y_{\mathrm{i}}) v^k(t_{\mathrm{i}})}{\lambda^k(x_{\mathrm{i}}, y_{\mathrm{i}}, t_{\mathrm{i}})}, \tag{3.10}$$

and

$$p_{\mathrm{ij}}^k = \frac{g^k(x_{\mathrm{i}} - x_{\mathrm{j}}, y_{\mathrm{i}} - y_{\mathrm{j}}, t_{\mathrm{i}} - t_{\mathrm{j}})}{\lambda^k(x_{\mathrm{i}}, y_{\mathrm{i}}, t_{\mathrm{i}})}, \tag{3.11}$$

where for each event i from dataset $A$, we have that $\sum_{k=1}^{K} \left( \sum_{t_{\mathrm{i}} \geq t_{\mathrm{j}}} p_{\mathrm{ij}}^k \right) = 1$, and for event i from dataset $B$ we have that $p_{\mathrm{ij}}^{\tilde{k}} = 0$ for all events j where $t_{\mathrm{i}} \geq t_{\mathrm{j}}$ and $\tilde{k}$ is not the group to which event i belongs.

The parameters are then estimated using $P^k$:

$$
\begin{aligned}
K_0^k &= \sum_{t_i > t_j} p_{ij}^k \Big/ \sum_{i,j} p_{ij}^k, \\
w^k &= \sum_{t_i > t_j} p_{ij}^k \Big/ \sum_{t_i > t_j} p_{ij}^k \cdot (t_i - t_j), \\
\sigma^k &= \sqrt{\sum_{t_i > t_j} p_{ij}^k \cdot [(x_i - x_j)^2 + (y_i - y_j)^2] \Big/ \left( 2 \cdot \sum_{t_i > t_j} p_{ij}^k \right)}, \\
\mu_0^k &= \sum p_{ii}^k.
\end{aligned}
$$

and the EM algorithm is iterated to convergence.

## 3.3 Results

### 3.3.1 Synthetic Data

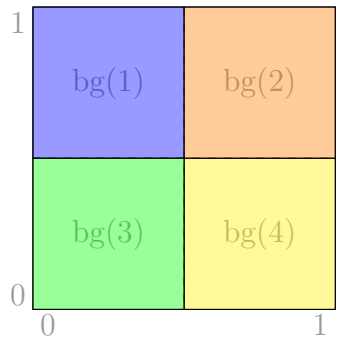



**Figure 3.2.** Simulation of events' location: for each background event, probabilities of falling in the purple, orange, green, and yellow regions are $bg(1)$, $bg(2)$, $bg(3)$, $bg(4)$, respectively.

To validate our methodology, we simulate point process data where data set $B$ has $K = 4$ groups with parameters given by those in Table 3.1 and 3.2. The background rate for each group is heterogeneous in space, with different rates in each quadrant in the unit square and homogeneous in time. Figure 3.2 and Table 3.1 illustrate how the background events are simulated: different background rates are assigned to each of the four different regions. Table 3.2 contains the true parameters for each group.

**Table 3.1.** Background rates of synthetic data.

| group | bg(1) | bg(2) | bg(3) | bg(4) |
|---|---|---|---|---|
| 1 | 0.1 | 0.2 | 0.3 | 0.4 |
| 2 | 0.4 | 0.3 | 0.2 | 0.1 |
| 3 | 0.4 | 0.4 | 0.1 | 0.1 |
| 4 | 0.1 | 0.4 | 0.1 | 0.4 |

**Table 3.2.** True parameters of synthetic data.

| group | w | $K_0$ | $\sigma$ | $\mu$ |
|---|---|---|---|---|
| 1 | 0.1 | 0.9 | 0.01 | 67 |
| 2 | 0.5 | 0.8 | 0.001 | 28 |
| 3 | 1 | 0.6 | 0.02 | 55 |
| 4 | 0.3 | 0.75 | 0.003 | 132 |

**Table 3.3.** Number of events from each group vs estimated number while dataset $A$ is 30% (left) and 90% (right) of all data.

| group | true # | estimated # |
|---|---|---|
| group 1 | 570 | 581 |
| group 2 | 154 | 145 |
| group 3 | 173 | 168 |
| group 4 | 431 | 434 |

| group | true # | estimated # |
|---|---|---|
| group 1 | 1197 | 1195 |
| group 2 | 71 | 56 |
| group 3 | 134 | 113 |
| group 4 | 380 | 418 |

We then simulate the missing data process by assigning 30% of the data to dataset $A$ (no label) and 70% to $B$. We find that the EM algorithm detailed above converges within 50 iterations.

We simulate 50 synthetic datasets and then estimate the true parameters, where the results are displayed in Figure 3.3. In the figure, the histograms of $w$, $K_0$, $\sigma$ and $\mu$ correspond to the estimates from the EM algorithm, where the red reference lines represent the average of the 50 results and the true value of the parameters are in blue. We find that our model is able to accurately recover both the true parameters and the event cluster membership up to the standard errors of the estimators.

In Table 3.3, we display the estimated number of events of each group (along with their actual values) when A has 30% of events as well as when 90% of events are assigned to $A$ (and thus unknown). We find in both experiments that the model is able to recover the cluster sizes accurately.

In Figure 3.4, we compare baseline models estimated only on $A$ or $B$ individually against the combined model. We also analyze the difference in performance versus the percentage of events assigned to dataset $A$. Here we find that the model estimated on both datasets always has higher likelihood than the models estimated only on one dataset.

### 3.3.2 Emergency Data and Toxicology Report

Next we analyze a dataset of drug overdose data from Marion County, Indiana (Indianapolis). The data spans the time period from January 14, 2010 to December 30, 2016. The fatal drug overdose dataset with toxicology reports (dataset $B$) consists of 969 events

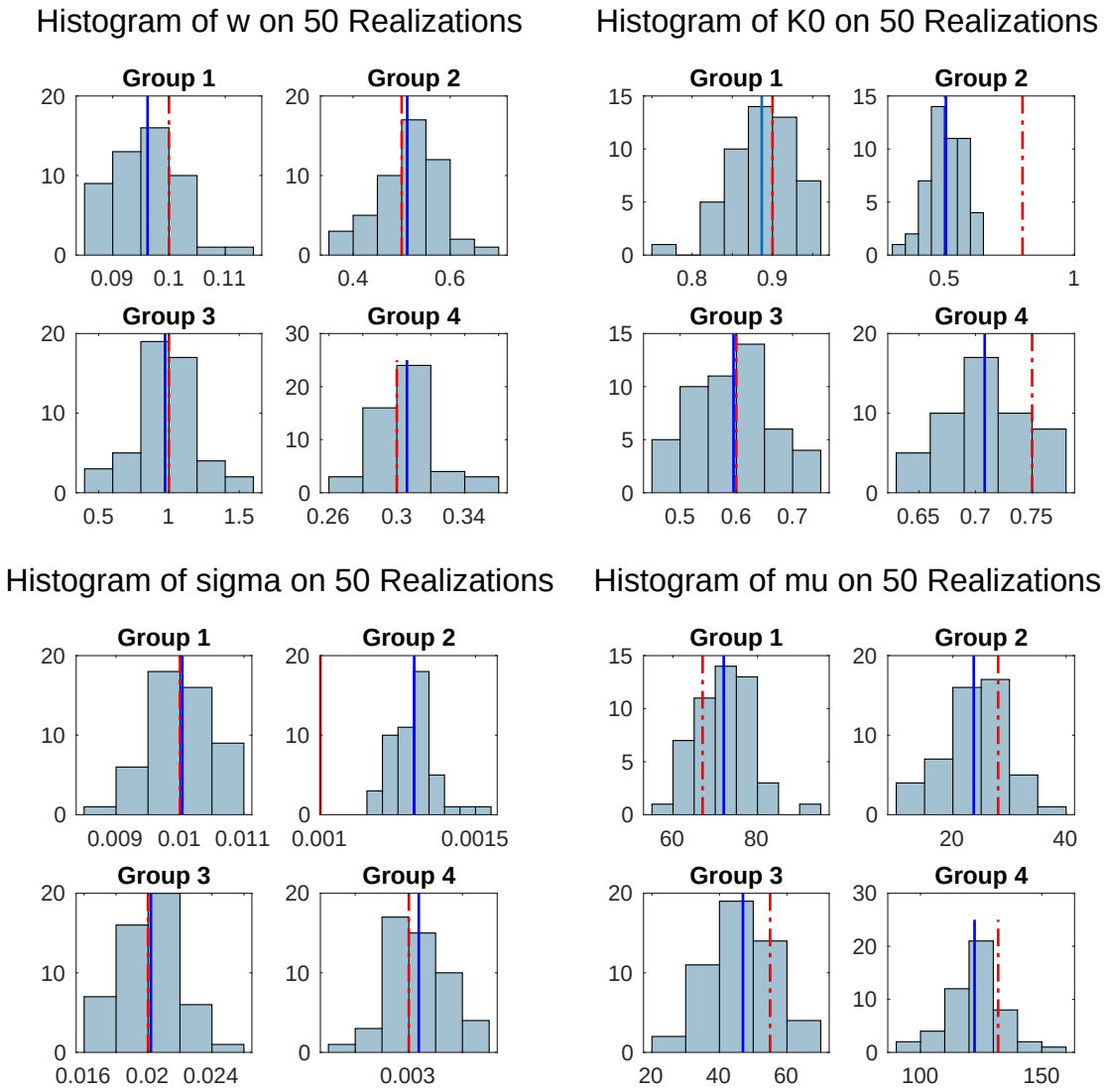



**Figure 3.3.** Parameters' true value (in red dash-dot line) and average of converged values (in blue solid line).

and the non-fatal, emergency medical calls for service dataset is 24 times bigger, with 22,049 unlabelled events.

We use NMF as described above to cluster the toxicology report data. We use coherence [117] to select the number of clusters, which we find to be $K = 4$ for our data (see Figure 3.5). In Table 3.4 we show the top 24 most frequent drugs and their frequencies present in the fatal overdose dataset and in Table 3.5 we display the top 5 most frequent drugs found in each NMF group. In Table 3.5 we find that the first group consists of illicit drugs (6-MAM and

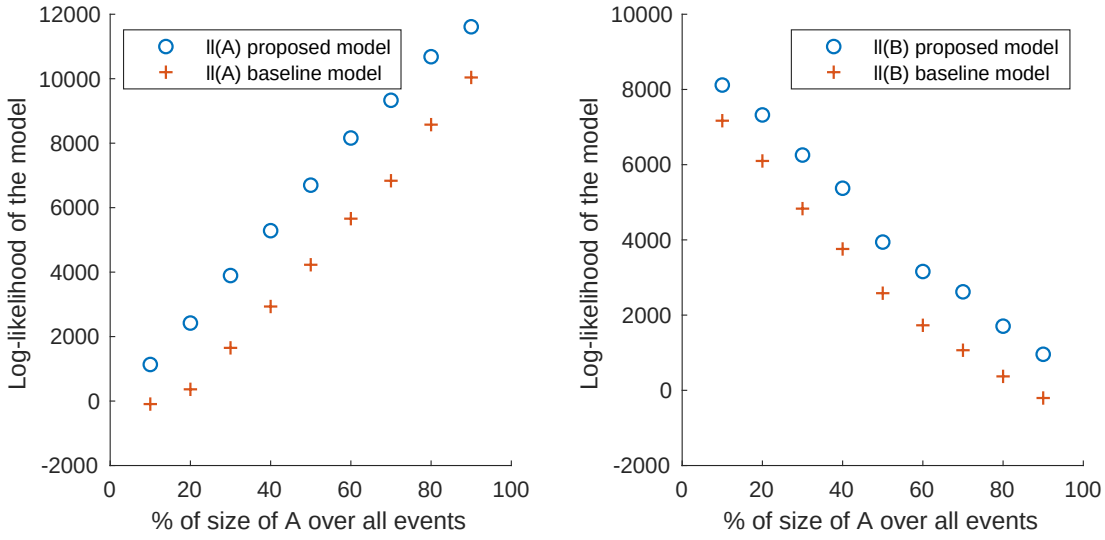



**Figure 3.4.** Log-likelihood of the model vs baseline model on individual datasets with different percentage of $A$. Left: likelihood evaluated on dataset $A$. Right: likelihood evaluated on dataset $B$.

heroin), whereas group 2 consists of mostly opioids that can be obtained via a prescription. Group 3 overdoses involve alcohol, whereas group 4 is fentanyl related overdoses.

Next we fit the point process model to the fatal and non-fatal overdose data. In Figure 3.6 we plot a heatmap of the inferred background events in space, disaggregated by group, along with the temporal trend of background events in Figure 3.7. We find that in time, the frequency of prescription opioid overdoses went down in Indianapolis, whereas illicit opioid overdoses, including the fentanyl group, increased over the same time period. In space, the illict drug hotspots are focused downtown, whereas the prescription opioid hotspots are more spread out in the city.

In Table 3.8 we display the estimated point process parameters. We see that for each group self-excitation plays a large role, where the branching ratio ranges from .72 to .98. In Table 3.6 and 3.7 we compare the log-likelihood values of the combined heterogeneous point process to baseline models estimated only on EMS or overdose death data. Here we find that including the EMS data improves the AIC values of the model for opioid overdose death, and the overdose death data improves the AIC of the model for EMS events.

**Table 3.4.** 24 most frequently present drugs.

| drug | frequency | drug | frequency |
|---|---|---|---|
| Hypnotic | 0.9617 | 11-Nor-9-carboxy-THC | 0.8113 |
| Lidocaine | 0.5588 | 11-Hydroxy-THC | 0.4856 |
| Phenobarbital | 0.4762 | Gastrointestinal | 0.3841 |
| Eszopiclone | 0.3841 | THC-Aggregate | 0.3580 |
| Promethazine | 0.3566 | Alcohol | 0.2451 |
| Ethanol | 0.2451 | Opioids | 0.2263 |
| Illicit | 0.2189 | Norfentanyl | 0.1773 |
| Amphetamine | 0.1760 | Acetylfentanyl | 0.1605 |
| Fentanyl | 0.1571 | Acetyl | 0.1343 |
| Methamphetamine | 0.1162 | Morphine | 0.1162 |
| Delta-9-THC | 0.0907 | 6-MAM | 0.0604 |
| Diazapam | 0.0537 | THC | 0.0524 |

**Table 3.5.** Top 5 drugs from each group.

| drug | group 1 | group 2 | group 3 | group 4 |
|---|---|---|---|---|
| 1 | 6-MAM | Benzodiazepine | Ethanol | Fentanyl |
| 2 | Heroin | Hydrocodone | Alcohol | Norfentanyl |
| 3 | Codeine | Oxycodone | Cocaine | Opioids |
| 4 | Morphine | Hydromorphone | Illicits | Amphetamine |
| 5 | Illicit | Oxymorphone | Benzodiazepine | Methamph. |

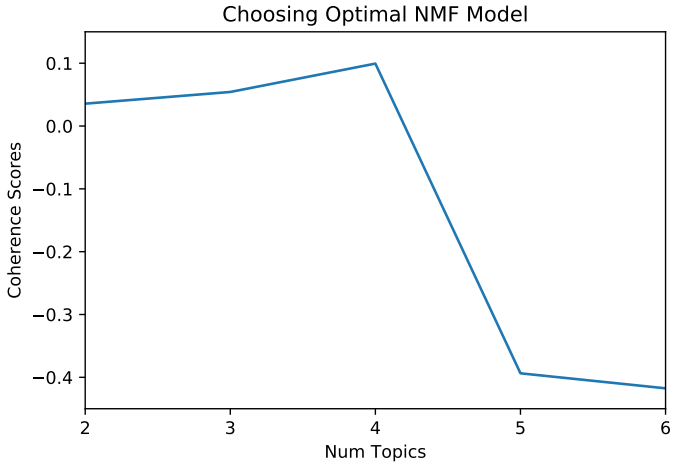



**Figure 3.5.** NMF coherence scores of drug overdose clusters vs number of topic clusters $K$.

To assess the model with a metric that better mirrors how interventions might work, we run the following experiment. For each day in January 15, 2010 to December 30, 2016, we estimate the point process intensity in each of 50x50 grid cells covering Indianapolis. We then rank the cells by the intensity and assign labels for whether an overdose occurs (1) or does not occur (0) during the next day. We then compute the area under the curve (AUC) of this ranking for the baseline and the proposed method. In practice, a point process model could be used to rank the top hotspots where overdoses are likely to occur and then those areas could be the focus of targeted interventions, such as distribution of naloxone that reverses the effects of an overdose.

In Table 3.6 and 3.7 we find that the AUC of the combined model evaluated on overdose death data is .85, compared to .81 for the model utilizing only overdose data. However adding overdose death data to the EMS data impairs the model in terms of AUC. The heterogeneous model has an AUC of .72 compared to .8 for the EMS data model (though the overdose death data does improve the AIC of the EMS data model).

**Table 3.6.** Different measurement results on EMS data.

| model | log-likelihood | df | AIC | AUC |
|---|---|---|---|---|
| baseline model | $4.9892 \times 10^4$ | 4 | $-9.9774 \times 10^4$ | **0.8032** |
| proposed model | $5.5752 \times 10^4$ | 16 | $\mathbf{-1.1147 \times 10^5}$ | 0.7159 |

**Table 3.7.** Different measurement results on Opioid overdose death data.

| model | log-likelihood | df | AIC | AUC |
|---|---|---|---|---|
| baseline model | $-3.6110 \times 10^3$ | 16 | $7.2540 \times 10^3$ | 0.8088 |
| proposed model | $1.7165 \times 10^3$ | 16 | $\mathbf{-3.4009 \times 10^3}$ | **0.8524** |

## 3.4 Chapter Summary

Heterogeneous data integration for model improvement promotes several policy and intervention benefits. Research using emergency medical services data has shown that persons who experience repeat non-fatal drug overdoses have a significantly higher mortality rate as compared to individuals without repeat events [118]. As our results suggest, toxicology data can be leveraged to model overdose diffusion across space and time, and diffusion varies across geographies. Taken together, integration of large-scale event data and overdose diffusion can sharpen policy interventions designed to reduce substance abuse and substance-related deaths. One such policy example is the deployment of nasal naloxone by police and EMS agencies which mitigates overdose effects [119].

Integration of heterogeneous data sources also help to contextualize and better understand the nuances of how social harms may affect different populations of people. As our study illustrates, prescription drug overdoses occur at higher rates in areas further from downtown Indianapolis, while illicit drug overdoses are more concentrated around the urban core of the city. These results underscore societal differences of opioid drug use. Consistent with community explanations of crime and social disadvantage [120], we observe that illicit drugs, which are more likely to result in mortality, may disproportionately impact minority communities. Current evidence indicates these trends are driven by heroin and synthetic

**Table 3.8.** Parameters of estimated model for each group.

| Group # | $K_0$ | w | $\mu$ | $\sigma$ |
|---|---|---|---|---|
| 1 | 0.9609 | 0.0153 | 4.0517 | 0.0148 |
| 2 | 0.9864 | 0.0170 | 2.8304 | 0.0313 |
| 3 | 0.7257 | 0.0094 | 28.7279 | 0.0044 |
| 4 | 0.9214 | 0.0143 | 4.4550 | 0.0091 |

opioid-related deaths as well as growing use of fentanyl-laced cocaine among African Americans [121], [122]. Moreover, these trends persist despite evidence that African Americans are less likely to be prescribed opioids for pain relative to Caucasians [123], which has been identified as a primary pathway to illicit opioid use [124]. Together, current evidence suggests the

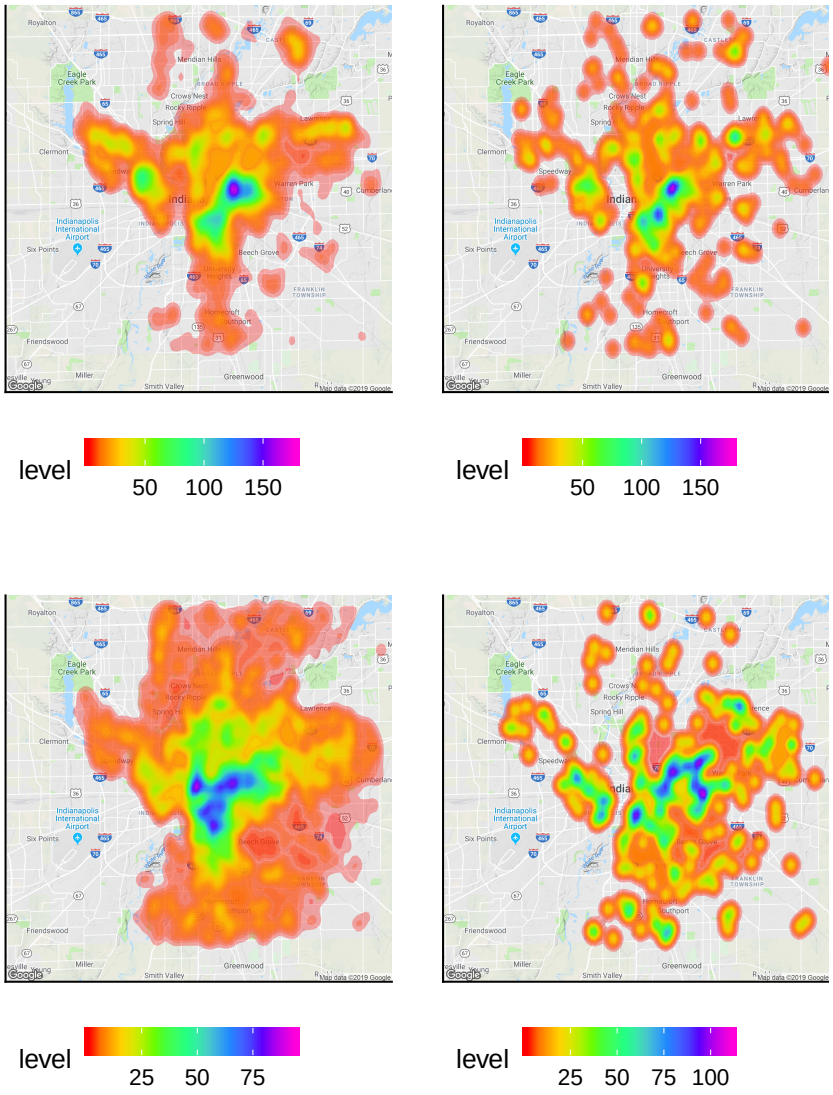



**Figure 3.6.** Heatmaps of non-fatal overdose events (left) and fatal overdose events (right). Top row: group 1; bottom row: group 2.

epidemiology of opioid use, especially illicit opioid use, is not well-defined for racial-ethnic minorities. Heterogeneous data integration is likely the most appropriate path forward to improve our understanding of this issue.

Our work here is also related to the analysis of free text data that accompanies crime reports [125]–[127] and other types of incidents, for example railway accidents [6]. While the majority of point process focused studies of crime and social harm use only location, time, and incident category as input into the model, we believe future research efforts on incorporating auxilliary, high-dimensional information into these models may yield improvements in model accuracy and also provide insight into the underlying causal mechanisms in space-time event contagion.

We do note that disentangling contagion patterns from other types of spatio-temporal clustering is challenging due to seasonal and exogeneous trends [76], [128]. Future work should also focus on investigating the extent to which drug overdose triggering found in the present study can be detected across cities and model specifications.

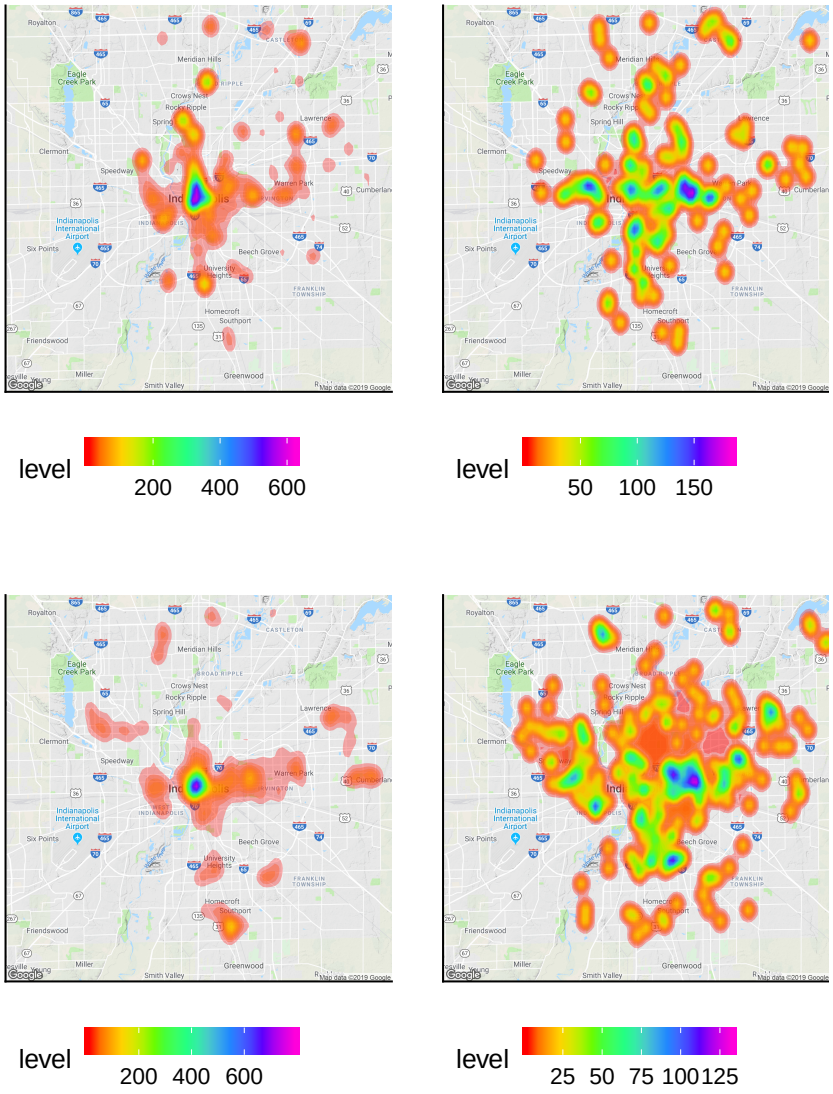



**Figure 3.6.** Heatmaps of non-fatal overdose events (left) and fatal overdose events (right) (cont.). Top row: group 3; bottom row: group 4.

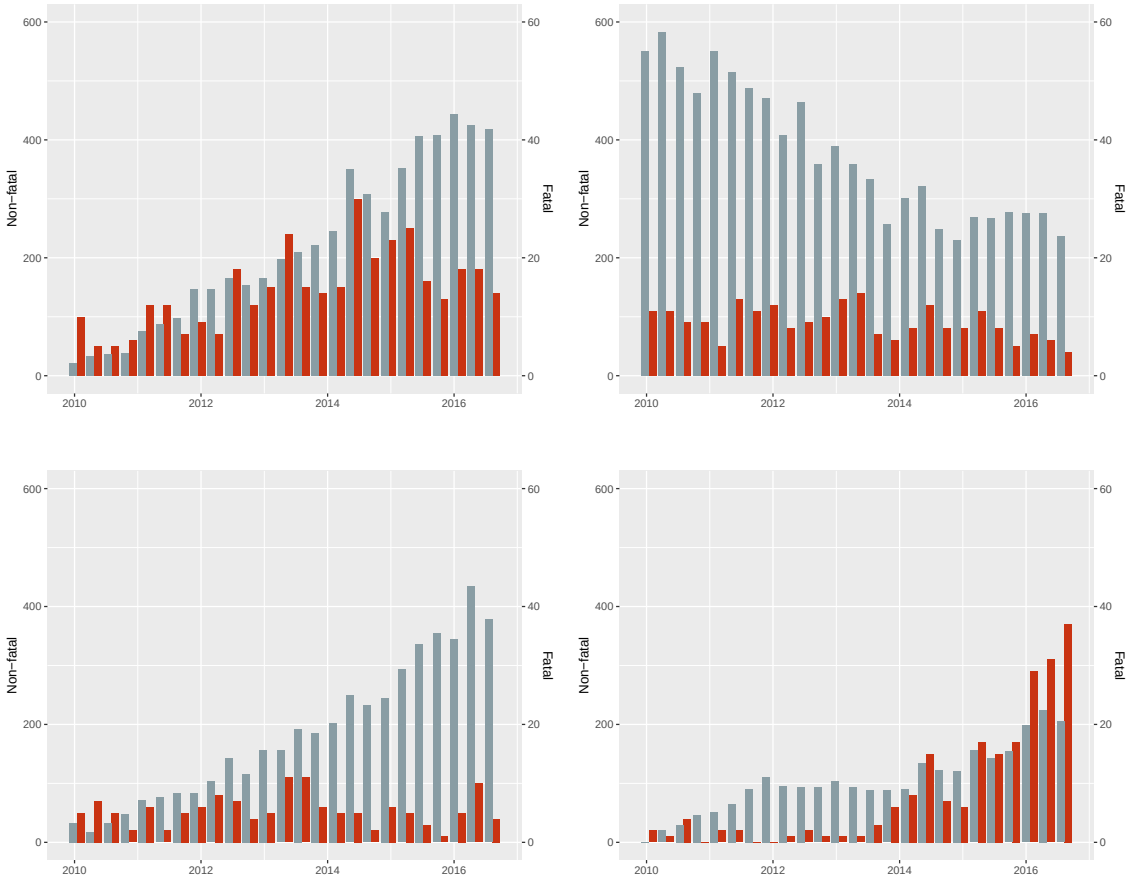



**Figure 3.7.** Histograms of non-fatal (grey) and fatal (red) overdose events for each group over time: group 1 (top left), group 2 (top right), group 3 (lower left), and group 4 (lower right).

# 4. TIME-TO-EVENT INTERVAL MODELING

A version of this chapter is ready to submit. Liu, X., Fang, S., Mohler, G., Carson, J., Xiao, Y. Time-to-event modeling of subreddit transitions to r/suicidewatch.

## 4.1 Introduction

In 2019, approximately 47,500 deaths in the U.S. were attributed to suicide by the Center for Disease Control [44]. Given that suicide can be preventable by early intervention, recent data mining research has focused on the analysis of social media text, content and networks to identify suicide ideation and to better understand social media user risk, trajectories, interactions, and potential interventions.

One line of recent research focuses on detecting suicide ideation in online user content on sites such as Twitter and Reddit [45]–[47]. Other research has focused on modeling data from text messages [48] and surveys [49], [50]. While some studies utilized text based features input into classical machine learning models, more recently deep learning has been used to detect suicide ideation in text data [51]–[53]. A comprehensive survey on machine learning for suicide detection can be found in [54], and [55] provides a survey on mining social networks to improve suicide prevention.

Reddit in particular has been the focus of recent data mining research on suicide, as several subreddits such as 'r/suicidewatch' provide forums for individuals thinking about suicide, drug addiction, and/or depression and who may be seeking help from others online. In [56], the authors analyzed discourse patterns of posts and comments on four Reddit online communities including r/depression, r/suicidewatch, r/anxiety and r/bipolar. In [57], detection methods were developed for suicide ideation in text on r/suicidewatch and related subreddits and in [58], the authors showed how to improve detection on r/suicidewatch by combining graph and language models. Other work has focused on determining the impact of the COVID-19 pandemic on suicide ideation on Reddit [59], creating an automated question answering system for suicide risk assessment using posts and comments extracted from r/suicidewatch [60], and predicting the degree of suicide risk on r/suicidewatch and related subreddits [61].

While a great deal of work has focused on detecting suicide ideation in online posts, there has been limited research on the temporal dynamics of users and suicide ideation. For example, a user who has suicidal thoughts may post on social media, at which point another may be able to intervene and provide mental health support. However, it is possible that earlier posts may have contained early indicators that could also have been points for interventions. In this work our goal is to better understand these earlier events through time-to-event survival analysis of transitions from other subreddit forums to r/suicidewatch.

In Figure 4.1, we show three example post sequences from Reddit that illustrate the type of dynamics we would like to model in the present paper. The first user posts on r/offmychest several times, indicating that they feel sad and are having relationship problems, and then later post on r/suicidewatch. Our goal is to identify which subreddits have a higher association with users transitioning to posting on r/suicidewatch, which text based features are associated with such transitions, and the time between posts from other forums and the first post on r/suicidewatch. We note that temporal dynamics of suicide ideation on Reddit were considered in [62], however the authors analyzed day of week and hour of day trends in the times of posts rather than analyzing the inter-event time dynamics of transitions to r/suicidewatch.

The outline of the paper is as follows. In Section 4.2, we describe our Cox proportional hazards modeling approach. In Section 4.3, we provide details on the data we collected from Reddit (r/suicidewatch and connected subreddits). In Section 4.4, we present our results of time-to-event modeling of transitions to r/suicidewatch, including the important features that indicate transitions. In Section 4.5, we discuss our results and directions for future work.

## 4.2 Model

Survival analysis [129] is a statistical method for analyzing the expected duration until an event occurs. The survival function $S(t)$ [14], defined as $S(t) = P(T \geq t)$, gives the probability that the time to the event occurs later than an observed time $t$. The cumulative

distribution function (CDF) of the time to event gives the cumulative probability for a given $t$:

$$F(t) = P(T < t) = 1 - S(t)$$

The hazard function $h(t)$ is defined as the probability that an event will occur in the time interval $[t, t + \Delta t)$ given that the event has not occurred before time t [14]:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)},$$

where $f(t)$ is the probability density function (PDF) of the time to event.

One feature of survival analysis is censoring of event times, as some users observation windows may not be large enough to have fully observed an event outcome. If for a given user an event of interest has occurred, then the survival time is known (fully observed), whereas for those that the events has not (yet) occurred, we only know that the waiting time exceeds the observation time [130]. These events with unknown survival time are referred to as censored data. In this study we restrict our analysis to users who post or comment on r/suicidewatch at least once.

The Cox proportional hazards model [20] is a standard Survival model that allows for the incorporation of covariates. The idea behind the Cox model is that the log-hazard of an individual is a linear function of a covariate vector $\boldsymbol{x}$ and parameter vector $\boldsymbol{\beta}$ and a population-level baseline hazard $h_0(t)$ that changes over time. The Cox's proportional hazard model has the form,

$$h(t|\mathbf{x}) = h_0(t)\exp[g(\boldsymbol{x})], \quad g(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}, \tag{4.1}$$

and has been used previously to model transitions to drug addiction and recovery on Reddit [131].

The Cox model in Equation 4.1 is fit to data in two steps[132]. First, the exponential part is fitted by maximizing the Cox partial likelihood (Equation 4.2), which does not depend on the baseline hazard, then the baseline hazard $h_0(t)$ is estimated using Breslow's method.

For individual i, let $T_{\mathrm{i}}$ denote the censored event time and $R_{\mathrm{i}}$ denote the set of all observations at risk at time $T_{\mathrm{i}}$. The Cox partial likelihood is defined as

$$L_{cox} = \prod_{\mathrm{i}} \left( \frac{\exp[g(\mathbf{x_i})]}{\sum_{\mathrm{j} \in R_{\mathrm{i}}} \exp[g(\mathbf{x_j})]} \right)^{D_{\mathrm{i}}}, \tag{4.2}$$

and the negative partial log-likelihood, which can be used as a loss function, is

$$\mathrm{ll}_{cox} = \sum_{\mathrm{i}} D_{\mathrm{i}} \log \left( \sum_{\mathrm{j} \in R_{\mathrm{i}}} \exp[g(\mathbf{x_j}) - g(\mathbf{x_i})] \right). \tag{4.3}$$

Let

$$S_x(t) = S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x}) \tag{4.4}$$

be the survival probability at time $t$, then the baseline probability is defined as follows:

$$S_0(t) = \mathrm{e}^{-\int_0^t h_0(t')dt'} = \mathrm{e}^{-H_0(t)}. \tag{4.5}$$

For an individual with features $X$,

$$S_x(t) = \mathrm{e}^{-\int_0^t h(t'|x)dt'} = [S_0(t)]^{\exp(\boldsymbol{\beta}^T \boldsymbol{x})}. \tag{4.6}$$

Let $\hat{\boldsymbol{\beta}}$ be the value of $\boldsymbol{\beta}$ that optimizes (4.2) and (4.3). Then the cumulative baseline hazard function can be estimated by the Breslow estimator[133]:

$$\widehat{H_0(t)} = \sum_{\mathrm{i}=1}^{n} \frac{D_{\mathrm{i}}}{\sum_{\mathrm{j} \in R_{\mathrm{i}}} \exp[g(\mathbf{x_j})]}. \tag{4.7}$$

Note here $D_{\mathrm{i}}$ is an indicator variable that event i is uncensored, as defined in (2.22).

We use the lifelines CoxPHFitter[1] in Python to fit the Cox model and estimate coefficients $\beta$ and baseline hazard.

[1]↑https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html

## 4.3 Data

The data is collected from Reddit[2], using PushShift[3] and PRAW[4] APIs. We first obtain a list of users who posted on r/suicidewatch between 1/1/2019 and 12/31/2021. We then randomly sample 2000 users and download their posts over the 3-year period, along with comments from these users that posted on r/suicidewatch and their comments and posts from other subreddits.

After dealing with exceptions on PRAW API and removing deleted posts, we collected more than 163k posts from over 1k users. We retained information including user name, post time, post content, post title, and on which subreddit the post was made. We then filtered out users with only one post, and posts that occurred after the user already posted on r/suicidewatch. We further cleaned the data by removing special tokens, detecting and translating posts in foreign languages into English, and performing spell check and making corrections. The data we use for analysis throughout this paper contains over 61k posts from 751 users.

We cut the data at 2020/12/31 23:59:59, and assign posts and comments prior to this time as training data, the rest being the test one. For the users who has posted by the cutoff time but post on r/SuicideWatch afterwards, their corresponding posts in the training data are labeled as `censored`. There are 129 censored users and more than 7k posts.

### 4.3.1 Suicide ideation detection model

For each post, we estimated a probability score as to whether a post contained language associated with suicide ideation using a pre-trained model [134]. The model is trained on text data collected from r/suicidewatch and r/depression and utilizes a LSTM neural network based on text embeddings with 'suicidal' vs. 'non-suicidal' binary labels. We use this model to estimate a score between 0 and 1 that represents the probability that a post in our dataset is associated with suicide. We then define posts to be 'high risk' if the score is higher than 0.95, and low otherwise.

[2]↑https://www.reddit.com
[3]↑https://github.com/pushshift/api
[4]↑https://praw.readthedocs.io/en/stable/

### 4.3.2 Summary statistics and figures

In Table 4.1, we provide average suicidal scores of some most frequently posted subreddits, where on r/suicidewatch, the score is noticeably higher. In Table 4.2a we provide a summary of the number of posts of each user, where the average number of posts is 53.4 per user. In Table 4.2b we provide a summary of the length of posts with high and low suicidal scores, where we find that high risk scores are associated with longer posts. Figure 4.2 displays the posting frequencies on each day of week. Posts with high suicidal scores are more likely to occur on Monday.

Figure 4.3 shows a histogram of users' posts, where 10% of users have only 2 posts and 76% of users have less than 50 posts each. Figure 4.4 shows the distribution of inter-event times between posts on other sub-reddits and r/suicidewatch. More than 3000 of the posts were made within 3.5 days of posting on r/suicidewatch. The longest waiting time is 698 days.

**Table 4.1.** Average suicidal score of posts on popular subreddits.

| Subreddit | Average suicidal score |
|---|---|
| AskReddit | 0.1730 |
| teenagers | 0.1455 |
| **SuicideWatch** | **0.8036** |
| memes | 0.1485 |
| depression | 0.6200 |
| AskOuija | 0.1416 |
| relationship_advice | 0.3657 |
| unpopularopinion | 0.1425 |
| dankmemes | 0.1631 |
| AmItheAsshole | 0.2027 |
| trees | 0.0958 |
| FortNiteBR | 0.0963 |
| selfharm | 0.4478 |
| awakened | 0.4240 |
| Advice | 0.4374 |
| NoFap | 0.2781 |

**Table 4.2.** Summary of data 1

| | |
|---|---|
| Max | 838 |
| Min | 2 |
| Mean | 81.34 |

(a) Number of posts and comments per user.

| | High | Low |
|---|---|---|
| Mean length | 539.18 | 164.38 |
| % on weekend | 27.69 | 27.94 |

(b) Posts and comments grouped by suicidal scores.

### 4.3.3 Topic models

We analyze the topic models of text data by first utilizing SentenceTransformers [135] - a BERT-based pretrained model that derives semantically meaningful sentence embeddings. We then perform KMeans with the embeddings that associate with high risk posts (i.e. suicidal score $> 0.95$). We search for the optimal $K$ using the "elbow" method, which suggests $K = 4$. Moreover, we extract keywords of each topic with the help of spaCy library in Python. The keywords are displayed in Table 4.3.

**Table 4.3.** Keywords extracted from posts with high suicidal score.

| Topic | Keywords |
|---|---|
| Topic 1 | right, tear, know |
| Topic 2 | think, sending, affected, died, unjustified, death, help, threaten, pills, life, traumatising, emotionally,... |
| Topic 3 | tried, turning, fix, way, find |
| Topic 4 | longer, like, want, future, realized, world, care, depression, time, method, meant, planned, fought, struggled, torture, hope,... |

Figure 4.5 demonstrates popularity of each topic over the 3-year observed timeframe.

### 4.3.4 Feature Selection

#### Keyword expansion

We use keyword expansion to determine a list of keywords related to suicide. Starting with a manually selected list of 40 keywords, we then use cosine similarity of word vectors to find the 10 most similar words to each. We use the word2vec implementation in gensim [136] to create the 100-dimensional word vector representations. This process results in a

keyword list of length 331. We next create dummy variables that indicate if a post contains each of these keywords. Figure 4.6 shows the histogram of number of keywords present in a post or comment. 40 most frequently occurred keywords are displayed in Table 4.4.

**Table 4.4.** Top 40 most frequent keywords.

| word | count | word | count | word | count | word | count |
|---|---|---|---|---|---|---|---|
| like | 8823 | god | 696 | health | 448 | red | 251 |
| good | 3462 | mental | 690 | kid | 423 | cry | 238 |
| way | 2708 | women | 589 | important | 381 | therapist | 228 |
| life | 2220 | using | 569 | death | 372 | account | 224 |
| thanks | 1899 | soon | 565 | eat | 356 | upset | 198 |
| work | 1657 | pain | 550 | type | 354 | weed | 186 |
| friends | 1254 | relationship | 534 | bring | 336 | ending | 182 |
| friend | 985 | damn | 527 | abuse | 329 | toxic | 174 |
| idea | 799 | check | 457 | depressed | 326 | emotional | 169 |
| place | 778 | anxiety | 455 | therapy | 306 | party | 164 |

#### Sources connecting to subreddit r/suicidewatch

We select the top 50 frequent subreddits (excluding r/suicidewatch) and create dummy variables that indicate if a post is from one of these subreddits. In Figure 4.7, we show the most frequently posted 15 subreddits. On average, the data contains 13.75 posts and comments from each subreddit.

#### Index of topics

The topic indices obtained from part 4.3.3 are transferred into dummy variables.

### 4.4 Results

We fit a Cox proportional hazards model to our data, with a penalizer term of 5, and within the summary table of results, we focus on the variables with a p-value less than 0.05. In Table 4.5, we show the coefficients of these statistically significant variables. The subreddits with the highest coefficients (indicating sooner transition to r/suicidewatch) include r/selfharm, r/Wishlist, r/awakened, r/BreakUps and r/MadeOfStyrofoam (which is a forum

for selfharm discussion). Subreddits associated with longer time-to-event intervals between an initial post and a subsequent post on r/suicidewatch include r/LivestreamFail, r/AvPD, r/ftm and r/PurplePillDebate (a forum to discuss sex and gender issues). While a majority of the keywords were not statistically significant, both 'pain' and 'life' were associated with shorter time-to-event periods, as was the high risk category based on the suicide detection model described above. Keyword 'women' appeared to be associated with longer time-to-event interval. Topic feature is not statistically significant.

In Figure 4.8, we display the estimated Kaplan Meier curve for the distribution of time-to-event disaggregated by high vs. low suicidal scores. Here we find that the time-to-event distribution has a shorter tail for higher risk scores, indicating that posts with high risk scores are associated with subsequent r/suicidewatch posts occurring sooner. In Figure 4.10, we display the transition network from other subreddits (yellow indicating a positive association, blue indicating a negative association) to r/suicidewatch along with the average transition time between the final post preceding a post on r/suicidewatch and their first post on r/suicidewatch. On each edge, the number represents the average number of days between subreddit and r/suicidewatch posts when the suicidal score is high (low). No post is found from r/cats, r/MortalKomat, r/sweden and r/Eminem with a high suicidal score.

We predict expected remaining lifetime of each censored user and compute the concordance between prediction and ground truth, the model obtains a concordance index of 0.5123. We also compute AUC by dividing the entire future into 30-day interval. We label an interval 1 if transition to r/SuicideWatch occurred in the interval. This gives us an AUC of 0.8214.

We plot predicted survival curve of each user in Figure 4.1 and show the results in Figure 4.9. Each curve represents the predicted survival probability from the time of the post.

## 4.5 Chapter Summary

In this research we collected a large corpus of suicide related posts from r/suicidewatch, along with earlier posts made by users on other subreddits. We then fit a Cox proportional hazards model to predict the time-to-event between earlier posts and later posts on r/suicide-

**Table 4.5.** Coefficients of significant variables.

| **Subreddit indicators** | | | |
|---|---|---|---|
| depression | 0.06 | LivestramFail | -0.28 |
| teenagers | 0.05 | fireemblem | -0.08 |
| relationship_advice | 0.05 | MortalKombat | -0.12 |
| awakened | 0.15 | AvPD | -0.37 |
| selfharm | 0.09 | Sweden | -0.23 |
| MadeOfStyrofoam | 0.11 | Traaa...nnnns | -0.1 |
| Wishlist | 0.17 | ftm | -0.19 |
| BPD | 0.08 | PurplePillDebate | -0.29 |
| Eminem | 0.14 | | |
| cats | 0.1 | | |
| unpopularopinion | 0.06 | | |
| BreakUps | 0.12 | | |
| **Keyword indicators** | | **Suicidal score** | |
| "pain" | 0.04 | score | 0.04 |
| "women" | -0.07 | | |
| "life" | 0.02 | | |

watch. We found statistically significant features using indicators for subreddit, keyword, or suicide risk. While some patterns match existing intuition, for example r/teenagers and r/relationship_advice are positively associated with posting sooner on r/suicidewatch, others were more surprising. For example, the average time between a high risk post on r/Wishlist and a consecutively following post on r/suicidewatch is 10.2 days (less than the 97.2 day average time between events on r/depression and r/suicidewatch). Our results indicate potential points of earlier intervention and analysis of associated subreddits to suicide may yield new hypotheses for suicide researchers to investigate.

Future research may improve upon our work in several ways. While we used a Cox proportional hazards model, a deep learning based survival model [137] may yield improvements to accuracy. Also, we did not explore the social network of individuals and how interactions on the network may be predictive of transitions to suicide ideation. In future work we hope to analyze posting behavior of network connections and people whom a given person interacts with, and determine how those interactions may act to protect against or lead to higher risk of suicide ideation observed online.

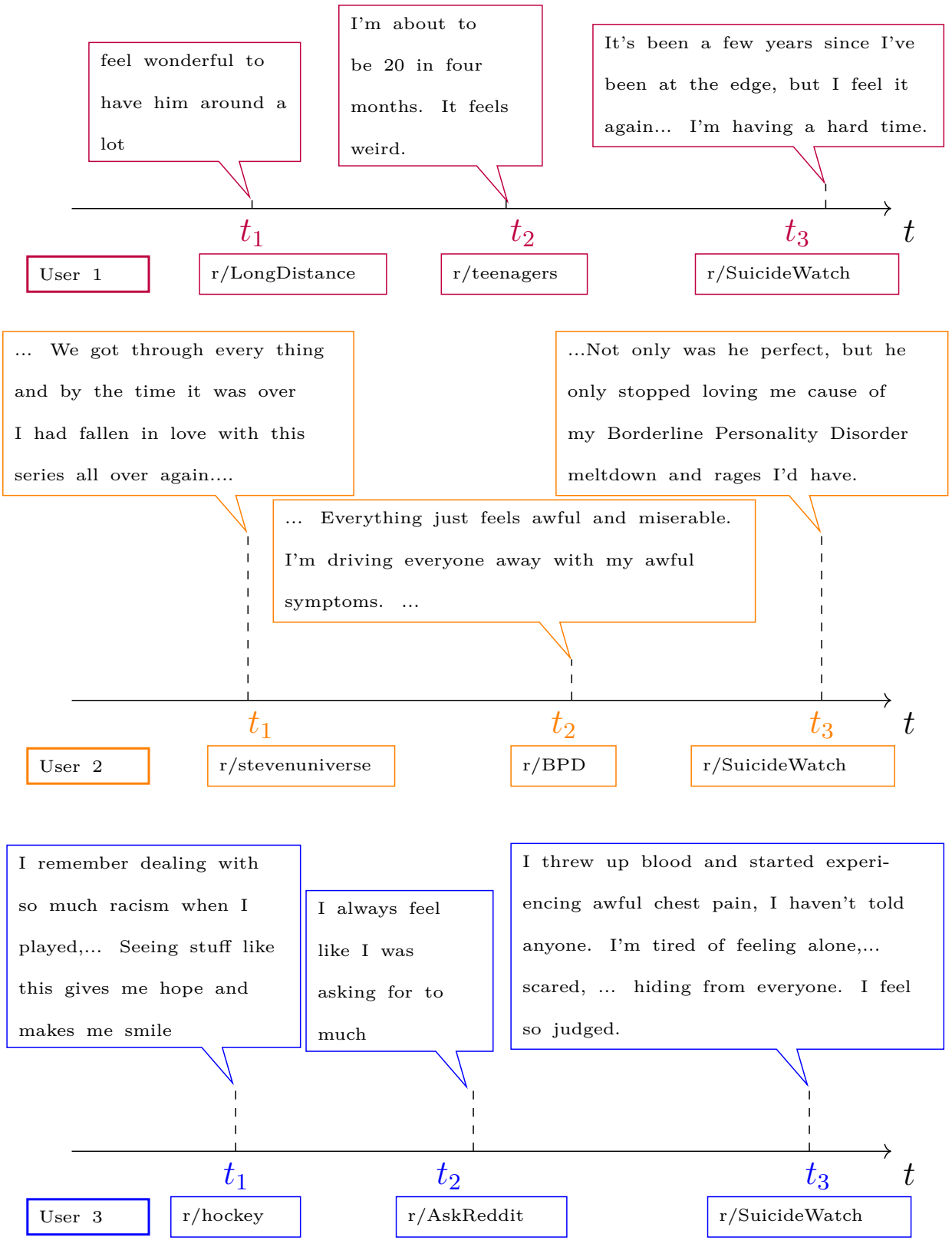



**Figure 4.1.** Posting sequences of 3 Reddit users.

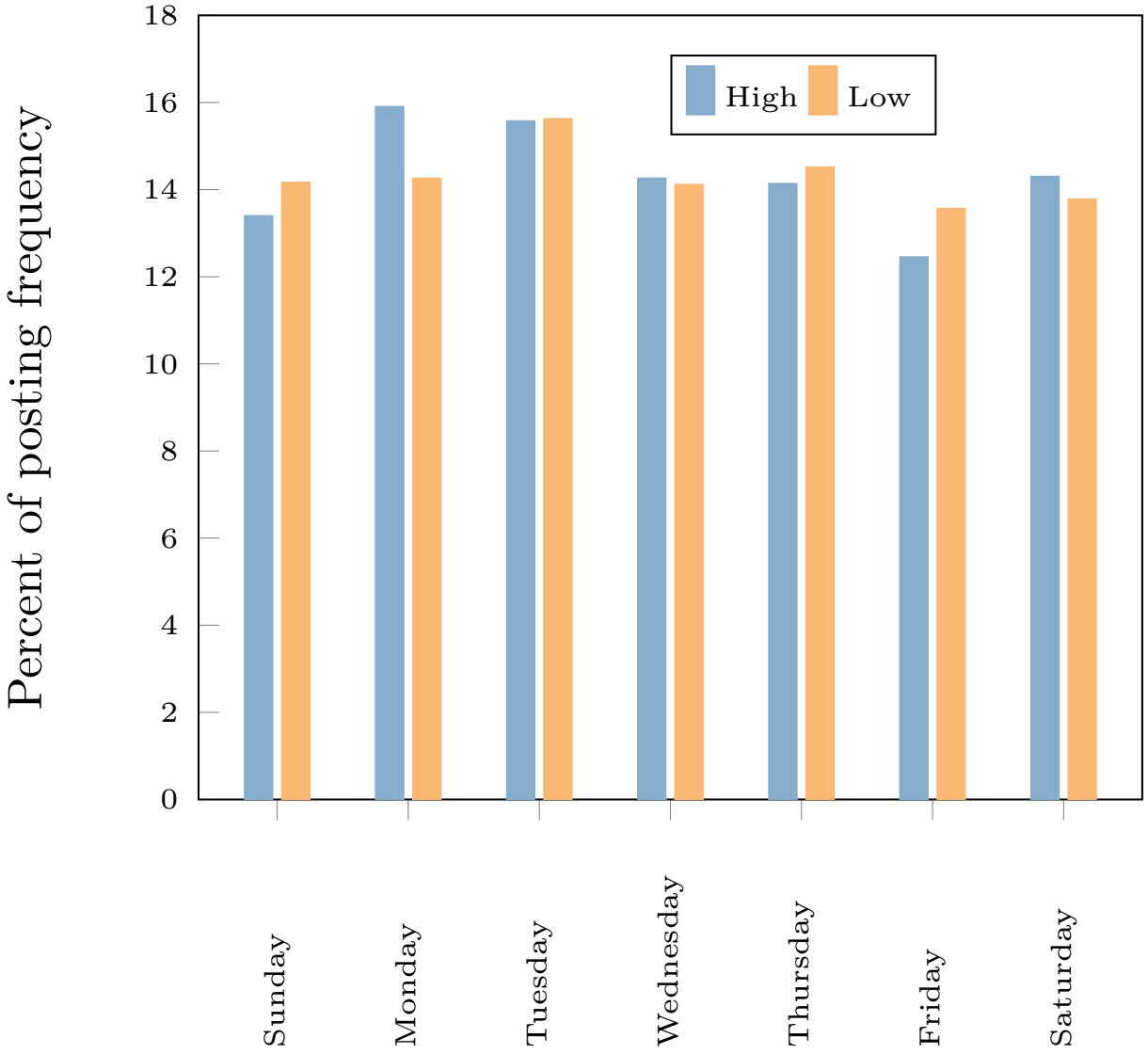



**Figure 4.2.** Posting frequency on day of week by suicidal score group.

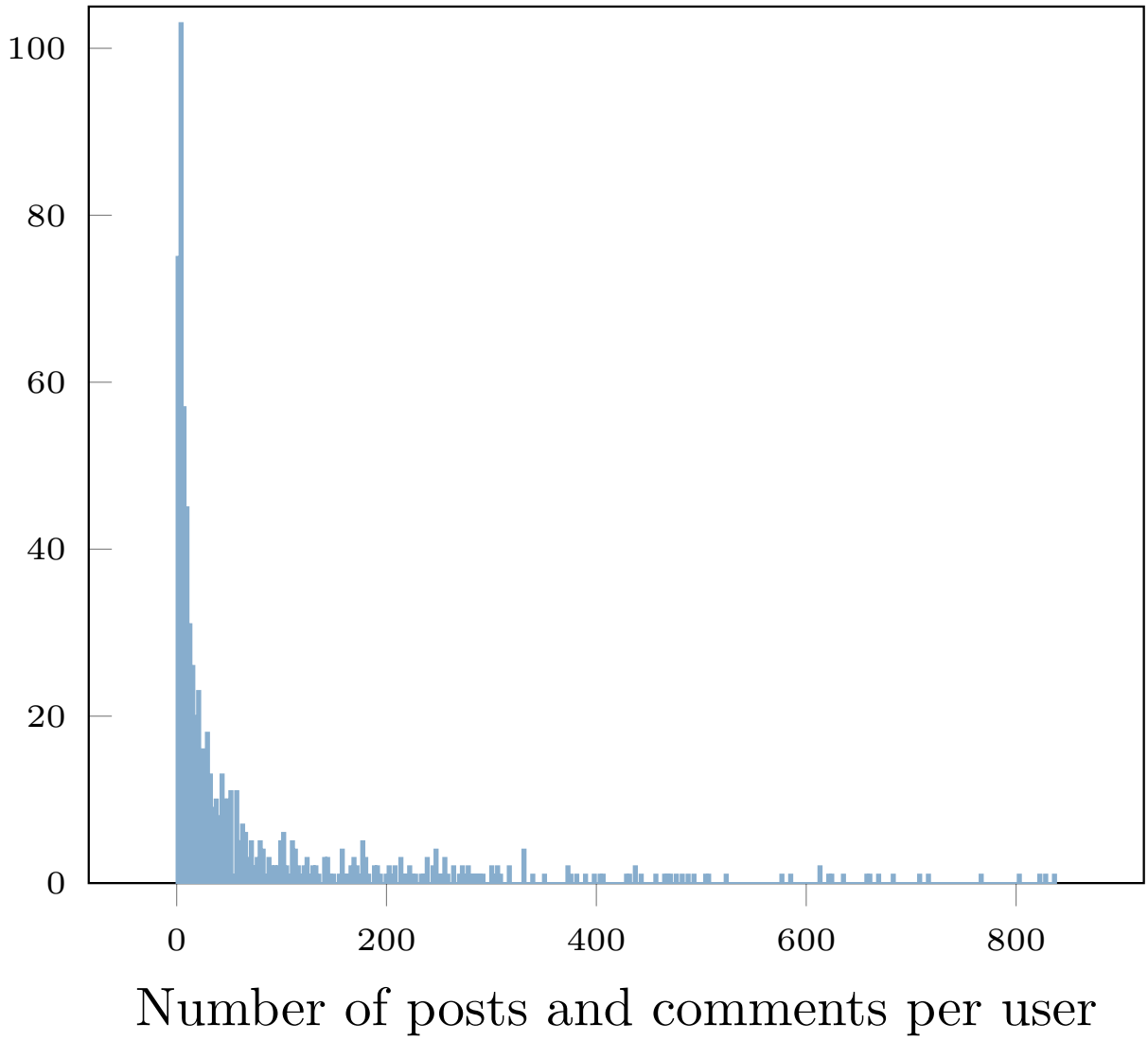



**Figure 4.3.** Histogram of users' posts and comments.

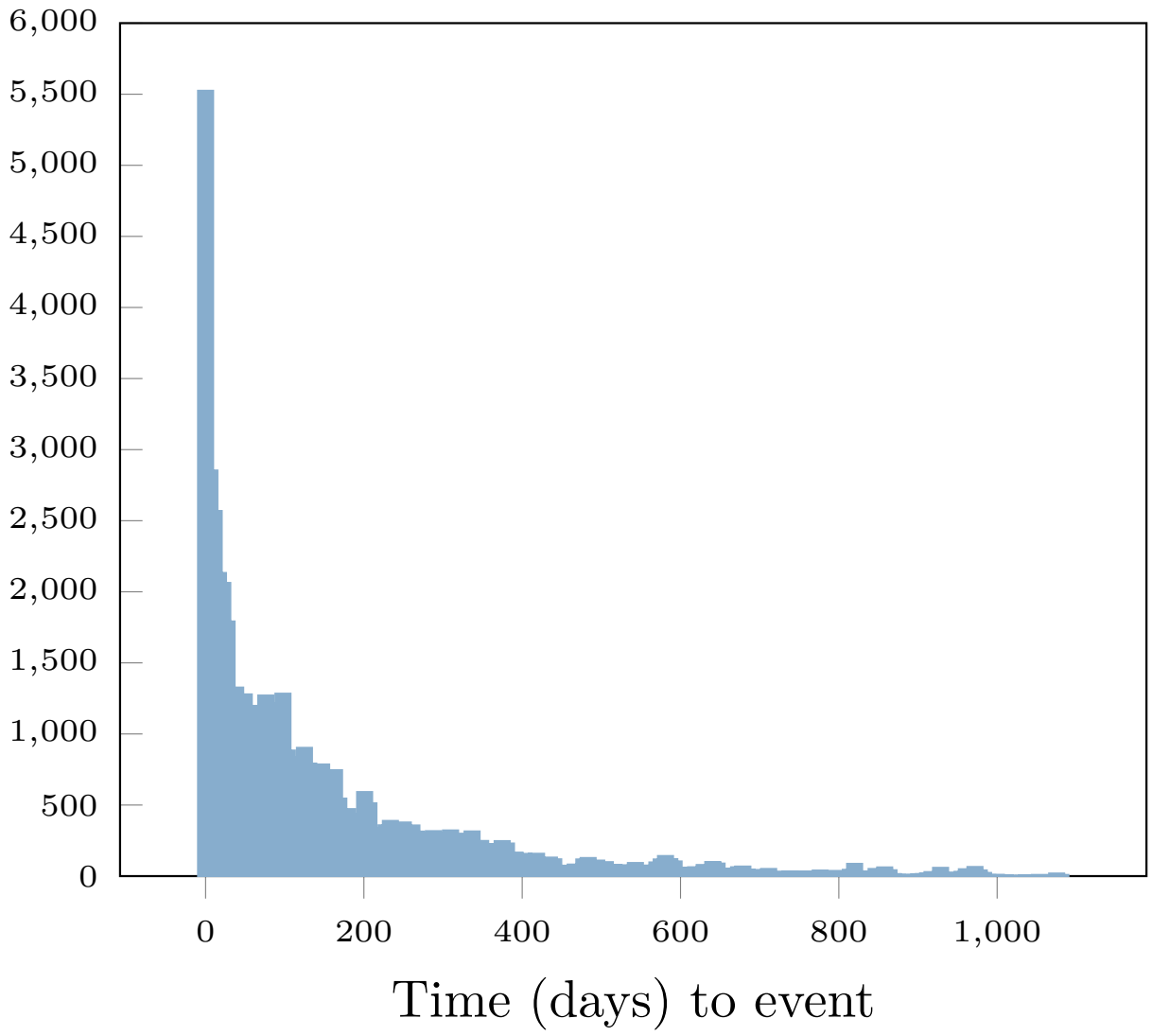



**Figure 4.4.** Histogram of time to event.

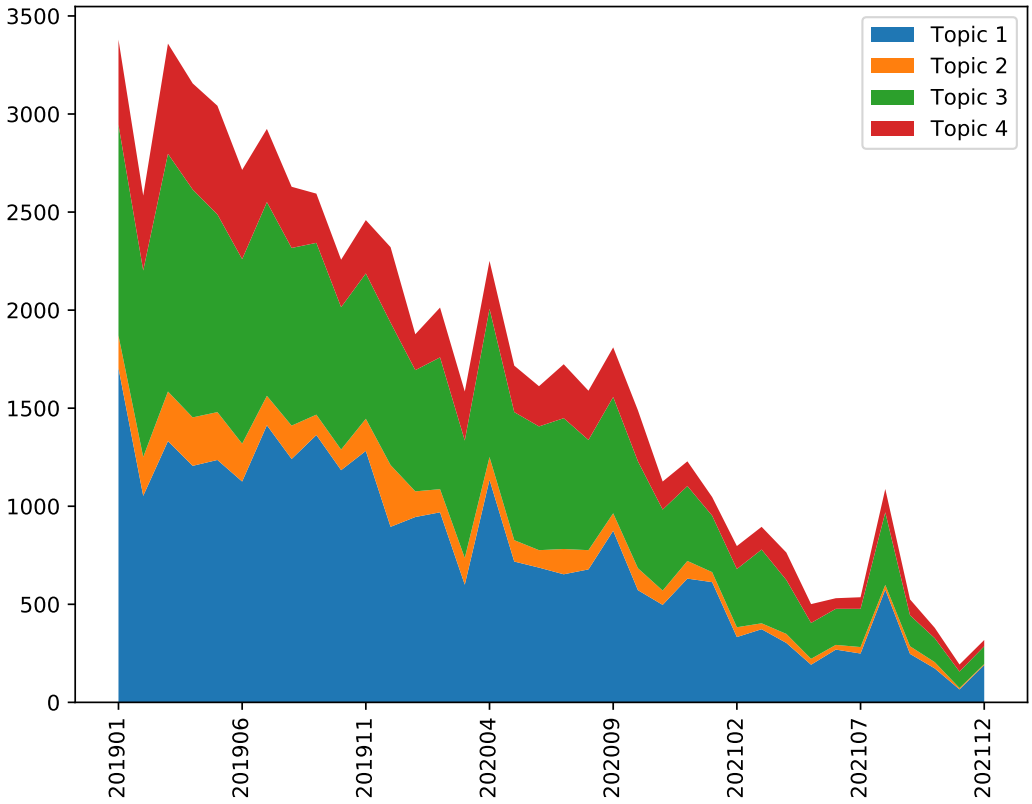



**Figure 4.5.** Popularity of each topic over time (month).

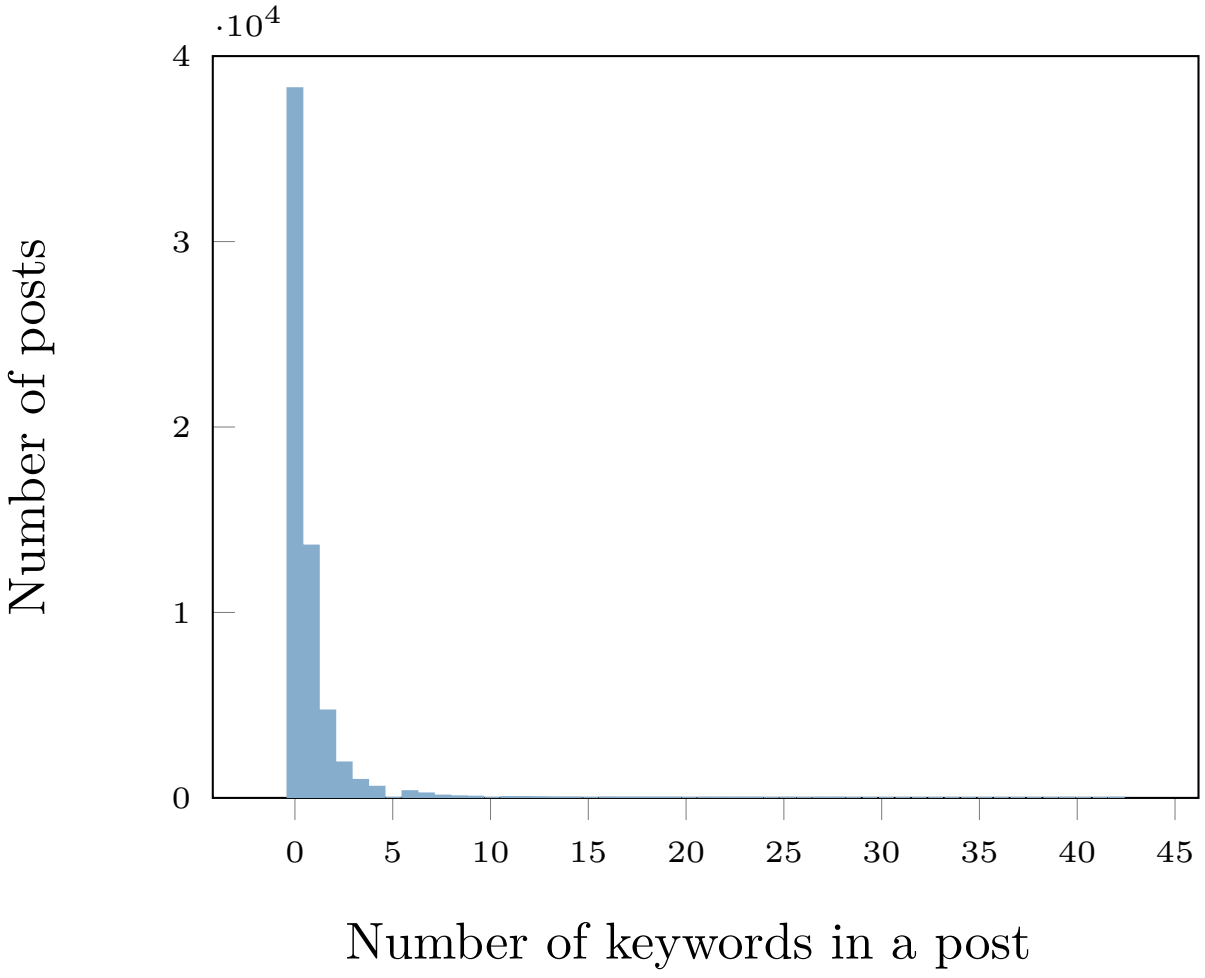



**Figure 4.6.** Histogram of number of keywords.

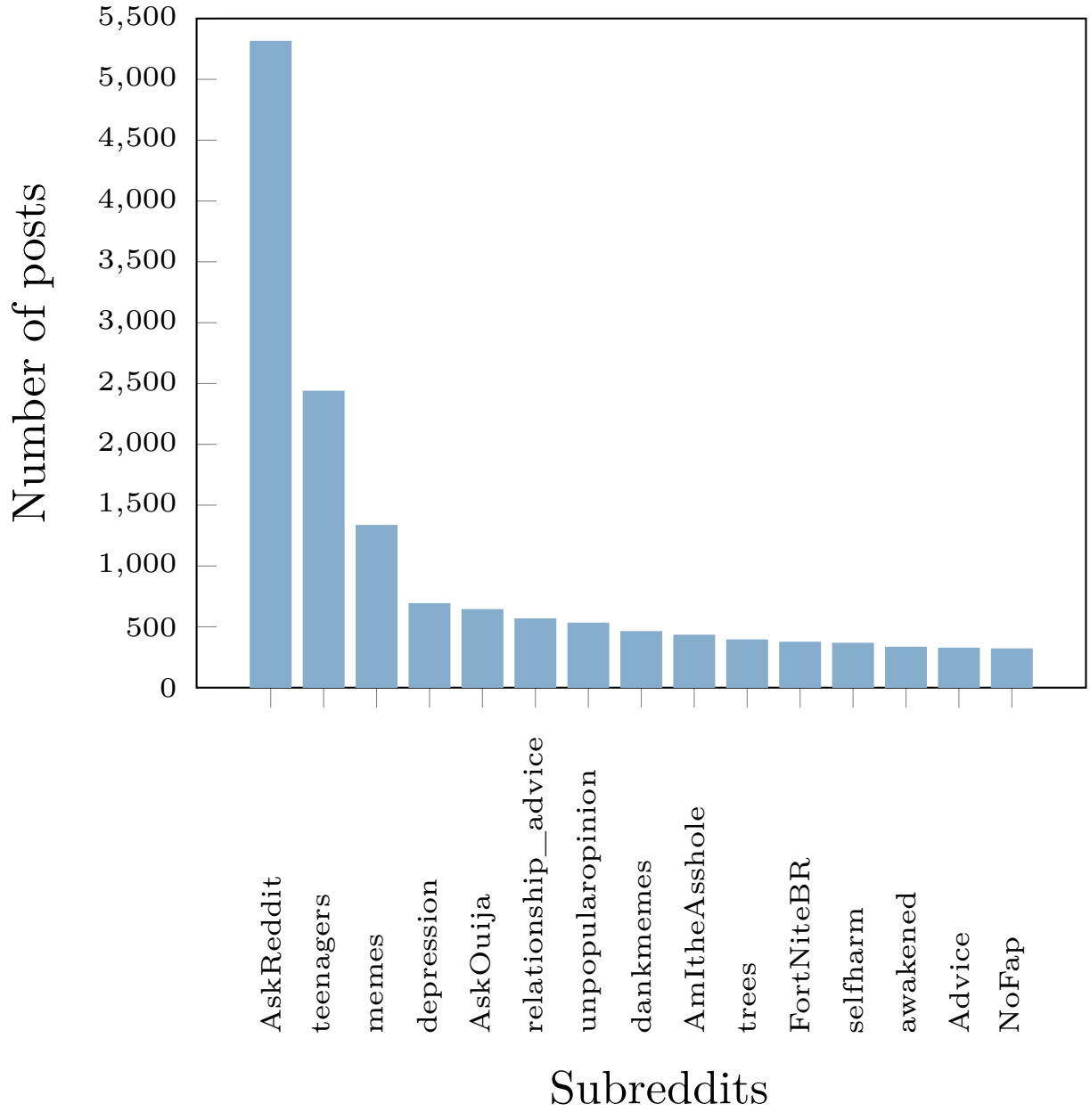



**Figure 4.7.** 15 mostly posted subreddits.

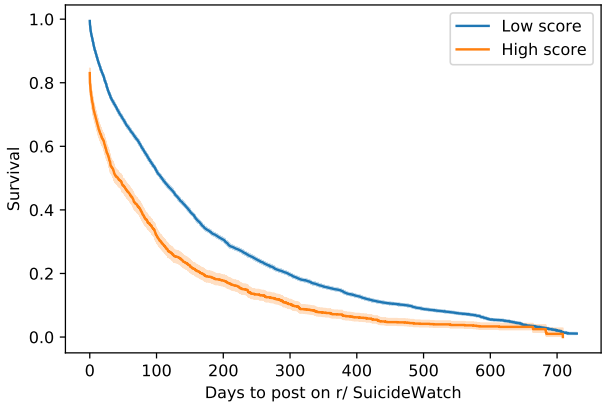
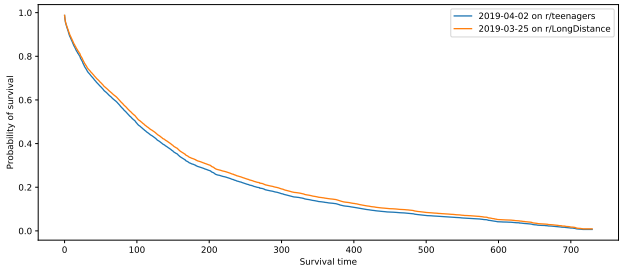
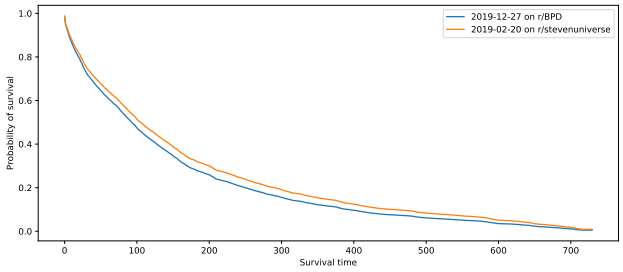


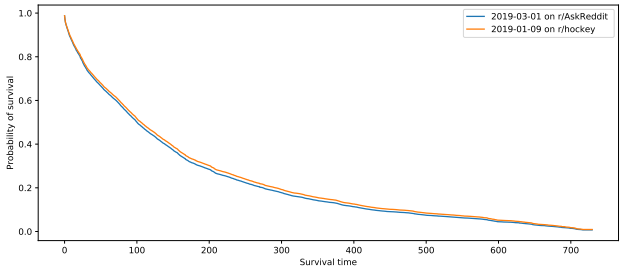

**Figure 4.8.** Kalpan Meier estimates by suicidal score




(a) User 1




(b) User 2




(c) User 3

**Figure 4.9.** Predicted survival curve at time of each post


**Figure 4.10.** Average days from the most recent subreddit transitions to r/ SuicideWatch with a high (low) suicidal score

# 5. SUMMARY AND DISCUSSION

In this thesis, we studied several specialized forms of counting processes. We introduced the concept of counting processes. As a most used form of a counting process, a Poisson process was discussed and extended to a Hawkes process as its background component. We provided intensity functions and likelihood functions of these point processes. We also reviewed the Cox model and how this model could be extended to a model with covariates that relate to the intensity process of counting processes [21].

In Chapter 3, we showed that by integrating heterogeneous datasets with semi-parametric spatial-temporal Hawkes processes, we improved model accuracy to parameters estimation, and simultaneously inferred the missing overdose category for the nonfatal overdose EMS data. In Chapter 4, we collected a large corpus of suicide related posts from r/suicidewatch, along with earlier posts made by users on other subreddits. We then fit a Cox proportional hazards model to predict the time-to-event between earlier posts and later posts on r/suicide- watch. We found statistically significant features using indicators for subreddit, keyword, or suicide risk. Our results indicated potential points of earlier intervention and analysis of associated subreddits to suicide may yield new hypotheses for suicide researchers to investigate.

We see several potential directions for future work. For the semi-parametric Hawkes modeling on drug overdose events in Chapter 3, we could focus on investigating the extent to which drug overdose triggering found in the present study can be detected across cities and model specifications.

We would also like to extend the Cox model in Chapter 4, so that we could take inter-post time as another feature, or model the inter-post time in a Hawkes temporal model for each user, and take the intensity as another feature. A deep learning-based survival model [137] may yield improvements to accuracy. Also, we did not explore the social network of individuals and how interactions on the network may be predictive of transitions to suicide ideation. In future work we hope to analyze posting behavior of network connections and people whom a given person interacts with, and determine how those interactions may act to protect against or lead to higher risk of suicide ideation observed online.

# REFERENCES


[1] *What is an electronic health record (ehr)?* https://www.healthit.gov/faq/what-electronic-health-record-ehr, Accessed: 2010-09-30.

[2] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, "Temporal event sequence simplification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2227–2236, 2013. DOI: 10.1109/TVCG.2013.200.

[3] M. L. Mauriello, B. Shneiderman, F. Du, S. Malik, and C. Plaisant, "Simplifying overviews of temporal event sequences," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '16, San Jose, California, USA: Association for Computing Machinery, 2016, pp. 2217–2224, ISBN: 9781450340823. DOI: 10.1145/2851581.2892440. [Online]. Available: https://doi.org/10.1145/2851581.2892440.

[4] A. Veen and F. P. Schoenberg, "Estimation of space–time branching process models in seismology using an em–type algorithm," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 614–624, 2008. DOI: 10.1198/016214508000000148. eprint: https://doi.org/10.1198/016214508000000148. [Online]. Available: https://doi.org/10.1198/016214508000000148.

[5] R. A. Clements, F. Paik Schoenberg, and D. Schorlemmer, "Residual analysis methods for space–time point processes with applications to earthquake forecast models in California," *arXiv e-prints*, arXiv:1202.6487, arXiv:1202.6487, Feb. 2012. arXiv: 1202.6487 [stat.AP].

[6] M. Heidarysafa, K. Kowsari, L. Barnes, and D. Brown, "Analysis of railway accidents' narratives using deep learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2018, pp. 1446–1453.

[7] B. Green, T. Horel, and A. V. Papachristos, "Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014," *JAMA Internal Medicine*, vol. 177, no. 3, pp. 326–333, Mar. 2017, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2016.8245. eprint: https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2594804/jamainternal_green_2017_oi_160109.pdf. [Online]. Available: https://doi.org/10.1001/jamainternmed.2016.8245.

[8] C. Loeffler and S. Flaxman, "Is gun violence contagious? a spatiotemporal test," *Journal of Quantitative Criminology*, vol. 34, no. 4, pp. 999–1017, 2017. DOI: 10.1007/s10940-017-9363-8.

[9] M. Farajtabar, J. Yang, X. Ye, H. Xu, R. Trivedi, E. Khalil, S. Li, L. Song, and H. Zha, "Fake news mitigation via point process based intervention," in *International conference on machine learning*, PMLR, 2017, pp. 1097–1106.

[10] Y. Li, N. Du, and S. Bengio, "Time-dependent representation for neural event sequence prediction," *ArXiv*, vol. abs/1708.00065, 2018.

[11] M. K. Kumwenda, C. C. Johnson, A. T. Choko, W. Lora, W. Sibande, D. Sakala, P. Indravudh, R. Chilongosi, R. C. Baggaley, R. Nyirenda, M. Taegtmeyer, K. Hatzold, N. Desmond, and E. L. Corbett, "Exploring social harms during distribution of hiv self-testing kits using mixed-methods approaches in malawi," *Journal of the International AIDS Society*, vol. 22, no. S1, e25251, 2019. DOI: https://doi.org/10.1002/jia2.25251. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jia2.25251.

[12] F. P. Schoenberg, M. Hoffmann, and R. J. Harrigan, "A recursive point process model for infectious diseases," *Annals of the Institute of Statistical Mathematics*, vol. 71, no. 5, pp. 1271–1287, Oct. 2019. DOI: 10.1007/s10463-018-0690-9. [Online]. Available: https://ideas.repec.org/a/spr/aistmt/v71y2019i5d10.1007_s10463-018-0690-9.html.

[13] N. Ramakrishnan, S. Tadepalli, L. T. Watson, R. F. Helm, M. Antoniotti, and B. Mishra, "Reverse engineering dynamic temporal models of biological processes and their relationships," *Proceedings of the National Academy of Sciences*, vol. 107, no. 28, pp. 12 511–12 516, 2010. DOI: 10.1073/pnas.1006283107. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1006283107. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1006283107.

[14] O. O. Aalen, Ø. Borgan, and S. Gjessing, *Survival and event history analysisa process point of view*, ser. Statistics for Biology and Health. Springer Science & Business Media, 2008.

[15] P. Olofsson, "Counting process," in. Jan. 2013. DOI: 10.1002/9780470057339.vnn067.

[16] P. Guttorp and T. L. Thorarinsdottir, "What happened to discrete chaos, the quenouille process, and the sharp markov property? some history of stochastic point processes," *International Statistical Review*, vol. 80, no. 2, pp. 253–268, 2012. DOI: https://doi.org/10.1111/j.1751-5823.2012.00181.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2012.00181.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2012.00181.x.

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

[18] P. J. Laub, Y. Lee, and T. Taimre, "The elements of hawkes processes," 2021.

[19] R. Li, X. Bi, and S. Zhang, "Web renewal counting processes and their applications in insurance," *Journal of Inequalities and Applications*, vol. 2018, no. 1, pp. 1–15, 2018.

[20] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972, ISSN: 00359246. [Online]. Available: http://www.jstor.org/stable/2985181.

[21] P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: A large sample study," *The Annals of Statistics*, vol. 10, no. 4, pp. 1100–1120, 1982, ISSN: 00905364. [Online]. Available: http://www.jstor.org/stable/2240714.

[22] A. Calì, D. Lembo, R. Rosati, and M. Ruzzi, "Experimenting data integration with dis@dis," in *Advanced Information Systems Engineering*, A. Persson and J. Stirna, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 51–66, ISBN: 978-3-540-25975-6.

[23] A. Halevy, A. Rajaraman, and J. Ordille, "Data integration: The teenage years," in *Proceedings of the 32nd international conference on Very large data bases*, 2006, pp. 9–16.

[24] P. Geeleher, A. Nath, F. Wang, Z. Zhang, A. N. Barbeira, J. Fessler, R. L. Grossman, C. Seoighe, and R. S. Huang, "Cancer eqtls can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity," *bioRxiv*, 2018. DOI: 10.1101/366922. eprint: https://www.biorxiv.org/content/early/2018/07/10/366922.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2018/07/10/366922.

[25] R. Beaubrun, "Integration of heterogeneous wireless access networks," in *Heterogeneous Wireless Access Networks*, Springer, 2008, pp. 1–18.

[26] L. Wang, "Heterogeneous data and big data analytics," *Automatic Control and Information Sciences*, vol. 3, no. 1, pp. 8–15, 2017, ISSN: 2375-1630. DOI: 10.12691/acis-3-1-3. [Online]. Available: http://pubs.sciepub.com/acis/3/1/3.

[27] B. O. Muthén, "Latent variable modeling in heterogeneous populations," *Psychometrika*, vol. 54, no. 4, pp. 557–585, 1989.

[28] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Embedding heterogeneous data using statistical models," in *Proceedings of The National Conference on Artificial Intelligence*, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, vol. 21, 2006, p. 1605.

[29] K. S. Berlin, N. A. Williams, and G. R. Parra, "An Introduction to Latent Variable Mixture Modeling (Part 1): Overview and Cross-Sectional Latent Class and Latent Profile Analyses," *Journal of Pediatric Psychology*, vol. 39, no. 2, pp. 174–187, Nov. 2013. DOI: 10.1093/jpepsy/jst084. eprint: https://academic.oup.com/jpepsy/article-pdf/39/2/174/14103457/jst084.pdf. [Online]. Available: https://doi.org/10.1093/jpepsy/jst084.

[30] N. Ali, D. Neagu, and P. Trundle, "Classification of heterogeneous data based on data type impact on similarity," in *Advances in Computational Intelligence Systems*, A. Lotfi, H. Bouchachia, A. Gegov, C. Langensiepen, and M. McGinnity, Eds., Cham: Springer International Publishing, 2019, pp. 252–263, ISBN: 978-3-319-97982-3.

[31] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," *Pattern Recognition*, vol. 107, p. 107 501, 2020.

[32] Y. Ogata, K. Katsura, and M. Tanemura, "Modelling heterogeneous space–time occurrences of earthquakes and its residual analysis," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 52, no. 4, pp. 499–509, 2003. DOI: https://doi.org/10.1111/1467-9876.00420. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9876.00420. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00420.

[33] C. R. Bhat, "A new generalized heterogeneous data model (ghdm) to jointly model mixed types of dependent variables," *Transportation Research Part B: Methodological*, vol. 79, pp. 50–77, 2015, ISSN: 0191-2615. DOI: https://doi.org/10.1016/j.trb.2015.05.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261515001198.

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

[35] K.-L. Tsui, S. Y. Wong, W. Jiang, and C.-J. Lin, "Recent research and developments in temporal and spatiotemporal surveillance for public health," *IEEE Transactions on Reliability*, vol. 60, no. 1, pp. 49–58, 2011.

[36] H.-L. Yu, A. Kolovos, G. Christakos, J.-C. Chen, S. Warmerdam, and B. Dev, "Interactive spatiotemporal modelling of health systems: The seks–gui framework," *Stochastic Environmental Research and Risk Assessment*, vol. 21, no. 5, pp. 555–572, 2007.

[37] D. Weisburd, "The law of crime concentration and the criminology of place," *Criminology*, vol. 53, no. 2, pp. 133–157, 2015.

[38] J. Hibdon and E. R. Groff, "What you find depends on where you look: Using emergency medical services call data to target illicit drug use hot spots," *Journal of contemporary criminal justice*, vol. 30, no. 2, pp. 169–185, 2014.

[39] J. Hibdon, C. W. Telep, and E. R. Groff, "The concentration and stability of drug activity in seattle, washington using police and emergency medical services data," *Journal of quantitative criminology*, vol. 33, no. 3, pp. 497–517, 2017.

[40] J. G. Carter, G. Mohler, and B. Ray, "Spatial concentration of opioid overdose deaths in indianapolis: An application of the law of crime concentration at place to a public health epidemic," *Journal of Contemporary Criminal Justice*, p. 1 043 986 218 803 527, 2018.

[41] M. Townsley, R. Homel, and J. Chaseling, "Infectious burglaries. a test of the near repeat hypothesis," *British Journal of Criminology*, vol. 43, no. 3, pp. 615–633, 2003.

[42] A. M. Zeoli, J. M. Pizarro, S. C. Grady, and C. Melde, "Homicide as infectious disease: Using public health methods to investigate the diffusion of homicide," *Justice quarterly*, vol. 31, no. 3, pp. 609–632, 2014.

[43] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham, "Randomized controlled field trials of predictive policing," *Journal of the American statistical association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[44] *https://www.cdc.gov/mmwr/volumes/70/wr/mm7008a1.htm*.

[45] A. Mbarek, S. Jamoussi, A. Charfi, and A. B. Hamadou, "Suicidal profiles detection in twitter," in *15th International Conference on Web Information Systems and Technologies*, Jan. 2019.

[46] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015, ISSN: 2214-7829. DOI: https://doi.org/10.1016/j.invent.2015.03.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214782915000160.

[47] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, vol. 2018, pp. 1–10, 2018. DOI: https://doi.org/10.1155/2018/6157249.

[48] A. Nobles, J. J. Glenn, K. Kowsari, B. Teachman, and L. E. Barnes, "Identification of imminent suicide risk among young adults using text messages," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[49] A. M. May, E. K. Czyz, and B. T. West, "Differentiating adolescent suicide attempters and ideators: A classification tree analysis of risk behaviors," *Journal of Adolescent Health*, vol. 67, no. 6, pp. 837–850, 2020, ISSN: 1054-139X. DOI: https://doi.org/10.1016/j.jadohealth.2020.04.018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1054139X20302056.

[50] Y. Xiao, M. Romanelli, and M. Lindsey, "A latent class analysis of health lifestyles and suicidal behaviors among us adolescents," English (US), *Journal of Affective Disorders*, vol. 255, pp. 116–126, Aug. 2019, Publisher Copyright: © 2019 Elsevier B.V., ISSN: 0165-0327. DOI: 10.1016/j.jad.2019.05.031.

[51] Y. Ophir, R. Tikochinski, C. S. C. Asterhan, I. Sisso, and R. Reichart, "Deep neural networks detect suicide risk from textual facebook posts," *Scientific Reports*, no. 1, p. 16 685, 2020. DOI: 10.1038/s41598-020-73917-0. [Online]. Available: https://doi.org/10.1038/s41598-020-73917-0.

[52] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," *Algorithms*, vol. 13, no. 1, 2020, ISSN: 1999-4893. DOI: 10.3390/a13010007. [Online]. Available: https://www.mdpi.com/1999-4893/13/1/7.

[53] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, P. Limbachiya, S. C. Guntuku, and H. A. Schwartz, "Suicide risk assessment with multi-level dual-context language and BERT," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 39–44. DOI: 10.18653/v1/W19-3005. [Online]. Available: https://www.aclweb.org/anthology/W19-3005.

[54] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, 2021. DOI: 10.1109/TCSS.2020.3021467.

[55] J. Lopez-Castroman, B. Moulahi, J. Azé, S. Bringay, J. Deninotti, S. Guillaume, and E. Baca-Garcia, "Mining social networks to improve suicide prevention: A scoping review," *Journal of neuroscience research*, vol. 98, no. 4, pp. 616–625, Apr. 2020, ISSN: 0360-4012. DOI: 10.1002/jnr.24404. [Online]. Available: https://doi.org/10.1002/jnr.24404.

[56] B. Silveira Fraga, A. P. Couto da Silva, and F. Murai, "Online social networks in health care: A study of mental disorders on reddit," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018, pp. 568–573. DOI: 10.1109/WI.2018.00-36.

[57] A. E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, "Detecting suicidal ideation on forums: Proof-of-concept study," *J Med Internet Res*, vol. 20, no. 6, e215, Jun. 2018, ISSN: 1438-8871. DOI: 10.2196/jmir.9840.

[58] A. Ruch, "Can x2vec save lives? integrating graph and language embeddings for automatic mental health classification," *Journal of Physics: Complexity*, no. 3, 2020.

[59] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study," *J Med Internet Res*, vol. 22, no. 10, e22635, Oct. 2020, ISSN: 1438-8871. DOI: 10.2196/22635. [Online]. Available: http://www.jmir.org/2020/10/e22635/.

[60] A. Alambo, M. Gaur, U. Lokala, U. Kursuncu, K. Thirunarayan, A. Gyrard, A. Sheth, R. S. Welton, and J. Pathak, "Question answering for suicide risk assessment using reddit," in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 468–473. DOI: 10.1109/ICOSC.2019.8665525.

[61] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead, "CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 24–33. DOI: 10.18653/v1/W19-3003. [Online]. Available: https://www.aclweb.org/anthology/W19-3003.

[62] R. Dutta, G. Gkotsis, S. Velupillai, I. Bakolis, and R. Stewart, "Temporal and diurnal variation in social media posts to a suicide support forum," *BMC Psychiatry*, no. 1, p. 259, 2021. DOI: 10.1186/s12888-021-03268-1. [Online]. Available: https://doi.org/10.1186/s12888-021-03268-1.

[63] F. P. Schoenberg, "Introduction to point processes," in *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Ltd, 2011, ISBN: 9780470400531. DOI: https://doi.org/10.1002/9780470400531.eorms0425. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470400531.eorms0425. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470400531.eorms0425.

[64] U. Eden, *Chapter 2: Introduction to Point Processes*, http://www.stat.columbia.edu/~liam/teaching/neurostat-fall17/uri-eden-point-process-notes.pdf, 2017.

[65] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*. New York: Springer, 2003.

[66] J. Kingman, *Poisson Processes*, ser. Oxford Studies in Probability. Clarendon Press, 1992, ISBN: 9780191591242. [Online]. Available: https://books.google.com/books?id=VEiM-OtwDHkC.

[67] Y. Xu, S. Kim, S. M. Salapaka, C. L. Beck, and T. P. Coleman, "Clustering large networks of parametric dynamic generative models," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, IEEE, 2012, pp. 5248–5253.

[68] J. G. Rasmussen, "Lecture notes: Temporal point processes and the conditional intensity function," *arXiv preprint arXiv:1806.00221*, 2018.

[69] F. P. Schoenberg, "Introduction to point processes," *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

[70] E. Lewis and G. Mohler, "A nonparametric em algorithm for multiscale hawkes processes," *Journal of Nonparametric Statistics*, vol. 1, no. 1, pp. 1–20, 2011.

[71] J. G. Rasmussen, "Bayesian inference for hawkes processes," *Methodology and Computing in Applied Probability*, vol. 15, no. 3, pp. 623–642, 2013.

[72] R. C. Lambert, C. Tuleau-Malot, T. Bessaih, V. Rivoirard, Y. Bouret, N. Leresche, and P. Reynaud-Bouret, "Reconstructing the functional connectivity of multiple spike trains using hawkes models," *Journal of Neuroscience Methods*, vol. 297, pp. 9–21, 2018, ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2017.12.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027017304442.

[73] A. Bonnet, C. Dion-Blanc, F. Gindraud, and S. Lemler, "Neuronal network inference and membrane potential model using multivariate hawkes processes," *Journal of Neuroscience Methods*, vol. 372, p. 109 550, 2022, ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2022.109550. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027022000772.

[74] G. Mohler, M. B. Short, F. Schoenberg, and D. Sledge, "Analyzing the impacts of public policy on covid-19 transmission: A case study of the role of model and dataset selection using data from indiana," *Statistics and Public Policy*, vol. 8, no. 1, pp. 1–8, 2021.

[75] W.-H. Chiang, X. Liu, and G. Mohler, "Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates," *International journal of forecasting*, vol. 38, no. 2, pp. 505–520, 2022.

[76] J. Zhuang and J. Mateu, "A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Dec. 2018. DOI: 10.1111/rssa.12429.

[77] P. Embrechts, T. Liniger, and L. Lin, "Multivariate hawkes processes: An application to financial data," *Journal of Applied Probability*, vol. 48, no. A, pp. 367–378, 2011. DOI: 10.1239/jap/1318940477.

[78] A. G. Hawkes, "Hawkes processes and their applications to finance: A review," *Quantitative Finance*, vol. 18, no. 2, pp. 193–198, 2018.

[79] A. Reinhart, "A review of self-exciting spatio-temporal point processes and their applications," *Statistical Science*, vol. 33, no. 3, pp. 299–318, 2018.

[80] T. Utsu, "A statistical study on the occurrence of aftershocks.," *Geophysical magazine*, vol. 30, pp. 521–605, 1961.

[81] Y. Ogata, "Statistical models for earthquake occurrences and residual analysis for point processes," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 9–27, 1988. DOI: 10.1080/01621459.1988.10478560. eprint: https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478560. [Online]. Available: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1988.10478560.

[82] G. Mohler, "Marked point process hotspot maps for homicide and gun crime prediction in chicago," *International Journal of Forecasting*, vol. 30, no. 3, pp. 491–497, 2014.

[83] B. Yuan, H. Li, A. L. Bertozzi, P. J. Brantingham, and M. A. Porter, "Multivariate spatiotemporal hawkes processes and network reconstruction," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 2, pp. 356–382, 2019.

[84] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 100–108, 2011. DOI: 10.1198/jasa.2011.ap09546. eprint: https://doi.org/10.1198/jasa.2011.ap09546. [Online]. Available: https://doi.org/10.1198/jasa.2011.ap09546.

[85] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[86] V. Bewick, L. Cheek, and J. Ball, "Statistics review 12: Survival analysis," *Critical care*, vol. 8, no. 5, pp. 1–6, 2004.

[87] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-meier estimate," *International journal of Ayurveda research*, vol. 1, no. 4, p. 274, 2010.

[88] D. Cox and D. Oakes, *Analysis of Survival Data (1st ed.)* Ser. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 1984. [Online]. Available: https://doi.org/10.1201/9781315137438.

[89] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[90] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part i: Basic concepts and first analyses," *British journal of cancer*, vol. 89, no. 2, pp. 232–238, 2003.

[91] W. Nelson, "Hazard plotting for incomplete failure data," *Journal of Quality Technology*, vol. 1, no. 1, pp. 27–52, 1969.

[92] W. Nelson, "Theory and applications of hazard plotting for censored failure data," *Technometrics*, vol. 14, no. 4, pp. 945–966, 1972.

[93] O. O. Aalen, *Statistical inference for a family of counting processes.* University of California, Berkeley, 1975.

[94] K. Lasslett, "Crime or social harm? a dialectical perspective," *Crime, Law and Social Change*, vol. 54, pp. 1–19, 2010.

[95] M. Innes and C. Leigh, *Mapping and measuring the social harms of crime and anti-social behaviour: Toward an outcomes-based approach to community safety in wales*, 2011. [Online]. Available: https://gov.wales/sites/default/files/statistics-and-research/2019-08/130121-mapping-measuring-social-harms-crime-anti-social-behaviour-en.pdf.

[96] *Prevention strategies*, https://www.cdc.gov/suicide/prevention/index.html, 2008.

[97] A. FB, C. JA, R. LM, and S. P, *Provisional drug overdose death counts*, https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm, 2022.

[98] D. Ciccarone, "Fentanyl in the us heroin supply: A rapidly changing risk environment," *International Journal of Drug Policy*, vol. 46, pp. 107–111, 2017.

[99] T. J. Cicero, M. S. Ellis, H. L. Surratt, and S. P. Kurtz, "The changing face of heroin use in the united states: A retrospective analysis of the past 50 years," *JAMA psychiatry*, vol. 71, no. 7, pp. 821–826, 2014.

[100] R. A. Rudd, L. J. Paulozzi, M. J. Bauer, R. W. Burleson, R. E. Carlson, D. Dao, J. W. Davis, J. Dudek, B. A. Eichler, J. C. Fernandes, *et al.*, "Increases in heroin overdose deaths — 28 states, 2010 to 2012," *MMWR. Morbidity and mortality weekly report*, vol. 63, no. 39, p. 849, 2014.

[101] G. K. Strickler, K. Zhang, J. M. Halpin, A. S. Bohnert, G. Baldwin, and P. W. Kreiner, "Effects of mandatory prescription drug monitoring program (pdmp) use laws on prescriber registration and use and on risky prescribing," *Drug and Alcohol Dependence*, 2019.

[102] R. M. Gladden, "Fentanyl law enforcement submissions and increases in synthetic opioid-involved overdose deaths - 27 states, 2013–2014," *MMWR. Morbidity and mortality weekly report*, vol. 65, 2016.

[103] C. M. Jones, E. B. Einstein, and W. M. Compton, "Changes in synthetic opioid involvement in drug overdose deaths in the united states, 2010-2016," *Jama*, vol. 319, no. 17, pp. 1819–1821, 2018.

[104] D. B. Kandel, M.-C. Hu, P. Griesler, and M. Wall, "Increases from 2002 to 2015 in prescription opioid overdose deaths in combination with other substances," *Drug and alcohol dependence*, vol. 178, pp. 501–511, 2017.

[105] C. McCall Jones, G. T. Baldwin, and W. M. Compton, "Recent increases in cocaine-related overdose deaths and the role of opioids," *American journal of public health*, vol. 107, no. 3, pp. 430–432, 2017.

[106] *Drug overdose deaths in the u.s. top 100,000 annually*, https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2021/20211117.htm#:~:text=For%20Immediate%20Release%3A%20November%2017%2C%202021&text=The%20new%20data%20documents%20that,from%2056%2C064%20the%20year%20before., 2021.

[107] D. C. Daley, "Family and social aspects of substance use disorders and treatment," *Journal of food and drug analysis*, vol. 21, no. 4, S73–S76, 2013.

[108] J. P. Smith, *The social impact of drug abuse*, https://www.unodc.org/pdf/technical_series_1995-03-01_1.pdf, 1995.

[109] *Preventing suicide*, https://www.cdc.gov/suicide/facts/, 2022.

[110] *Cdc wonder: Underlying cause of death, 1999–2019*, https://wonder.cdc.gov/Deaths-by-Underlying-Cause.html, 2020.

[111] *Community health and program services: Health disparities among racial/ethnic populations*, https://www.cdc.gov/minorityhealth/CHDIReport.html, 2008.

[112] E. D. Klonsky, A. M. May, B. Y. Saffer, *et al.*, "Suicide, suicide attempts, and suicidal ideation," *Annu Rev Clin Psychol*, vol. 12, no. 1, pp. 307–30, 2016.

[113] P. Seth, L. Scholl, R. A. Rudd, and S. Bacon, "Overdose deaths involving opioids, cocaine, and psychostimulants — united states, 2015–2016," *Morbidity and Mortality Weekly Report*, vol. 67, no. 12, p. 349, 2018.

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

[115] J. Zhuang, Y. Ogata, and D. Vere-Jones, "Stochastic declustering of space-time earthquake occurrences," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 369–380, 2002, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/3085650.

[116] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[117] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, 2012, pp. 952–961.

[118] B. R. Ray, E. M. Lowder, A. J. Kivisto, P. Phalen, and H. Gil, "Ems naloxone administration as non-fatal opioid overdose surveillance: 6-year outcomes in marion county, indiana," *Addiction*, vol. 113, no. 12, pp. 2271–2279, 2018.

[119] R. Fisher, D. O'Donnell, B. Ray, and D. Rusyniak, "Police officers can safely and effectively administer intranasal naloxone," *Prehospital Emergency Care*, vol. 20, no. 6, pp. 675–680, 2016.

[120] R. J. Sampson and W. B. Groves, "Community structure and crime: Testing social-disorganization theory," *American journal of sociology*, vol. 94, no. 4, pp. 774–802, 1989.

[121] M. J. Alexander, M. V. Kiang, and M. Barbieri, "Trends in black and white opioid mortality in the united states, 1979–2015," *Epidemiology (Cambridge, Mass.)*, vol. 29, no. 5, p. 707, 2018.

[122] H. Jalal, J. M. Buchanich, M. S. Roberts, L. C. Balmert, K. Zhang, and D. S. Burke, "Changing dynamics of the drug overdose epidemic in the united states from 1979 through 2016," *Science*, vol. 361, no. 6408, eaau1184, 2018.

[123] S. H. Meghani, E. Byun, and R. M. Gallagher, "Time to take stock: A meta-analysis and systematic review of analgesic treatment disparities for pain in the united states," *Pain Medicine*, vol. 13, no. 2, pp. 150–174, 2012.

[124] S. G. Mars, P. Bourgois, G. Karandinos, F. Montero, and D. Ciccarone, ""every 'never'i ever said came true": Transitions from opioid pills to heroin injecting," *International Journal of Drug Policy*, vol. 25, no. 2, pp. 257–266, 2014.

[125] D. Kuang, P. J. Brantingham, and A. L. Bertozzi, "Crime topic modeling," *Crime Science*, vol. 6, no. 1, p. 12, 2017.

[126] R. Pandey and G. O. Mohler, "Evaluation of crime topic models: Topic coherence vs spatial crime concentration," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, IEEE, 2018, pp. 76–78.

[127] G. Mohler and P. J. Brantingham, "Privacy preserving, crowd sourced crime hawkes processes," in *2018 International Workshop on Social Sensing (SocialSens)*, IEEE, 2018, pp. 14–19.

[128] G. Mohler *et al.*, "Modeling and estimation of multi-source clustering in crime and security data," *The Annals of Applied Statistics*, vol. 7, no. 3, pp. 1525–1539, 2013.

[129] J. Klein and M. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[130] G. Rodríguez, "Survival models," in *Lecture Notes on Generalized Linear Models*, 2007. [Online]. Available: https://data.princeton.edu/wws509/notes/c7.pdf.

[131] J. Lu, S. Sridhar, R. Pandey, M. A. Hasan, and G. Mohler, "Investigate transitions into drug addiction through text mining of reddit data," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2367–2375.

[132] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression," *arXiv preprint arXiv:1907.00825*, 2019.

[133] "Discussion on professor cox's paper," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 202–220, 1972. DOI: https://doi.org/10.1111/j.2517-6161.1972.tb00900.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00900.x. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00900.x.

[134] A. Singh, *Suicidal thought detection*, https://www.kaggle.com/abhijitsingh001/suicidal-thought-detection/notebook.

[135] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[136] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," English, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, May 2010, pp. 45–50.

[137] Y. Zhao, Q. Hong, X. Zhang, Y. Deng, Y. Wang, and L. Petzold, "Bertsurv: Bert-based survival models for predicting outcomes of trauma patients," *arXiv preprint arXiv:2103.10928*, 2021.

# VITA

Xueying Liu received her B.Sc. in Mathematics and Statistics from University of Edinburgh in July, 2014. She then received a M.S. in Statistics from George Washington University in May, 2016. She joined the Department of Computer and Information Science of Indiana University Purdue University Indianapolis in August, 2017 to pursue her Ph.D. degree in computer science under the supervision of Dr. George Mohler. Her research focused on statistical and machine learning approaches to solving problems in time series prediction, natural language processing, healthcare and social harms.