Anytime User Engagement Prediction in Information Cascades for Arbitrary Observation Periods

Akshay Aravamudan , Xi Zhang, Georgios C. Anagnostopoulos

Department of Computer Engineering & Sciences, Florida Institute of Technology, Melbourne, FL, USA.
aaravamudan2014@my.fit.edu, zhang2012@my.fit.edu, georgio@fit.edu

Abstract

Predicting user engagement – whether a user will engage in a given information cascade – is an important problem in the context of social media, as it is useful to online marketing and misinformation mitigation just to name a few major applications. Based on split population multi-variate survival processes, we develop a discriminative approach that, unlike prior works, leads to a single model for predicting whether individual users of an information network will engage a given cascade for arbitrary forecast horizons and observation periods. Being probabilistic in nature, this model retains the interpretability of its generative counterpart and renders count prediction intervals in a disciplined manner. Our results indicate that our model is highly competitive, if not superior, to current approaches, when compared over varying observed cascade histories and forecast horizons.

Introduction

As of late, the study of information diffusion across the Internet has been an active field of research. An *information cascade* – the propagation trace of a piece of information shared among users/agents of a communication network – forms as a result of users engaging a particular piece of content. An important problem is that of predicting whether or not a user will engage a given information cascade. User engagement prediction garners benefits to those who have a vested interest in knowing whether content will become popular/viral over time, especially within the realm of social media. A few other examples include marketing companies promoting product adoption (Bei, Chen, and Linchi 2011), political campaigns leveraging public opinion (Farajtabar et al. 2016), and even social media companies engaged in content moderation and rumor control (Chen et al. 2022; Farajtabar et al. 2017).

While the terms *information diffusion prediction* and *user engagement prediction* fall under the broader umbrella of *(content) popularity prediction*, there have been various interpretations within existing works. In this work, we refer to the latter as the task of predicting the number of new users engaging a specified information cascade – by reacting to or resharing content at least once – for a given observation period and forecast horizon. This is functionally distinct from works that seek to identify the timings and identity of the next user to engage the cascade such as (Islam et al. 2018; Yang et al. 2018; Cao et al. 2017; Lamprier 2019).

Popularity prediction in the literature has primarily been approached via the use of a generative or a discriminative objective (Zhou et al. 2021). Generative works model popularity using temporal point processes. This includes either using a univariate temporal point process to represent the dynamics of the process (by disregarding user identities) (Chen and Tan 2018) or a multivariate point process (Gomez-Rodriguez, Balduzzi, and Schölkopf 2011; Gomez-Rodriguez, Leskovec, and Krause 2012; Farajtabar et al. 2015) that takes into consideration complex user-level interactions. The benefit of such an approach is that it provides interpretable models of the underlying information diffusion mechanism and, at the same time, enabling us to produce cascade size counts for varying prediction time intervals and/or forecast horizons in a principled way. However, they oftentimes provide lackluster performance on account of generative models not being trained for prediction (Mishra, Rizoiu, and Xie 2016; Zhou et al. 2021; Cao et al. 2017).

On the other hand, works with discriminative/predictive objectives aim solely to produce counts and may not necessarily be interested in discovering the underlying dynamics of the process. While some works do attempt to imbue point process based assumptions (Cao et al. 2017), most of them extract handcrafted features from information cascades for prediction. Recent developments in the realm of recurrent and graph neural networks have facilitated the development of popularity prediction models that consider user interactions as well. In specific, the successes of deep learning models in processing multi-modal information such as graph structure, user, and topic information have shown great promise in improving performance (Wang, Zhou, and Kong 2020). A shortcoming of these models is that, while they allow for user-level analysis and produce satisfactory results, they often require training a model per observation period $[0, t_c]$ and/or forecast horizon $(t_c, \Delta t]$ while not being easily interpretable.

In this work, we bridge the divide by developing a single user-level model across all censoring times and forecast horizons called DANTE (**D**iscriminative probabilistic **AN**ytime user **E**ngagement prediction). We continue to use

point processes due to their rich properties, however, we do so under a lens of a discriminative approach. We develop a discriminative model for user engagement by using split population multi-variate survival processes. This was derived from its generative counterpart found in multivariate survival process (Gomez-Rodriguez, Balduzzi, and Schölkopf 2011) and tweaked to model the sequence of prediction probabilities rather than event occurrences. This helps us highlight user engagement within an information cascade for arbitrary observation periods and forecast horizons from a single trained model. We use a split population assumption for a multivariate survival processes in order to relax the assumption of survival processes that all users eventually manifest an event, which can prove unrealistic for real-world settings. Importantly, since our model is derived from a generative setting and is highly interpretable, we can provide prediction intervals for the number of users that will engage online conversations. We do not make any assumptions about the nature of user-features/platform and our model is therefore able to accommodate various kinds of information networks. The contributions of our work are listed below.

- We provide a single model to predict user engagement in an information cascade for all censoring times t_c and forecast horizons Δt.
- We show with our synthetic experiments that our probabilistic discriminative approach as well as the split population formulation yield benefits over traditional generative modeling of multi-variate survival processes, especially for larger values of t_c.
- We show via real world experiments that our single model performs competitively, if not superior to models who, at the least, require training per observation time t_c.
- We provide prediction intervals for the number of users that engage in an information cascade for varying observation times and forecast horizons.

The rest of this paper is organized as follows. In the next section we describe works related to user engagement. In Preliminaries we provide the reader with some background about survival processes. Novel Formulation describes our modeling framework called DANTE. Finally, in the later sections, we detail our datasets, experimental methodology and comment on results.

Related Work

User engagement prediction, as we have framed it, can quite straightforwardly be cast as a cascade size prediction problem. We do note that among the following works, our treatment of user engagement wherein a user appears only once in a cascade may not always be true. Nevertheless, they are relevant to our task since user-engagement prediction can be treated as a cascade size prediction problem by only considering the first instance of each user. Cascade size prediction tasks can be roughly categorized into two groups.

Macro-level works: This group of works generally is not concerned with engagement of an individual user in a cascade. They might adopt user features such as number of followers, gender, participating communities, etc. as inputs. However, they do not try to predict whether or not individual users will participate in a cascade. Within this group, we can further categorize the approaches into generative or discriminative models. Point process based generative models, such as (Shen et al. 2014; Zhao et al. 2015; Chen and Tan 2018; Tan and Chen 2021; Zhang, Aravamudan, and Anagnostopoulos 2022), model a cascade's information diffusion process first by specifying intensity functions of the process. Then, the conditional mean or median of the counting process is typically adopted as an estimate for cascade size predictions. Additionally, these works are geared towards specific social networks (such as Twitter) and make use of features such as user follower counts which may not always be available for all social media datasets. On the other hand, the discriminative models directly predict cascade size with either hand-crafted features (Cheng et al. 2014; Martin et al. 2016) or features learned from deep network models (Cao et al. 2017; Chen et al. 2019; Li et al. 2017; Xu et al. 2021).

User-level works: This group of works explicitly consider individual users in the construction of models. To predict the cascade size, such an approach starts by forecasting each user's engagement of a cascade, and then aggregates them to find the overall cascade size. The majority of these works adopt a generative approach, which model the behavior of individual users (Gomez-Rodriguez, Leskovec, and Krause 2012; Gomez-Rodriguez, Balduzzi, and Schölkopf 2011; Myers and Leskovec 2010; Kempe, Kleinberg, and Tardos 2003; Yang and Zha 2013) by considering interactive behaviors among each other, such as an influencer driving other users to participate. (Yu et al. 2017) proposed a multi-variate survival process model, where the dynamic process of an infected node's neighbors getting infected by a cascade is modeled as a survival process, and their model NEWER is proposed for predicting cascading processes by effectively aggregating these behavioral dynamics. Recently, (Yang et al. 2019) proposed a deep learning discriminative model which directly predicts individual user engagement and additionally predicts the cascade size with the same model via a reinforcement learning objective.

Preliminaries

To address our problem setting, we will employ a special type of multi-variate *survival processes*. In this section, we provide some basic, relevant background. The interested reader may want to consult the textbook of (Aalen, Borgan, and Gjessing 2008) for more in-depth coverage of this material. In what follows, $[\![\cdot]\!]$ stands for the Iverson bracket, which evaluates to 1, if its argument is true, and to 0, if otherwise.

A *survival process* that commences at time $t_o \in \mathbb{R}$ is a stationary temporal point process with conditional intensity

$$\lambda(t \mid \mathcal{H}_t) \triangleq \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}\{N(t+\Delta t) - N(t) \mid \mathcal{H}_t\}}{\Delta t} = \\ = [\![N(t) = 0]\!]\, h_e(t \mid \mathcal{H}_t) \quad (1)$$

Note that a temporal process is uniquely specified by its conditional intensity. Above, $N(\cdot)$ is the associated counting process, which counts the number of events generated

by the process up to a specified time. It holds that $N(t) = 0$ a.s. for $t < t_o$ and $N(t) \leq 1$ a.s. for $t \geq t_o$. Also, $\mathcal{H}_t$ refers to the process' *history*, *i.e.*, the set of events that have occurred by time t. For example, when $N(t) = 0$, $\mathcal{H}_t = \{t_o\}$ for $t \geq t_o$. However, as we shall see later in the multi-variate setting, $\mathcal{H}_t$ may include additional events that are external to the process. Finally, the non-negative function $h_e(\cdot \mid \mathcal{H}_t)$ is referred to as the process' *hazard rate* and is defined as

$$h_e(t \mid \mathcal{H}_t) \triangleq \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}\{T_e \leq t + \Delta t \mid T_e > t, \mathcal{H}_t\}}{\Delta t} \tag{2}$$

where $T_e \geq 0$ a.s. is the Random Variable (RV) representing the process' event time. The hazard rate reflects the instantaneous event rate at time t, given that the process has not yet yielded an event. If

$$H_e(t \mid \mathcal{H}_t) \triangleq \int_{t_o}^{t} h_e(\tau \mid \mathcal{H}_t)\, d\tau \tag{3}$$

is the *cumulative (integrated) hazard*, then the process' *survival function* is defined as

$$S_e(t \mid \mathcal{H}_t) \triangleq \mathbb{P}\{T_e > t\} = e^{-H_e(t \mid \mathcal{H}_t)} \tag{4}$$

If the distribution of T_e is absolutely continuous, then it will have a density given by

$$f_e(t \mid \mathcal{H}_t) = -\frac{dS_e(t \mid \mathcal{H}_t)}{dt} = h_e(t \mid \mathcal{H}_t)\, S_e(t \mid \mathcal{H}_t) \tag{5}$$

Right-censoring formulation. Event times are typically unbounded and, hence, one may stop observing a survival process, which has not yet generated an event, after a *right-censoring* time $T_{\mathrm{RC}} > 0$ a.s. In this setting, one can think of observing $T \triangleq \min\{T_e, T_{\mathrm{RC}}\}$ and $\Delta \triangleq [\![T_e \leq T_{\mathrm{RC}}]\!]$, instead of T_e. It is usually the case that T_e and T_{RC} are independent RVs (*independent right-censoring assumption*) and, moreover, that $T_{\mathrm{RC}} = t_{\mathrm{RC}} > 0$ a.s. (*fixed time right-censoring assumption*).

Split population formulation. As it will become clearer later, in our context, it will be useful to think that some survival processes will never generate an event. This leads to the notion of a *split-population* survival process, whose realization can be thought of as being drawn as follows: a RV $R \sim \mathrm{Bernoulli}(\pi)$ is sampled, where $\pi \in [0, 1]$ is a *susceptibility probability*. If $R = 1$, then T_e is drawn from a distribution with density $f_e(t \mid R = 1, \mathcal{H}_t)$ and, otherwise ($R = 0$), the process will never generate an event. In the latter case, it is convenient to regard that $T_e = +\infty$ and, if right-censoring is employed, one has that $T = T_{\mathrm{RC}}$ a.s. For this setting, one can show that the joint distribution of (T, Δ) is given as

$$p_G(t, \delta \mid \mathcal{H}_t) \triangleq [\pi\, f_e(t \mid R = 1, \mathcal{H}_t)]^{\delta} \cdot [S_e(t \mid \mathcal{H}_t)]^{1-\delta} \cdot \\ \cdot [f_{\mathrm{RC}}(t)]^{1-\delta} [S_{\mathrm{RC}}(t)]^{\delta} \tag{6}$$

where, assuming that T_{RC} has an absolutely continuous distribution, $f_{\mathrm{RC}}(\cdot)$ and $S_{\mathrm{RC}}(\cdot)$ are its density and survival function respectively and where

$$S_e(t \mid \mathcal{H}_t) \triangleq \pi\, S_e(t \mid R = 1, \mathcal{H}_t) + (1 - \pi) \tag{7}$$

Since in the vast majority of applications we are not interested in modelling any aspect of T_{RC}, the last two terms of (6), being constants, are almost always omitted. Note that, when $\delta = 1$ is observed, then the observed t corresponds to an observed event time t_e, while, when $\delta = 0$, the observed t corresponds to a right-censoring time t_{RC}. Also, note that, if $\pi = 1$, one obtains a conventional right-censored survival process, for which (2) through (5) apply. Finally, the quantity

$$h_e(t \mid \mathcal{H}_t) \triangleq \frac{\pi f_e(t \mid R = 1, \mathcal{H}_t)}{S_e(t \mid \mathcal{H}_t)} \tag{8}$$

has qualities of a hazard function; one can show that, for $0 \leq \pi < 1$ (split population), it is integrable and, hence, it converges to 0 as t becomes unbounded, while, when $\pi = 1$ (conventional population), it diverges, and, hence, is not integrable.

Prediction probability. Now, assume that we observe the process during the period $[0, t_c]$ of *observation duration* $t_c > 0$ and that we do not record any event. Then, for a forecast window $(t_c, t_c + \Delta t]$ with *forecast horizon* $\Delta t > 0$, we define the *prediction probability* $\mathrm{pp}(t_c, \Delta t)$ as the probability that an event of the process will occur within the forecast window, if right-censoring has not occurred by time t_c, *i.e.*,

$$\mathrm{pp}(t_c, \Delta t) \triangleq \mathbb{P}\{T \leq t_c + \Delta t, \Delta = 1 \mid T > t_c, \mathcal{H}_{t_c}\} \tag{9}$$

Under a fixed right-censoring time assumption and assuming that $\Delta t \leq t_{\mathrm{RC}} - t_c$, this probability can be computed as

$$\mathrm{pp}(t_c, \Delta t) = \frac{S_e(t_c \mid \mathcal{H}_{t_c}) - S_e(t_c + \Delta t \mid \mathcal{H}_{t_c})}{S_e(t_c \mid \mathcal{H}_{t_c})} \tag{10}$$

In specific, this is shown as follows:

$$\mathrm{pp}(t_c, \Delta t) \overset{(9)}{=} \frac{\mathbb{P}\{t_c < T \leq t_c + \Delta t, \Delta = 1 \mid \mathcal{H}_{t_c}\}}{\mathbb{P}\{T > t_c \mid \mathcal{H}_{t_c}\}} \tag{11}$$

$\Delta = 1$ implies that $T = T_e \leq T_{\mathrm{RC}}$ and, thus, (11) can be re-written as

$$\mathrm{pp}(t_c, \Delta t) = \frac{\mathbb{P}\{t_c < T_e \leq t_c + \Delta t, T_e \leq T_{\mathrm{RC}} \mid \mathcal{H}_{t_c}\}}{\mathbb{P}\{T > t_c \mid \mathcal{H}_{t_c}\}} \tag{12}$$

Since $R = 0$ implies $\Delta = 0$ and, hence, that $T_{\mathrm{RC}} < T_e = +\infty$, (12)'s numerator can be written as

$$\begin{aligned} &\mathbb{P}\{t_c < T_e \leq t_c + \Delta t, T_e \leq T_{\mathrm{RC}} \mid \mathcal{H}_{t_c}\} = \\ &= \mathbb{P}\{0 < T_e - t_c \leq \Delta t, T_e \leq T_{\mathrm{RC}} \mid R = 1, \mathcal{H}_{t_c}\}\, \pi + \\ &+ \underbrace{\mathbb{P}\{0 < T_e - t_c \leq \Delta t, T_e \leq T_{\mathrm{RC}} \mid R = 0, \mathcal{H}_{t_c}\}}_{=0}(1 - \pi) = \\ &= \pi\, \mathbb{P}\{0 < T_e - t_c \leq \Delta t, T_e \leq T_{\mathrm{RC}} \mid R = 1, \mathcal{H}_{t_c}\} = \\ &= \pi\, \mathbb{E}\big\{S^{+}_{\mathrm{RC}}(T_e)\, [\![0 < T_e - t_c \leq \Delta t]\!] \mid R = 1, \mathcal{H}_{t_c}\big\} \end{aligned} \tag{13}$$

where in the last step we leveraged the independence of T_e and T_{RC} and where we define $S_{\mathrm{RC}}(t) \triangleq \mathbb{P}\{T_{\mathrm{RC}} > t\}$ and $S^{+}_{\mathrm{RC}}(t) \triangleq S_{\mathrm{RC}}(t) + \mathbb{P}\{T_{\mathrm{RC}} = t\}$.

On the other hand, recalling that $T \triangleq \min\{T_e, T_{\mathrm{RC}}\}$ and using once again the independence of T_e and T_{RC}, (12)'s

denominator can be written as

$$\mathbb{P}\{T > t_c \mid \mathcal{H}_{t_c}\} = \mathbb{P}\{T_e > t_c, T_{\mathrm{RC}} > t_c \mid \mathcal{H}_{t_c}\} = \\ = \mathbb{P}\{T_e > t_c \mid \mathcal{H}_{t_c}\} \underbrace{\mathbb{P}\{T_{\mathrm{RC}} > t_c\}}_{=S_{\mathrm{RC}}(t_c)} \quad (14)$$

where

$$\mathbb{P}\{T_e > t_c \mid \mathcal{H}_{t_c}\} = \underbrace{\mathbb{P}\{T_e > t_c \mid R=1, \mathcal{H}_{t_c}\}}_{=S_e(t_c \mid R=1, \mathcal{H}_{t_c})} \pi + \\ + \underbrace{\mathbb{P}\{T_e > t_c \mid R=0, \mathcal{H}_{t_c}\}}_{=S_e(t_c \mid R=0, \mathcal{H}_{t_c})=1} (1-\pi) = \\ = \pi S_e(t_c \mid R=1, \mathcal{H}_{t_c}) + (1-\pi) \quad (15)$$

Substituting (15) into (14) yields

$$\mathbb{P}\{T > t_c \mid \mathcal{H}_{t_c}\} = [\pi S_e(t_c \mid R=1, \mathcal{H}_{t_c}) + (1-\pi)] \cdot \\ \cdot S_{\mathrm{RC}}(t_c) \quad (16)$$

Substituting (13) and (16) into (12) gives the prediction probability under the independent right-censoring assumption, which reads

$$\mathrm{pp}(t_c, \Delta t) = \frac{\mathbb{E}\{S^+_{\mathrm{RC}}(T_e)\, [\![0 < T_e - t_c \leq \Delta t]\!] \mid R=1, \mathcal{H}_{t_c}\}}{[S_e(t_c \mid R=1, \mathcal{H}_{t_c}) + r]\, S_{\mathrm{RC}}(t_c)} \quad (17)$$

where $r \triangleq (1-\pi)/\pi$. If we, now, assume a fixed right-censoring time $T_{\mathrm{RC}} = t_{\mathrm{RC}}$ a.s., then, for $0 < \Delta t \leq t_{\mathrm{RC}} - t_c$, (i) if $t_c \leq T_e < t_c + \Delta t$, $S_{\mathrm{RC}}(T_e) = 1$ and (ii) $S_{\mathrm{RC}}(t_c) = 1$. Hence,

$$\mathbb{E}\{S^+_{\mathrm{RC}}(T_e)\, [\![t_c \leq T_e < t_c + \Delta t]\!] \mid R=1, \mathcal{H}_{t_c}\} = \\ = S_e(t_c | R=1, \mathcal{H}_{t_c}) - S_e(t_c + \Delta t | R=1, \mathcal{H}_{t_c}) \quad (18)$$

Substituting (18) into (17) and using the definition of (7) finally yields (10).

Multi-variate survival process. In this work, we consider the case of a collection of mutually-interacting survival processes, where observed events of some of them influence the rest of them. Such a collection is referred to as a *multi-variate survival process*. The aforementioned interaction is achieved by enforcing the dependence of the i^{th} process' hazard rate $h^i_e(t \mid R=1, \mathcal{H}_t)$ on the history $\mathcal{H}_t$ of all processes up until time t, which includes the timings of observed events of all processes in the collection. Note that the constituent processes are conditionally independent given $\mathcal{H}_t$. This means that the joint distribution of all observed times from all processes of an N-variate survival process would consist of factors of the form $p_G(t^i, \delta^i \mid \mathcal{H}_{t^i})$ for $i \in \{1, 2, \ldots, N\}$.

Anytime Prediction Learning

In this work, we are interested in predicting whether a process event will be encountered in a forecast window $(t_c, t_c + \Delta t]$ for $t_c \in [t_o, t_{\mathrm{RC}})$ and $\Delta t \in (0, t_{\mathrm{RC}} - t_c]$ and where $t_{\mathrm{RC}} > t_o$ is a fixed right-censoring time, given that no such process event has been observed by time t_c. One obvious approach towards this is to estimate the various event time distributions by maximizing the likelihood function based off $p_G(t, \delta \mid \mathcal{H}_t)$ in (6) and then employ these estimates to predict the occurrence of process events via the prediction probability $\mathrm{pp}(t_c, \Delta t)$ of (9). We will refer to this approach as *generative learning*, since, as a byproduct, it estimates distributions of event times, which would allow one to fully simulate the underlying multi-variate process. This estimation approach works well, when training data are in abundance. In contrast, when training data are scarce, *predictive* or *discriminative learning* can be more effective; for example, see (Zhang, Aravamudan, and Anagnostopoulos 2022) in the context of event counts of marked Hawkes processes. In this work, we put forward such an approach, which employs an upper bound of the generative likelihood function – based on $p_G(t, \delta \mid \mathcal{H}_t)$ – as its objective function and consists of factors of the form:

$$p_D(t, \delta \mid \mathcal{H}_t) \triangleq \frac{p_G(t, \delta \mid \mathcal{H}_t)}{[S_e(t \mid \mathcal{H}_t)]^\delta} \propto [h_e(t \mid \mathcal{H}_t)]^\delta\, [S_e(t \mid \mathcal{H}_t)]^{(1-\delta)} \quad (19)$$

Vis-à-vis learning using the $p_G(t, \delta \mid \mathcal{H}_t)$ factor, we see that $p_D(t, \delta \mid \mathcal{H}_t)$ emphasizes the effect of process events that are observed late, *i.e.*, that are close to the right-censoring time.

We motivate the choice of (19) as follows. First, let us assume a fixed right-censoring time $t_{\mathrm{RC}} > t_o$. If we define the RV $L \triangleq [\![t_c < T_e \leq t_c + \Delta t]\!]$ taking on values $\ell \in \{0, 1\}$, a straightforward path towards predictive learning would be to employ a likelihood consisting of factors of the form

$$\mathbb{P}\{L = \ell \mid t_c, \Delta t\} = \mathrm{pp}(t_c, \Delta t)^\ell\, [1 - \mathrm{pp}(t_c, \Delta t)]^{1-\ell} \quad (20)$$

for some fixed t_c and Δt. In an effort to render the predictive learning applicable to arbitrary pairs of $(t_c, \Delta t)$, *i.e.*, in order to achieve *anytime* predictions, one could attempt to derive a lower bound for the expression in (20), whose maximization can be performed independently of $(t_c, \Delta t)$. Thus, estimating parameters by maximizing this lower bound will also force an increase of the $\mathbb{P}\{L = \ell \mid t_c, \Delta t\}$ likelihood factor. Since $\ell = [\![t_c < t_e \leq t_c + \Delta t]\!]$, $1 - \delta = [\![t_c + \Delta t \leq t_{\mathrm{RC}} < t_e]\!]$, then $1 - \ell = [\![t_c + \Delta t < t_e]\!] = [\![t_c + \Delta t < t_e \leq t_{\mathrm{RC}}]\!] + (1 - \delta)$ and from (20) we get that

$$\mathbb{P}\{L = \ell \mid t_c, \Delta t\} = [\mathrm{pp}(t_c, \Delta t)]^{[\![t_c < t_e \leq t_c + \Delta t]\!]} \cdot \\ \cdot [1 - \mathrm{pp}(t_c, \Delta t)]^{[\![t_c + \Delta t < t_e \leq t_{\mathrm{RC}}]\!]}\, [1 - \mathrm{pp}(t_c, \Delta t)]^{1-\delta} \quad (21)$$

Since $[\![t_c < t_e \leq t_c + \Delta t]\!] + [\![t_c + \Delta t < t_e \leq t_{\mathrm{RC}}]\!] = \delta$, the first two factors can be lower-bounded by $\min\{\mathrm{pp}(t_c, \Delta t), 1 - \mathrm{pp}(t_c, \Delta t)\}^\delta$, which yields

$$\mathbb{P}\{L = \ell \mid t_c, \Delta t\} \geq \min\{\mathrm{pp}(t_c, \Delta t), 1 - \mathrm{pp}(t_c, \Delta t)\}^\delta \cdot \\ \cdot [1 - \mathrm{pp}(t_c, \Delta t)]^{1-\delta} \quad (22)$$

From (9), one can see that, for fixed t_c, $\mathrm{pp}(t_c, \Delta t)$ is an increasing function of Δt. Hence, if we choose $\Delta t \approx 0$, from (10) we obtain that $\mathrm{pp}(t_c, \Delta t) \approx h_e(t_c \mid \mathcal{H}_{t_c})\, \Delta t$, where we ignored $o(\Delta t)$ terms. This results in the first term of (22) to be approximately lower-bounded by $h_e(t_c \mid \mathcal{H}_{t_c})\, \Delta t$. Hence, (22) becomes

$$\mathbb{P}\{L = \ell \mid t_c, \Delta t\} \gtrapprox [h_e(t_c \mid \mathcal{H}_{t_c})\, \Delta t]^\delta\, [1 - \mathrm{pp}(t_c, \Delta t)]^{1-\delta} \quad (23)$$

The first term, corresponding to $\delta = 1$, can be lower-bounded by minimizing $h_e(t_c \mid \mathcal{H}_{t_c})$ over $t_c \in [t_o, t_e)$, which can prove to be difficult, as it is specific to the process' hazard function and how this process is affected by the events of the remaining processes in the ensemble. Nevertheless, asymptotically, $h_e(t_c \mid \mathcal{H}_{t_c})$ has to be a non-increasing function of t_c, since it is an integrable function. Therefore, we will assume that

$$\arg \inf_{t_c \in [t_o, t_e)} h_e(t_c \mid \mathcal{H}_{t_c}) \approx t_e \tag{24}$$

which will likely hold for $t_e \gg t_o$. For the second factor of (22), which corresponds to $\delta = 0$, we can lower-bound it by minimizing it first over $\Delta \in (0, t_{\mathrm{RC}} - t_c]$ and, next, over $t_c \in [t_o, t_{\mathrm{RC}})$ as follows

$$\begin{aligned}
&\inf_{t_c \in [t_o, t_{\mathrm{RC}})} \inf_{\Delta t \in (0, t_{\mathrm{RC}} - t_c]} [1 - \mathrm{pp}(t_c, \Delta t)] = \\
&= \inf_{t_c \in [t_o, t_{\mathrm{RC}})} \frac{S_e(t_{\mathrm{RC}} \mid \mathcal{H}_{t_c})}{S_e(t_c \mid \mathcal{H}_{t_c})} \geq \frac{\inf_{t_c \in [t_o, t_{\mathrm{RC}})} S_e(t_{\mathrm{RC}} \mid \mathcal{H}_{t_c})}{\sup_{t_c \in [t_o, t_{\mathrm{RC}})} S_e(t_c \mid \mathcal{H}_{t_c})} = \\
&= S_e(t_{\mathrm{RC}} \mid \mathcal{H}_{t_{\mathrm{RC}}})
\end{aligned} \tag{25}$$

since one can show that, for fixed t, $S_e(t \mid \mathcal{H}_{t'})$ is a non-increasing function of t' and $S_e(t \mid \mathcal{H}_t)$ is a non-increasing function of t. Consolidating the findings of (24) and (25) into (23) yields

$$\mathbb{P}\{L = \ell \mid t_c, \Delta t\} \gtrapprox [h_e(t \mid \mathcal{H}_{t_c}) \, \Delta t]^{\delta} \, [S_e(t \mid \mathcal{H}_t)]^{1-\delta} \tag{26}$$

which inspire us to use the factor $p_D(t, \delta \mid \mathcal{H}_t)$ of (19) in lieu of $p_G(t, \delta \mid \mathcal{H}_t)$ in the formulation of our predictive objective function.

Novel Formulation

In this section we introduce our novel modeling framework for anytime prediction of user engagement in information cascades for arbitrary observation periods and forecast horizons, which we call *Discriminative probabalistic ANyTime user Engagement prediction* (DANTE)[1].

We assume an information network of $N \geq 1$ users that mutually interact by posting, reacting to and (re)sharing content. We model such dynamic behavior as a right-censored multi-variate survival population with split populations – one process per user. The first time a user interacts with their social peers regarding a specific piece of content (*e.g.*, a meme, a piece of news, opinion on a particular subject, etc.) is deemed as the user's/process' *engagement event*. For a given content, the collection of all such engagement events will constitute an information cascade that will be of interest to us and, hence, a realization of the multi-variate process. The timing of the first event, which introduces such content and, hence, initiates a cascade, is deemed to be the common start time t_o of the multi-variate survival process. Furthermore, we will assume a right-censoring time t_{RC} that is common to all constituent processes.

By adopting the aforementioned multi-variate process, we make the following tacit assumptions: (i) each observed information cascade is an i.i.d. realization of the multi-variate process, (ii) users may or may not engage an information cascade depending on their susceptibility of doing so, (iii) if a user is susceptible, then the timing of her engagement may be influenced by past observed engagements of other network users, (iv) all users' behaviors are right-censored at a fixed time t_{RC}.

In particular, for DANTE, the i^{th} user's susceptibility of engaging an information cascade may be modeled via a common/global susceptibility probability π or, when user features (such as user embedding vectors) $\mathbf{x}^i \in \mathbb{R}^D$ are available for the i^{th} user, via a common/global susceptibility probability of the form

$$\pi\big(\mathbf{x}^i, \widetilde{\mathbf{w}}\big) \triangleq \mathbb{P}\big\{R = 1 \mid \mathbf{x}^i, \widetilde{\mathbf{w}}\big\} = \frac{1}{1 + e^{-\widetilde{\mathbf{w}}^T \widetilde{\mathbf{x}}^i}} \tag{27}$$

where $\widetilde{\mathbf{x}}^i \triangleq [(\mathbf{x}^i)^T \quad 1]^T$ and $\widetilde{\mathbf{w}} \in \mathbb{R}^{D+1}$ is a weight vector of parameters that is common to all users and that needs to be inferred.

Furthermore, DANTE assumes that users, which have already engaged a given information cascade, compete for causing the remaining users to engage as well. In particular, for the i^{th} user that has not engaged yet by time t, DANTE assumes a hazard rate given by

$$h_e^i(t \mid R = 1, \mathcal{H}_t) = \sum_{j : t_e^j \in \mathcal{H}_t} a_{i,j} \phi(t - (t_e^j - t_o)) \tag{28}$$

where $\{a_{i,j}\}_{i,j=1}^N$ is a set of non-negative parameters to be learned and which quantify the strength of causal influence that user j exerts on user i. Additionally, $\phi(\cdot)$ is a non-negative function called a *memory kernel*, which is common to all users and that is to be chosen by the modeler. The memory kernel specifies how user-to-user influence evolves over time. Table 1 contains some popular memory kernels that we experimented with in this work. Note that we assume that $\phi(\tau) = 0$ for $\tau < 0$. Finally, we point out that DANTE's modeling assumptions coincide with the ones of NETRATE (Rodriguez, Balduzzi, and Schölkopf 2011), when $\pi\big(\mathbf{x}^i, \widetilde{\mathbf{w}}\big) = 1$ for all users i.

Kernel type	Memory Kernel $\phi(t)$
Constant	1
Power-law	$\frac{[\![t \geq \beta]\!]}{t}$
Unit scale Weibull	$\gamma t^{\gamma - 1}$

Table 1: Memory kernel choices and their associated integrated memory kernels that will be used to define the relationship between historic events to the process. Note that the constants $\beta > 0$ and $\gamma > 0$ in the Power-law kernel and Weibull kernels respectively are determined via hyper-parameter search.

DANTE's Training

Assume a set $\mathcal{C}$ of $|\mathcal{C}|$ i.i.d. realizations (information cascades) of the right-censored split-population multivariate survival process, which we have focused on so far. The c^{th} cascade consists of observed pairs $\{(t^{i,c}, \delta^{i,c})\}_{i=1}^N$, where

[1]Python 3.9.12 code for DANTE can be found at https://github.com/aaravamudan2014/DANTE

$t^{i,c} = t_e^{i,c}$, when $\delta^{i,c} = 1$, and $t^{i,c} = t_{\text{RC}}^c$, when $\delta^{i,c} = 0$. DANTE's penalized negative log-likelihood is based off (19) and, finally, reads as

$$E(\mathbf{A}, \widetilde{\mathbf{w}}) \triangleq \sum_{i=1}^{N} E^i(\mathbf{a}^i, \widetilde{\mathbf{w}}) \tag{29}$$

where

$$E^i(\mathbf{a}^i, \widetilde{\mathbf{w}}) \triangleq -\sum_{c=1}^{|\mathcal{C}|} \Big[\delta^{i,c} \ln h_e^i\Big(t_e^{i,c} \mid \mathcal{H}_{t_e^{i,c}}\Big) + \\ +(1-\delta^{i,c}) \ln S_e^i\big(t_{\text{RC}}^c \mid \mathcal{H}_{t_{\text{RC}}^c}\big)\Big] + \nu \left\|\mathbf{a}^i\right\|_1 \tag{30}$$

and $\mathbf{A} \in \mathbb{R}_+^{N \times N}$ is the matrix that contains all $a_{i,j}$'s, $\mathbf{a}^i$ is $\mathbf{A}$'s i^{th} row, while $h_e^i(t \mid \mathcal{H}_t)$ and $S_e^i(t \mid \mathcal{H}_t)$ are the split population hazard rate and survival function respectively of the i^{th} process, both of which depend on $\mathbf{a}^i \in \mathbb{R}_+^N$ and $\widetilde{\mathbf{w}} \in \mathbb{R}^{D+1}$. Finally, $\nu \geq 0$ is a penalty parameter that is common to all constituent processes.

The minimization of the loss function in (29) can be viewed as a multi-task problem, since all constituent process share the weight vector $\widetilde{\mathbf{w}}$. In order to solve this minimization problem (for a fixed value of ν), DANTE employs a consensus Alternative Direction Method of Multipliers (ADMM) procedure (Boyd et al. 2011). Note that the ADMM algorithm's user/process sub-problems are minimized via a projected (onto the positive orthant) gradient descent with backtracking to guarantee the non-negativity of $a_{i,j}$ and obtain effective learning rate values during training.

DANTE was trained for at least 100 ADMM iterations and the best model was selected based on the mean SLE performance on the validation set. It was trained on an AMD Ryzen Threadripper 3970X 32-Core Processor and was parallelized using the Dask library [2].

Data Description

For showing the merits of our model, we chose real world social media datasets for whom there is a neat interpretation of user engagement. Some of the datasets are accompanied with a user network as well. Since we tackle the task of user engagement prediction, we only consider the first occurrence of each user in the information cascade. For the datasets that provide the friendship network, we utilize the graph embeddings as features that in turn influence the susceptibilty probability in (27). For all the datasets, we use relative times so that $t_o = 0$.

Irvine is a social media dataset collected from an online community of students from University of California, Irvine. This dataset contains information about user's activity on a public forum and was originally collected by (Opsahl 2013). It contains 893 users and 13,288 cascades. We do not consider any user features for this model.

LastFm is a music streaming platform. This data was originally collected in (Celma 2010) and contains the listening history for 1,000 users for 13,998 songs. An information cascade in this context represents a single song that propagates among the users. Much like (Yang et al. 2018) we ignore users who listen to less than 5 songs. This dataset also does not provide any user features.

Digg is a news aggregator that allows users to submit and rate news articles and was obtained from (Hogg and Lerman 2012). An information cascade in this context reflects an users engaging an individual news article. We only considered the most active 200 users and are provided with the friendship network of the users. There are a total of 3,554 cascades.

Memes is a dataset generating out of meme-tracking efforts done by (Leskovec, Backstrom, and Kleinberg 2009). In this dataset, each several websites publish "memes" which refer to content with similar context. So an information cascade consists of multiple such websites (taking on the role of users) publishing a particular piece of content. Here, we also have the underlying network formed if there is a hyper-link between websites. We only consider the top 200 popular websites in the dataset and this resulted in a total of 10,460 cascades.

Experiments and Evaluation

Predicting the Number of Engaged Users. Through the prediction probability of (10), a trained DANTE model is capable of predicting, whether a given, previously-inactive user i will engage an information cascade within a $(t_c, t_c + \Delta t]$ forecast window (the $\{L^i = 1\}$ event) based on how this cascade has evolved over the interval $[t_o, t_c]$. Furthermore, since the events of engaging a cascade (or not) are independent, when conditioned on a cascade's observed past, the distribution of the total count $M(t_c, t_c + \Delta t]$ of users engaging the cascade during the forecast window can be computed by convolving the distributions of the L^i's corresponding to these users, where $L^i \sim \text{Bernoulli}(\text{pp}^i(t_c, \Delta t))$:

$$\{\mathbb{P}\{M(t_c, t_c + \Delta t] = k\}\}_{k=0}^{N-M[t_o,t_c]} = \\ = \mathop{\ast}_{i:t_e^i \notin \mathcal{H}_{t_c}} \big\{\mathbb{P}\big\{L^i = \ell | t_c, \Delta t\big\}\big\}_{\ell=0}^{1} \tag{31}$$

where $M[t_o, t_c]$ is the number of users that already engaged the cascade during the $[t_o, t_c]$ time frame.

Comparative evaluation metric. In order to compare DANTE's predictive performance to competing methods, we will employ the Squared Log Error (SLE) metric, as used in (Yang et al. 2019) and (Cao et al. 2017), which, for a single cascade, is defined as

$$\text{SLE} = \Big[\ln m[t_o, t_c + \Delta t] - \widehat{\ln m}[t_o, t_c + \Delta t]\Big]^2 \tag{32}$$

where $m[t_0, t_c + \Delta t]$ is the actual number of users that have engaged the cascade during the $m[t_0, t_c + \Delta t]$ interval, while $\widehat{\ln m}[t_0, t_c + \Delta t]$ stands for the log-count of such users predicted by each model. In the case of DANTE, we used the point estimate

$$\widehat{\ln m}[t_o, t_c + \Delta t] = \arg\min_g \mathbb{E}\{(\ln M[t_0, t_c + \Delta t] - g)^2\} = \\ = \mathbb{E}\{\ln M[t_o, t_c + \Delta_t]\} = \\ = \sum_{k=0}^{N-m[t_o,t_c]} \ln(k + m[t_o, t_c])\, \mathbb{P}\{M(t_c, t_c + \Delta t] = k\} \tag{33}$$

[2] https://www.dask.org/

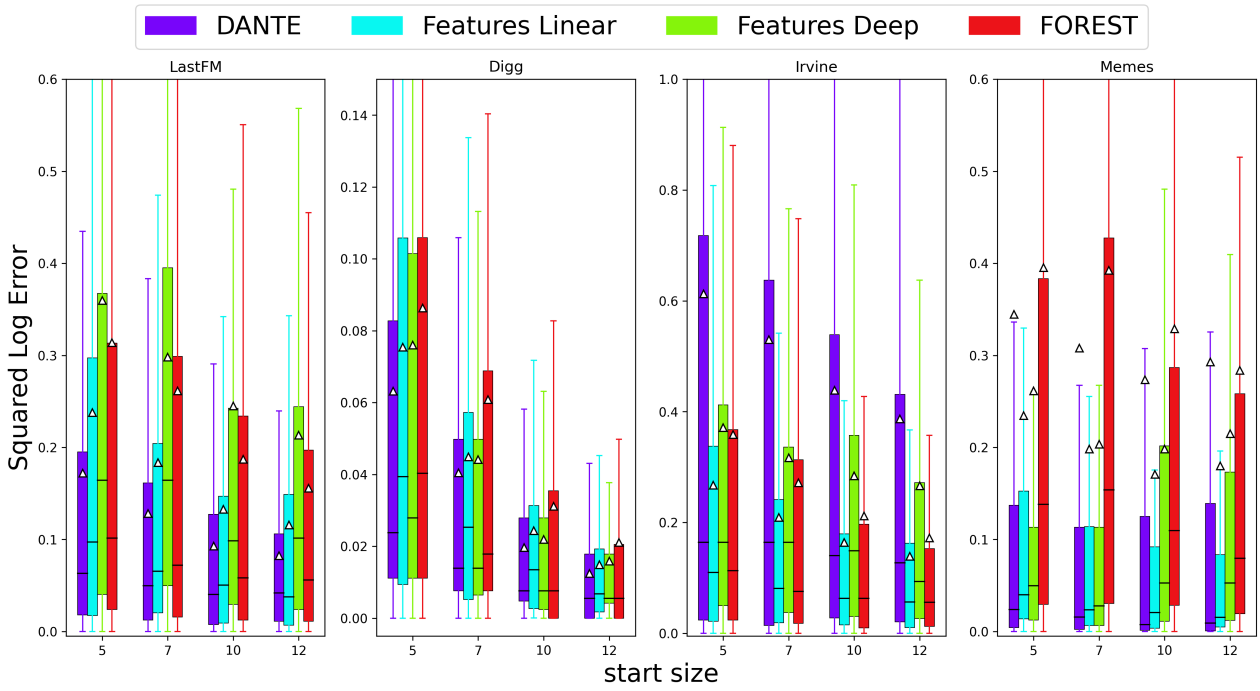

Figure 1: SLE results for final size prediction with varying start sizes of the cascades. The white triangle indicates the mean while the black line is the median SLE.

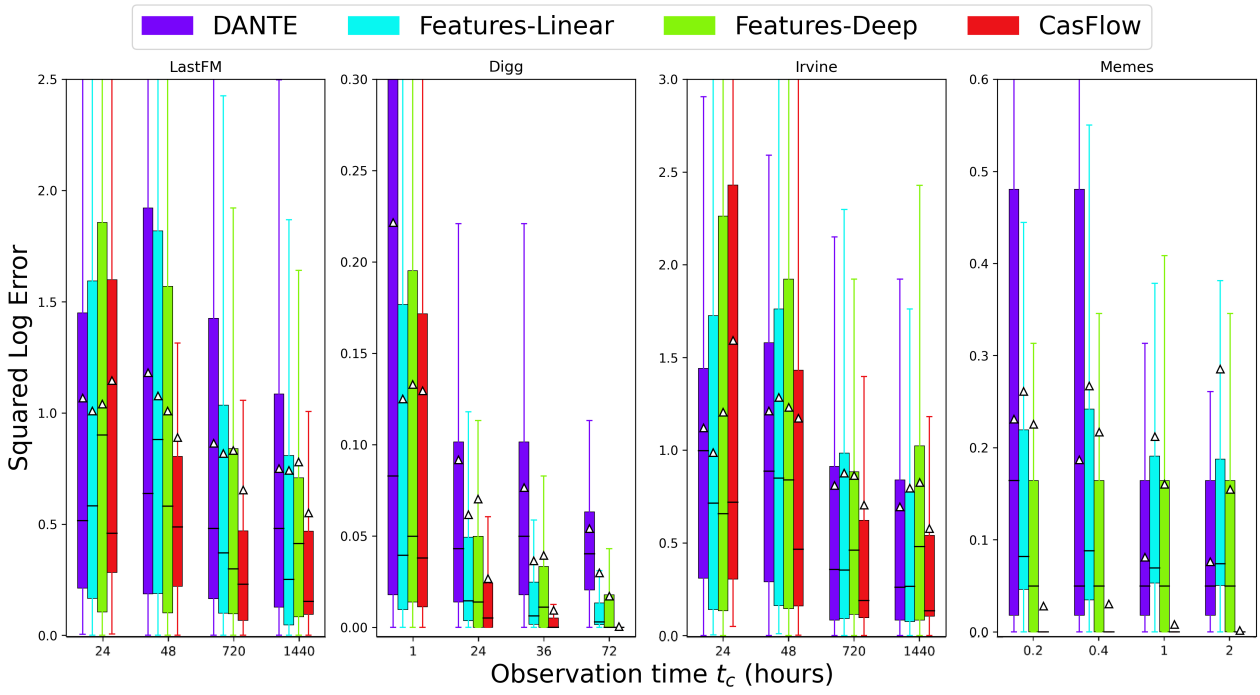

Figure 2: SLE results for final size prediction with varying values of observation time t_c.

while, for competing models, which directly provide a point estimate of the count $\widehat{m}[t_0, t_c + \Delta t]$, we used $\ln(\widehat{m}[t_0, t_c + \Delta t])$ in place of $\widehat{\ln m}[t_0, t_c + \Delta t]$ in (32).

Devising prediction intervals. Having generated the count distribution, the prediction intervals (for some confidence level α) can be generated via several parametric or

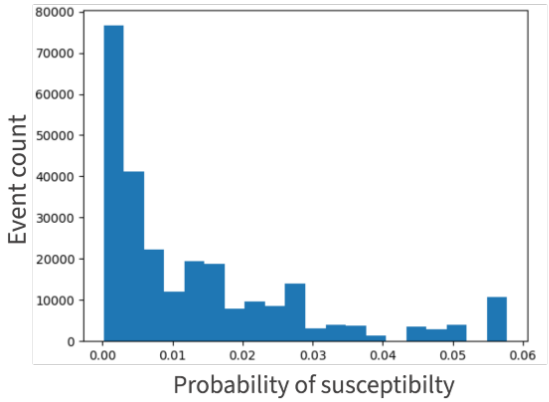

Figure 3: A histogram of the susceptibility probability per user appearing in information cascades for the LastFM dataset. Note that this susceptibility probability leans towards zero, hence motivating the use of split-population in modeling.

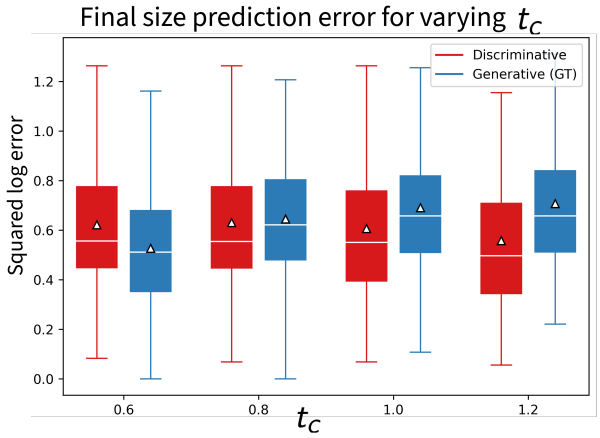

Figure 4: SLE values comparing the discriminative and generative models for final size prediction.

non-parametric methods. We use exact binomial confidence intervals (with a confidence level of 95%) to illustrate the benefits of the model.

Experimental settings. For a given model, we chose models for the *memory kernel* from a list of Power-law, Weibull and constant models with their respective parameters. We refer the reader to Table 1 for their expressions. The memory kernel, in addition to the parameters for the consensus ADMM algorithm comprise the hyper-parameters that were validated on a hold-out set. We found the Power-law kernel (where $\phi(t) = \frac{[\![t \geq \beta]\!]}{t}$ for fixed $\beta > 0$) to be the most effective choice.

Comparison Methods

We here list the models that we compare our proposed model against. While there have been several works listed in realm of popularity prediction for user-engagement, for comparison, we broadly group them into two categories (i) Feature based macro-level methods (ii) User-level deep learning methods. Additionally, a major factor while deciding the baseline methods was that works directly produced a count and did not resort to any simulation based strategy to derive these counts. Among these methods, we note that they either produce counts for a fixed observation time t_c or start size (number of observed users) of the cascade and require training a model per configuration of t_c and Δt.

Features-linear and **Features-deep**: These are feature-based methods simply aggregate temporal handcrafted features (Cheng et al. 2014). This included the cumulative popularity up until t_c, the time between reshares for the first and second half of the information cascade. Then, for each t_c and Δt, we trained two regression models, namely a log-linear regression model and a Multi-Layered Perceptron (MLP) with 1 hidden layer. These models can be used to produce counts based on both observed time t_c and start size of cascade.

FOREST (Yang et al. 2019) is a multi-scale deep learning based diffusion prediction model that combines sequential user prediction with count prediction formulated with a reinforcement learning objective. We were interested in this model even though it was used to predict the timings and identity of the next user because it additionally contained a reinforcement learning objective to predict the size of the cascade given the initial start size. We ran the model without using an initial network embedding. FOREST requires training per start size of the cascade.

CasFlow (Xu et al. 2021) is a state-of-the-art cascade prediction framework that utilises the latent representation of both the structural and temporal information to account for non-linear information diffusion. The model takes in user networks as input and predicts the incremental cascade size after observing up to t_c. CasFlow requires training a model per observation time t_c and prediction time interval Δt. We made no changes to the code apart from the loss function of the model from SLE of the incremental size of the cascade to the SLE of the final size of the cascade (after $t_c + \Delta t$). For the datasets without any network information, we built a network from the event sequences in the cascades.

Results and Discussion

First, we carried out experiments on synthetic data to show that our predictive model can provide us more control over prediction in various time intervals via additional hyper-parameters if needed. Synthetic data was simulated for 100 users via Ogata's thinning algorithm (Ogata 1981) for the generative point process model. DANTE was then trained on this data and was compared to the generative model for varying t_c. The results on synthetic data, found in Figure 4. We notice that the discriminative model tends to perform better in case of scarce training data and larger values of t_c.

The performance of our model in comparison with the baselines for the task of final size prediction with vary-

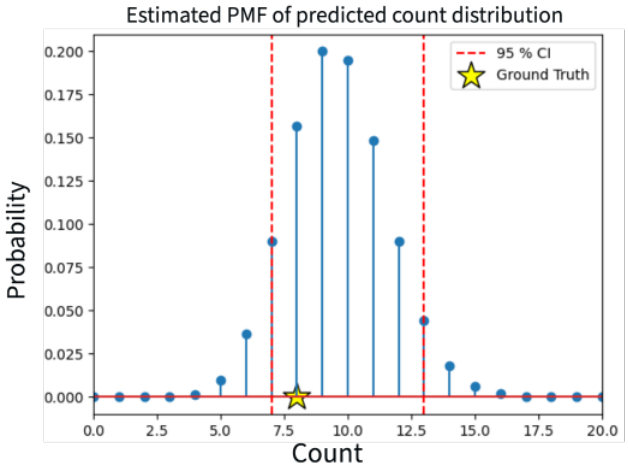

Figure 5: The Probability Mass Function (PMF) of the count distribution of a cascade from the Digg dataset.

ing start sizes – number of observed events – and observation times t_c can be found in Figure 1 and Figure 2 respectively. Note that for a single dataset, we use the same trained model for all the results presented here. In the case of varying start sizes, DANTE outperforms other models for the LastFM, Memes and Digg dataset with respect to the median. However, it performs comparably when considering the mean. For varying t_c, DANTE is highly competitive against Features-linear and Features-deep. Admittedly, CasFlow consistently beats DANTE with respect to both median and mean SLE. This, we attribute to the fact that CasFlow is trained per t_c and is optimized to directly minimize the mean SLE. Note that our model can produce predictions in scenarios for which there is no data samples, a feature that does not extend to the other baselines. In order to predict in a $(t_c, t_c + \Delta t)$ time interval, we do not require (for training) any events that fall in this interval since we use a continuous time point process model.

An added facet of this probabilistic model is that we can generate count prediction intervals via the estimated PMF of the predicted counts. Figure 5 shows an example of a cascade from the Memes dataset. Figure 3 provides a motivating example for adopting a split population formulation. Note that most of the values of π are closer to zero, indicating that most of the users do not engage every cascade. This probability leans closer to 1 for Digg and Memes, while it is 0.57 for Irvine, reinforcing the benefits of this data driven formulation.

Conclusions

In this work, we presented DANTE, a discriminative probabilistic model for predicting user engagement in information cascades. We adopted a split population formulation to account for users' proclivities, or lack thereof, to engage in an information cascade. Our point process based approach renders an interpretable model that can produce predictions for arbitrary observation period and prediction time intervals in a principled way. Additionally, such a perspective helps to provide prediction count intervals, thereby incorporating uncertainty to the model output. Our results for anytime user engagement prediction indicate promising performance against existing state-of-art cascade size prediction methods. Future works in anytime user engagement prediction can seek to capture more complex behaviors by assuming different functional, perhaps non-parametric, forms of the hazard function.

Acknowledgments

This work was supported in part by the U.S. Defense Advanced Research Projects Agency (DARPA) Grant No. FA8650-18-C-7823 under the Computational Simulation of Online Social Behavior (SocialSim) program of DARPA's Information Innovation Office and by the U.S. Air Force Research Laboratory (AFRL) Grant No. FA8650-21-C-1147. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the aforementioned agencies, or the U.S. Government.

References

Aalen, O. O.; Borgan, Ørnulf.; and Gjessing, H. K. 2008. *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag, statistics for biology and health edition.

Bei, Y.; Chen, M.; and Linchi, K. 2011. Toward Predicting Popularity of Social Marketing Messages. In John, S.; Yang, S. J.; Dana, N.; and Sun-Ki, C., eds., *Social Computing, Behavioral-Cultural Modeling and Prediction*, 317–324. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-19656-0.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.*, 3(1): 1–122.

Cao, Q.; Shen, H.; Cen, K.; Ouyang, W.; and Cheng, X. 2017. DeepHawkes: Bridging the Gap between Prediction and Understanding of Information Cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, 1149–1158. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349185.

Celma, O. 2010. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Berlin/Heidelberg, Germany: Springer Publishing Company, Incorporated, 1st edition. ISBN 3642132863.

Chen, F.; and Tan, W. H. 2018. Marked Self-Exciting Point Process Modelling of Information Diffusion on Twitter. *The Annals of Applied Statistics*, 12(4): 2175–2196.

Chen, T.; Rong, J.; Yang, J.; and Cong, G. 2022. Modeling Rumor Diffusion Process With the Consideration of Individual Heterogeneity: Take the Imported Food Safety Issue as an Example During the COVID-19 Pandemic. *Frontiers in public health*, 10: 781691.

Chen, X.; Zhou, F.; Zhang, K.; Trajcevski, G.; Zhong, T.; and Zhang, F. 2019. Information Diffusion Prediction via Recurrent Cascades Convolution. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 770–781.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can Cascades Be Predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, 925–936. New York, NY, USA: Association for Computing Machinery. ISBN 9781450327442.

Farajtabar, M.; Wang, Y.; Gomez-Rodriguez, M.; Li, S.; Zha, H.; and Song, L. 2015. COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Coevolution. *CoRR*, abs/1507.02293.

Farajtabar, M.; Yang, J.; Ye, X.; Xu, H.; Trivedi, R.; Khalil, E.; Li, S.; Song, L.; and Zha, H. 2017. Fake News Mitigation via Point Process Based Intervention. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1097–1106. JMLR.org.

Farajtabar, M.; Ye, X.; Harati, S.; Song, L.; and Zha, H. 2016. Multistage Campaigning in Social Networks. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Gomez-Rodriguez, M.; Balduzzi, D.; and Schölkopf, B. 2011. Uncovering the Temporal Dynamics of Diffusion Networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 561–568. Madison, WI, USA: Omnipress. ISBN 9781450306195.

Gomez-Rodriguez, M.; Leskovec, J.; and Krause, A. 2012. Inferring Networks of Diffusion and Influence. *ACM Trans. Knowl. Discov. Data*, 5(4).

Hogg, T.; and Lerman, K. 2012. Social dynamics of Digg. *EPJ Data Science*, 1(1): 5.

Islam, M. R.; Muthiah, S.; Adhikari, B.; Prakash, B. A.; and Ramakrishnan, N. 2018. DeepDiffuse: Predicting the 'Who' and 'When' in Cascades. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1055–1060. New Jersey, United States: IEEE.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the Spread of Influence through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, 137–146. New York, NY, USA: Association for Computing Machinery. ISBN 1581137370.

Lamprier, S. 2019. A Recurrent Neural Cascade-based Model for Continuous-Time Diffusion. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3632–3641. PMLR.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-Tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 497–506. New York, NY, USA: Association for Computing Machinery. ISBN 9781605584959.

Li, C.; Ma, J.; Guo, X.; and Mei, Q. 2017. DeepCas: An End-to-End Predictor of Information Cascades. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 577–586. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450349130.

Martin, T.; Hofman, J. M.; Sharma, A.; Anderson, A.; and Watts, D. J. 2016. Exploring Limits to Prediction in Complex Social Systems. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 683–694. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450341431.

Mishra, S.; Rizoiu, M.-A.; and Xie, L. 2016. Feature Driven and Point Process Approaches for Popularity Prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, 1069–1078. New York, NY, USA: Association for Computing Machinery. ISBN 9781450340731.

Myers, S. A.; and Leskovec, J. 2010. On the Convexity of Latent Social Network Inference. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, 1741–1749. Red Hook, NY, USA: Curran Associates Inc.

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

Opsahl, T. 2013. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2): 159 – 167. Special Issue on Advances in Two-mode Social Networks.

Rodriguez, M. G.; Balduzzi, D.; and Schölkopf, B. 2011. Uncovering the Temporal Dynamics of Diffusion Networks. In Getoor, L.; and Scheffer, T., eds., *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, 561–568. New York, NY, USA: ACM. ISBN 978-1-4503-0619-5.

Shen, H.; Wang, D.; Song, C.; and Barabási, A.-L. 2014. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, 291–297. AAAI Press.

Tan, W. H.; and Chen, F. 2021. Predicting the popularity of tweets using internal and external knowledge: an empirical Bayes type approach. *AStA Advances in Statistical Analysis*, 105(2): 335–352.

Wang, S.; Zhou, L.; and Kong, B. 2020. Information cascade prediction based on T-DeepHawkes model. *IOP Conference Series: Materials Science and Engineering*, 715(1): 12042.

Xu, X.; Zhou, F.; Zhang, K.; Liu, S.; and Trajcevski, G. 2021. CasFlow: Exploring Hierarchical Structures and Propagation Uncertainty for Cascade Prediction. *IEEE Transactions on Knowledge and Data Engineering*, 1.

Yang, C.; Sun, M.; Liu, H.; Han, S.; Liu, Z.; and Luan, H. 2018. Neural Diffusion Model for Microscopic Cascade Prediction. *CoRR*, abs/1812.08933.

Yang, C.; Tang, J.; Sun, M.; Cui, G.; and Liu, Z. 2019. Multi-scale Information Diffusion Prediction with Reinforced Recurrent Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 4033–4039. International Joint Conferences on Artificial Intelligence Organization.

Yang, S.-H.; and Zha, H. 2013. Mixture of Mutually Exciting Processes for Viral Diffusion. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, II–1–II–9. JMLR.org.

Yu, L.; Cui, P.; Wang, F.; Song, C.; and Yang, S. 2017. Uncovering and predicting the dynamic process of information cascades with survival model. *Knowledge and Information Systems*.

Zhang, X.; Aravamudan, A.; and Anagnostopoulos, G. C. 2022. Anytime Information Cascade Popularity Prediction via Self-Exciting Processes. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 26028–26047. PMLR.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 1513–1522. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642.

Zhou, F.; Xu, X.; Trajcevski, G.; and Zhang, K. 2021. A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances. *ACM Comput. Surv.*, 54(2).