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.

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.

1 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 primarily 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 process and, at the same time, enables 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. 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 addition, 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 this task (Wang, Zhou, and Kong 2020). A shortcoming of these kinds of 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 Δt 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 de-

velop 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. 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 as described in Section 7. 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. 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 untenable for real-world settings. 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 split population yield benefits over traditional generative modeling of 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 censoring 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 Section 3 we provide the reader with some preliminary background about survival processes. Section 4 describes our modeling framework called DANTE. Finally, in Section 6 and Section 7 we detail our experimental methodology and comment on results.

2 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 central assumption about a user appearing only once in a cascade may not always be true. Nevertheless, they are relevant to our task since we can cast user-engagement prediction as a cascade size prediction problem. Based on whether or not each users' engagement is predicted, such cascade size prediction problem can be roughly categorized into two groups, which we will discuss in detail below.

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 follower numbers, gender, etc. as part of input. However, they do not try to predict whether or not a user will participate in a cascade. Within the group, we can further categorize models 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 diffusion process first by specifying intensity functions of the process. Then, the conditional mean or median of the counting process is generally adopted as an estimate for cascade size prediction. Additionally, these works are geared towards specific social networks (Twitter) and make use of features such as user follower counts which may not be available for all 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 automatically 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 individual users, 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.

3 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 (Ghosh 2009) 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. Let us also note that, since the process is assumed stationary, we will henceforth take $t_o = 0$ and consider t to be the time relative to the process' start time. Finally, the non-negative function $h_e(\cdot \mid \mathcal{H}_t)$ is referred to as the process' *hazard rate* and is defined as

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

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

$$H_e(t \mid \mathcal{H}_t) \triangleq \int_0^t h_e(\tau \mid \mathcal{H}_t)\, d\tau \tag{3}$$

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

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

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 $f_e(\cdot \mid R = 1, \mathcal{H}_t)$ 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

$$\begin{aligned} p\,(t, \delta \mid \mathcal{H}_t) = &[\pi\, h_e(\cdot \mid R = 1, \mathcal{H}_t)\, S_e(t \mid R = 1, \mathcal{H}_t)]^{\delta} \cdot \\ &\cdot [(1 - \pi) + \pi\, S_e(t \mid R = 1, \mathcal{H}_t)]^{1-\delta} \cdot \\ &\cdot f_{\mathrm{RC}}(t)^{1-\delta} S_{\mathrm{RC}}(t)^{\delta} \end{aligned} \tag{6}$$

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. Since in the vast majority of applications we are not interested in modelling any aspect of T_{RC}, these last two terms are almost always omitted from (6). 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.

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.*,

$$\mathrm{pp}(t_c, \Delta t) \triangleq \mathbb{P}\{T < t_c + \Delta t, \Delta = 1 \mid T \geq t_c, \mathcal{H}_{t_c}\} \tag{7}$$

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

$$\begin{aligned} &\mathrm{pp}(t_c, \Delta t) = \\ &= \frac{S_e(t_c \mid R = 1, \mathcal{H}_{t_c}) - S_e(t_c + \Delta t \mid R = 1, \mathcal{H}_{t_c})}{S_e(t_c \mid R = 1, \mathcal{H}_{t_c}) + r} \end{aligned} \tag{8}$$

where $r \triangleq (1 - \pi)/\pi$. In specific, this is shown as follows:

$$\mathrm{pp}(t_c, \Delta t) \overset{(7)}{=} \frac{\mathbb{P}\{t_c \leq T < t_c + \Delta t, \Delta = 1\}}{\mathbb{P}\{T \geq t_c\}} \tag{9}$$

$\Delta = 1$ implies that $T = T_e \leq T_{\mathrm{RC}}$ and, thus, (9) can be re-written as

$$\mathrm{pp}(t_c, \Delta t) = \frac{\mathbb{P}\{T_e < t_c + \Delta t, T_e \leq T_{\mathrm{RC}}\}}{\mathbb{P}\{T \geq t_c\}} \tag{10}$$

Since $R = 0$ implies $\Delta = 0$ and, hence, that $T_{\mathrm{RC}} < T_e = +\infty$, (10)'s numerator can be written as

$$\begin{aligned} &\mathbb{P}\{T_e < t_c + \Delta t, T_e \leq T_{\mathrm{RC}}\} = \\ &= \mathbb{P}\{T_e < t_c + \Delta t, T_e \leq T_{\mathrm{RC}} \mid R = 1\} \underbrace{\mathbb{P}\{R = 1\}}_{=\pi} + \\ &+ \underbrace{\mathbb{P}\{T_e < t_c + \Delta t, T_e \leq T_{\mathrm{RC}} \mid R = 0\}}_{=0} \mathbb{P}\{R = 0\} = \\ &= \pi\, \mathbb{P}\{T_e < t_c + \Delta t, T_e \leq T_{\mathrm{RC}} \mid R = 1\} = \\ &= \pi\, \mathbb{E}\{S_{\mathrm{RC}}(T_e)\, [\![t_c \leq T_e < t_c + \Delta t]\!] \mid R = 1\} \end{aligned} \tag{11}$$

where in the last step we leveraged the independence of T_e and T_{RC} and where we define $S_{\mathrm{RC}}(t) \triangleq \mathbb{P}\{T_{\mathrm{RC}} \geq t\}$.

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}, (10)'s denominator can be written as

$$\begin{aligned} \mathbb{P}\{T \geq t_c\} &= \mathbb{P}\{T_e \geq t_c, T_{\mathrm{RC}} \geq t_c\} = \\ &= \mathbb{P}\{T_e \geq t_c\} \underbrace{\mathbb{P}\{T_{\mathrm{RC}} \geq t_c\}}_{=S_{\mathrm{RC}}(t_c)} \end{aligned} \tag{12}$$

where

$$\begin{aligned} \mathbb{P}\{T_e \geq t_c\} = &\underbrace{\mathbb{P}\{T_e \geq t_c \mid R = 1\}}_{=S_e(t_c \mid R=1, \mathcal{H}_{t_c})} \underbrace{\mathbb{P}\{R = 1\}}_{=\pi} + \\ &+ \underbrace{\mathbb{P}\{T_e \geq t_c \mid R = 0\}}_{=S_e(t_c \mid R=0, \mathcal{H}_{t_c})=1} \underbrace{\mathbb{P}\{R = 0\}}_{=1-\pi} = \\ &= \pi S_e(t_c | R = 1, \mathcal{H}_{t_c}) + (1 - \pi) \end{aligned} \tag{13}$$

Substituting (13) into (12) yields

$$\mathbb{P}\{T \geq t_c\} = [\pi S_e(t_c|R=1,\mathcal{H}_{t_c}) + (1-\pi)] \cdot \cdot S_{\mathrm{RC}}(t_c) \quad (14)$$

Substituting (11) and (14) into (10) gives the prediction probability under the independent right-censoring assumption, which reads

$$\mathrm{pp}(t_c,\Delta t) = \frac{\pi\,\mathbb{E}\{S_{\mathrm{RC}}(T_e)\,[\![t_c \leq T_e < t_c+\Delta t]\!] \mid R=1\}}{[\pi S_e(t_c|R=1,\mathcal{H}_{t_c}) + (1-\pi)]\,S_{\mathrm{RC}}(t_c)} \quad (15)$$

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,

$$\mathbb{E}\{S_{\mathrm{RC}}(T_e)\,[\![t_c \leq T_e < t_c+\Delta t]\!] \mid R=1\} = \\ = S_e(t_c|R=1,\mathcal{H}_{t_c}) - S_e(t_c+\Delta t|R=1,\mathcal{H}_{t_c}) \quad (16)$$

Substituting (16) into (15) finally yields (8).

Multi-variate survival process. One can 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 $h_e^i(\cdot \mid \mathcal{H}_t)$ 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, however, that the 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 is given as

$$p\big(\{(t^i,\delta^i)\}_{i=1}^N \mid \mathcal{H}\big) = \prod_{i=1}^N p\big(t^i,\delta^i \mid \mathcal{H}_{t^i}\big) \quad (17)$$

for a single realization of an information cascade.

Predictive Learning. When modelling survival data using a multi-variate survival process, it is typical to parameterize the processes' hazard rates and use a (potentially, penalized) maximum likelihood approach to estimate them. In particular, maximizing a likelihood function based off (6) or (17) will be referred to as *generative learning*, since the distributions of event times are eventually estimated, which allows us to fully simulate the multi-variate process.

However, in this work we are mostly interested in predicting future event counts and, therefore, a *predictive* or *discriminative learning* approach would be more suitable. Such an approach hinges on observing the RV $L^i \triangleq [\![t_c < T_e^i \leq t_c + \Delta t]\!]$ for the i^{th} process. This would lead to a joint distribution of the form

$$p\big(\{\ell^i\}_{i=1}^N \mid \mathcal{H}\big) = \prod_{i=1}^N \mathrm{pp}(t_c,\Delta t)^{\ell^i}\,[1-\mathrm{pp}(t_c,\Delta t)]^{1-\ell^i} \quad (18)$$

for a single realization of an information cascade.

Finally, note that, henceforth, we will be considering multi-variate survival processes, whose constituent processes are right-censored and that stem from a split population.

4 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).

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 a split population – 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 first event which introduces such content is deemed as the *starting event* of the information cascade.

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.

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

$$\pi\big(\mathbf{x}^i \mid \mathbf{w}\big) \triangleq \frac{1}{1+e^{-\widetilde{\mathbf{w}}^T\widetilde{\mathbf{x}}^i}} \quad (19)$$

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 (common to all users) to be learned.

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 as

$$h_e^i(t \mid R=1,\mathcal{H}_t) = \sum_{j:t_e^j \in \mathcal{H}_t} a_{i,j}\,\phi(t-t_e^j) \quad (20)$$

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 has on user i. Additionally, $\phi(\cdot)$ is a non-negative function called a *memory kernel*, which is common to all users and is to be chosen by the modeler. The memory kernel specifies how user-to-user influence evolves over time. Note that we assume that $\phi(\tau) = 0$ for $\tau < 0$. Finally, we point out that the modeling choice reflected in (20) is also adopted in NETRATE (Rodriguez, Balduzzi, and Schölkopf 2011).

DANTE aims to address the task of predicting whether a previously-inactive individual user will engage a given in-

formation cascade within a forecast window after having observed the past activity of other users in the network. Subsequently, by summing up all such predicted events, one can predict the future count of users that will engage this cascade. Such a problem could be tackled by learning the model via a *generative approach* by maximizing a (possibly, penalized) likelihood function based on (17) and, then, determine whether a previously inactive user will engage within the forecast window by means of the prediction probability (7). A *predictive approach* would, instead, base its training on a likelihood function with a quantity akin to (18), which involves the prediction probability directly. The probabilistic models obtained from these two approaches are inherently the same: they can be interpreted the same way and both feature the same parameters, albeit likely bearing different values, since they are estimated differently. This leads our intuition to expect that these models will coincide in the limit of infinite training realizations per user. However, on social media, it is common that the vast majority of users exhibit little to no activity in the online discourse, which means that there will be scant training data for the majority of users. Furthermore, a predictive learning approach is more befitting vis-à-vis the task at hand.

Nevertheless, employing a likelihood based on (18) for training would yield a predictive model for a fixed observation period duration t_c and a fixed forecast horizon Δt, which considerably limits its applicability. The main novel aspect of DANTE is in its discriminative learning approach: unlike prior works, which learn models for fixed/specific $(t_c, \Delta t)$ pairs, DANTE yields a single predictive model for arbitrary $(t_c, \Delta t)$, which we refer to as *anytime prediction*. Instead of maximizing a (potentially penalized) likelihood based on (18), which necessitates a fixed $(t_c, \Delta t)$ pair, DANTE accomplishes its task by employing

$$\widetilde{p}\big(\{\ell^i\}_{i=1}^N \mid \mathcal{H}\big) = \prod_{i=1}^{N} \left[\frac{f_e\big(t_e^i \mid R=1, \mathcal{H}_{t_e^i}\big)}{S_e\big(t_e^i \mid R=1, \mathcal{H}_{t_e^i}\big) + r}\right]^{\ell^i} \cdot \cdot \Big[\pi\, S_e\big(t_{\mathrm{RC}} \mid R=1, \mathcal{H}_{t_e^i}\big) + (1-\pi)\Big]^{1-\ell^i} \tag{21}$$

in its objective function instead.

Next, we provide the rationale behind this particular choice. As illustrated below, the main idea is to lower-bound the terms in (18) with a quantity that is independent of $(t_c, \Delta t)$. Consider the i^{th} term – corresponding to the i^{th} process/user – of (18), where we drop the process index i for notational simplicity:

$$p(\ell \mid \mathcal{H}_{t_c}) = \mathrm{pp}(t_c, \Delta t)^{\ell} \left[1 - \mathrm{pp}(t_c, \Delta t)\right]^{1-\ell} \tag{22}$$

where $\ell \triangleq [\![t_c < t_e \leq t_c + \Delta t]\!]$ and $\mathcal{H}_{t_c}$ stands for the multi-variate process' history that affects the specific user/process. We will refrain from showing the dependence of $p(\ell \mid \mathcal{H}_{t_c})$ on $(t_c, \Delta t)$, again, for notational simplicity. Define $\mathcal{S}_{\sup} \triangleq \{t_c \geq 0, 0 < \Delta t \leq t_{\mathrm{RC}} - t_e\}$ and $\mathcal{S}_{\inf} \triangleq \mathcal{S}_{\sup} \cap \{t_c \in (t_c, t_c + \Delta t]\}$ for a fixed value of $t_{\mathrm{RC}} > 0$. Then, a lower bound of (22) is given by

$$p_{\mathrm{LB}}(\ell \mid \mathcal{H}_{t_c}) \triangleq \left[\inf_{\substack{t_c \ \Delta t \\ (t_c, \Delta t) \in \mathcal{S}_{\inf}}} \inf p(\ell \mid \mathcal{H}_{t_c})\right]^{\ell} \cdot \cdot \left[1 - \sup_{\substack{t_c \ \Delta t \\ (t_c, \Delta t) \in \mathcal{S}_{\sup}}} \sup p(\ell \mid \mathcal{H}_{t_c})\right]^{1-\ell} \tag{23}$$

Using (7), the first term can be written as

$$\inf_{\substack{t_c \ \Delta t \\ (t_c, \Delta t) \in \mathcal{S}_{\inf}}} \inf p(\ell \mid \mathcal{H}_{t_c}) = = -\frac{\frac{S_e(t_c \mid R=1, \mathcal{H}_{t_c}) - S_e(t_c + \Delta t \mid R=1, \mathcal{H}_{t_c})}{\Delta t}}{S_e(t_c \mid R=1, \mathcal{H}_{t_c}) + r} \Delta t \tag{24}$$

Since $p(\ell \mid \mathcal{H}_{t_c})$ is a decreasing function of Δt, as $\Delta t \downarrow 0$, Δt becomes the infinitesimal dt, $t_c \to t_e$ and the numerator tends to the time derivative of $S_e(t \mid R=1, \mathcal{H}_{t_e})$ evaluated at $t = t_c$, which equals $-f_e(t_e \mid R=1, \mathcal{H}_{t_e})$. Hence,

$$\inf_{\substack{t_c \ \Delta t \\ (t_c, \Delta t) \in \mathcal{S}_{\inf}}} \inf p(\ell \mid \mathcal{H}_{t_c}) = \frac{f_e(t_e \mid R=1, \mathcal{H}_{t_e})}{S_e(t_e \mid R=1, \mathcal{H}_{t_c}) + r} dt = = h(t_e \mid \mathcal{H}_{t_e})\, dt \tag{25}$$

where $h(t \mid \mathcal{H}_t)$ for $t \geq 0$ is the hazard rate of the split-population survival process. Regarding the other term, since, again, $p(\ell \mid \mathcal{H}_{t_c})$ is a decreasing function of Δt, the expression is maximized for $\Delta t = t_{\mathrm{RC}} - t_c$

$$\sup_{\substack{t_c \ \Delta t \\ (t_c, \Delta t) \in \mathcal{S}_{\sup}}} \sup p(\ell \mid \mathcal{H}_{t_c}) = = \sup_{t_c \in [0, t_{\mathrm{RC}}]} \left[1 - \frac{S_e(t_{\mathrm{RC}} \mid R=1, \mathcal{H}_{t_{\mathrm{RC}}}) + r}{S_e(t_c \mid R=1, \mathcal{H}_{t_c}) + r}\right] \tag{26}$$

Since $S_e(t \mid R=1, \mathcal{H}_t)$ is a non-increasing function of t, the above maximization problem has optimal $t_c = 0$. Noting that $S_e(0 \mid R=1, \mathcal{H}_0) = 1$, we obtain that

$$\sup_{\substack{t_c \ \Delta t \\ (t_c, \Delta t) \in \mathcal{S}_{\sup}}} \sup p(\ell \mid \mathcal{H}_{t_c}) = = 1 - [\pi S_e(t_{\mathrm{RC}} \mid R=1, \mathcal{H}_{t_{\mathrm{RC}}}) + (1-\pi)] = = 1 - S(t_{\mathrm{RC}} \mid \mathcal{H}_{t_{\mathrm{RC}}}) \tag{27}$$

where $S(t \mid \mathcal{H}_t)$ is the survival function of the split-population survival process and which corresponds to the hazard rate $h(t \mid \mathcal{H}_t)$. Substituting (25) and (27) into (23), one obtains

$$p_{\mathrm{LB}}(\ell \mid \mathcal{H}_{t_c}) = \left[\frac{f_e(t_e \mid R=1, \mathcal{H}_{t_e})}{S_e(t_e \mid R=1, \mathcal{H}_{t_c}) + r} dt\right]^{\ell} \cdot \cdot [\pi S_e(t_{\mathrm{RC}} \mid R=1, \mathcal{H}_{t_{\mathrm{RC}}}) + (1-\pi)]^{1-\ell} = = \left[h(t_e \mid \mathcal{H}_{t_e})\, dt\right]^{\ell} S(t_{\mathrm{RC}} \mid \mathcal{H}_{t_{\mathrm{RC}}})^{1-\ell} \tag{28}$$

which leads to adopting (21) for use in training. It is important to note that our derivation naturally implies that $\ell = \delta$.

Let us finally note that, for DANTE's formulation, the use of a common vector parameter $\mathbf{w}$ in (19), which is shared among users/processes, will lead to a multi-task learning problem, whose optimization we addressed via a standard consensus ADMM (Alternative Direction Method of Multipliers) approach (Boyd et al. 2011).

4.1 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 $t^{i,c} = t_e^{i,c}$, 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 (21) and, finally, reads as

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

where

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

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 $h^i(t \mid \mathcal{H}_t)$ and $S^i(t \mid \mathcal{H}_t)$ are the 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 $\mathbf{w} \in \mathbb{R}^{D+1}$. 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 $\mathbf{w}$. 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), whose pseudocode is provided in the GitHub repository. Note that its i^{th} user sub-problem is addressed via a projected gradient descent with backtracking, where $a_{i,j}$ are projected into/onto the positive orthant, if they attain negative values.

DANTE[1] 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].

5 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 for the susceptibilty probability of (19).

Irvine is a social media dataset collected from an online community of students from University of California, Irvine. More specifically this dataset contains information about user's activity on a public forum. This dataset, originally collected by (Opsahl 2013), 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 allowed 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 publishing a particular piece of content. Here, we also have the underlying network formed if there is a link between websites. We only consider the top 200 popular websites in the dataset and this resulted in a total of 10,460 cascades.

6 Experiments and Evaluation

In order to compare various cascade size prediction methods, we evaluate the predictive performance of our model using the Squared Log Error metric (SLE), as utilised in (Yang et al. 2019; Cao et al. 2017). SLE is defined as follows

$$\{SLE\}_{c=1}^{C} = (\log S_c^{t_c+\Delta_t} - \log \widetilde{S}_c^{t_c+\Delta_t})^2$$

Where C is the total number of cascades, $S_c^{t_c+\Delta t}$ is the actual size of cascade c at time $t_c + \Delta t$, $\widetilde{S}_c^{t_c+\Delta t}$ is the prediction for the same.

Predicting Number of Events. We can devise a prediction distribution for cascade size given a fixed t_c and Δt by convolving the Bernoulli prediction probabilities as described in (8). This is formalized as follows.

$$\mathbb{P}\{S_c = k\} = *_{i:t_i \notin \mathcal{H}_{t_c}} \mathbb{P}\{C_i = 1\} \tag{31}$$

Where C_i is RV $C_i \sim \text{Bernoulli}(\text{pp}_\text{i}(t_c, \Delta t))$ The above expression evaluates the distribution of number of events inside the prediction time interval $[t_c, \Delta t]$. Note that we denote the set of users that have yet to participate in the cascade using $\{i : t_i \notin \mathcal{H}_{t_c}\}$. Since our evaluation metric is SLE, the probabilistic nature of our model allows us to derive predictions for cascade sizes that minimize the following metric.

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

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

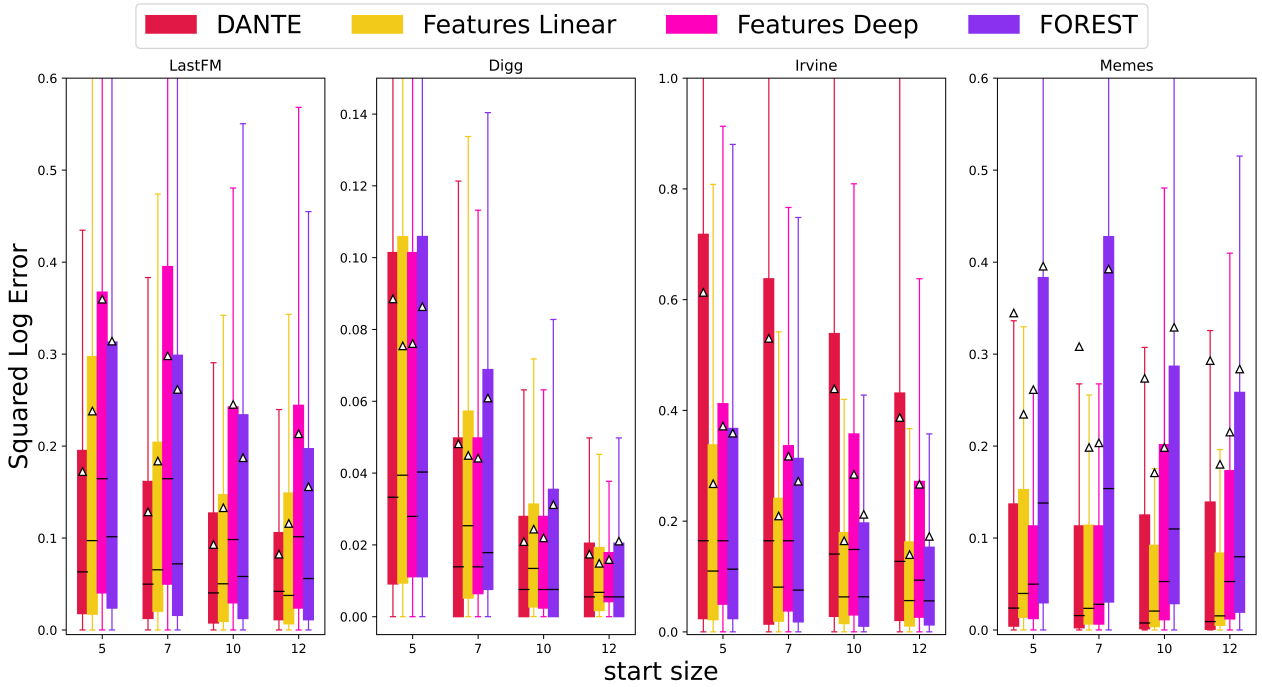

Figure 1: SLE results for final size prediction with varying start sizes of the cascades. The white triangle indicates the mean while the black line is the median SLE.

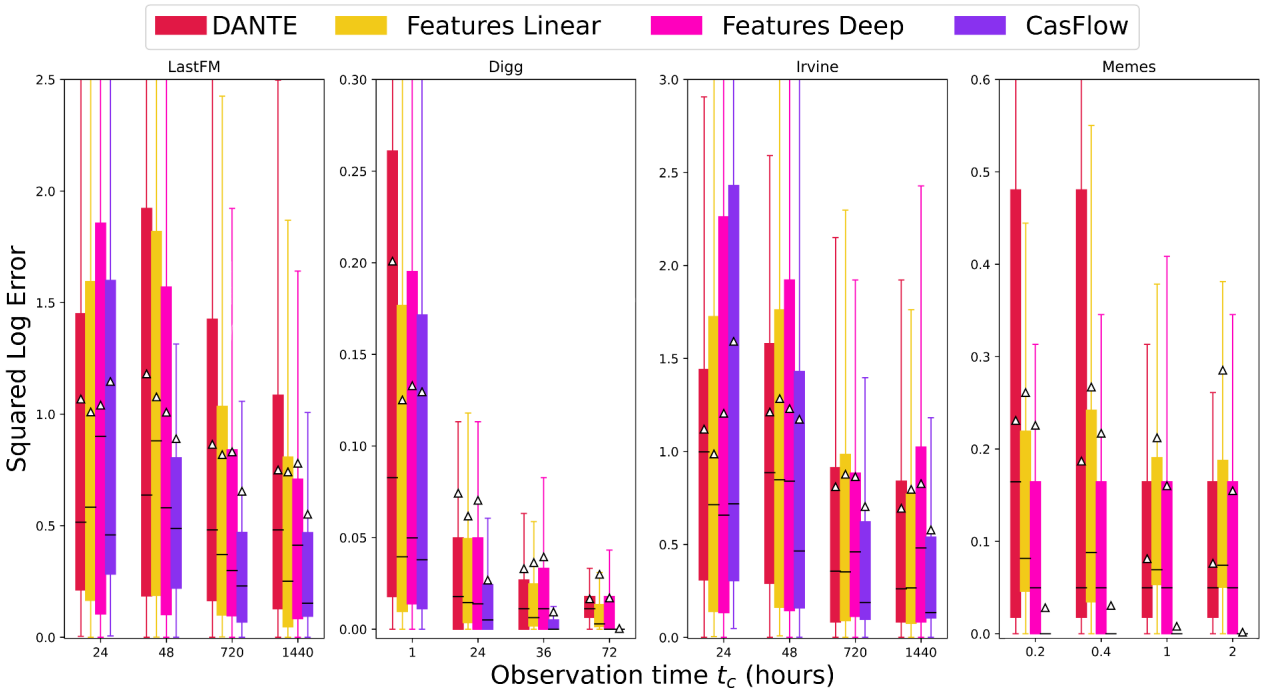

Figure 2: SLE results for final size prediction with varying values of observation time t_c.

$$\arg\min_g \mathbb{E}\{\left(\log S_c^{t_c+\Delta t} - g\right)^2 | C \geq S_c^{t_c}\} \quad (32)$$

$$= \mathbb{E}\{\log S_c^{t_c+\Delta_t} | S_c^{t_c+\Delta t} \geq S_c^{t_c}\} \quad (33)$$

$$= \sum_{k=0}^{D-S_c^{t_c}} \log(k + S_c^{t_c}) \mathbb{P}\{C_e = k\} \quad (34)$$

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

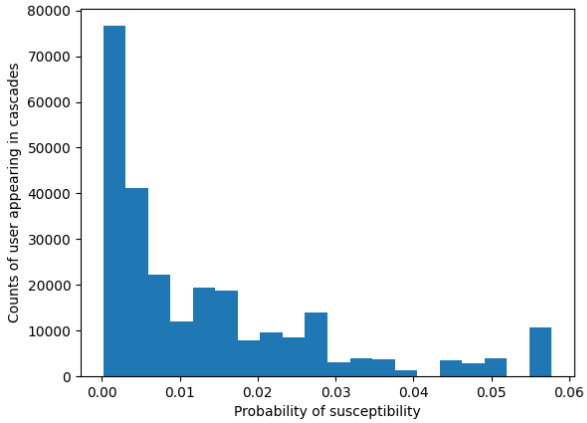

Figure 3: A histogram of the susceptibility probability per user appearing in information cascades for the LastFM dataset. Note that this susceptibility probability leans towards zero, hence motivating the use of split-population in modeling.

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 PowerLaw, Exponential and Rayleigh models with respective parameters. We refer the reader to Table 1 in (Gomez-Rodriguez, Balduzzi, and Schölkopf 2011) 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 PowerLaw kernel (where $\phi(t) = \frac{[\![t \geq \beta]\!]}{t}$ for fixed $\beta > 0$) to be an effective choice.

6.1 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 of the cascade and require training a model per configuration.

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 an additional 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 upto 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 $t_c + \Delta t$).

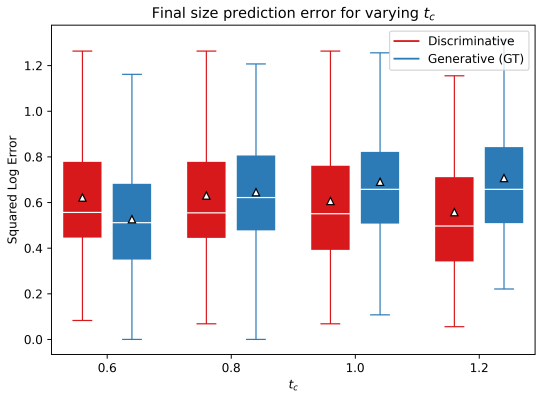

Figure 4: SLE values comparing the discriminative and generative models for final size prediction.

7 Results and Discussion

First, we carried out experiments on synthetic data to show that our predictive model performs better than a generative model for prediction. 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, presented in Figure 4 show that DANTE outperforms the generative model for larger values of t_c and shows its merits for long term prediction. More details on the synthetic experimental setup and results can be found in GitHub repository.

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

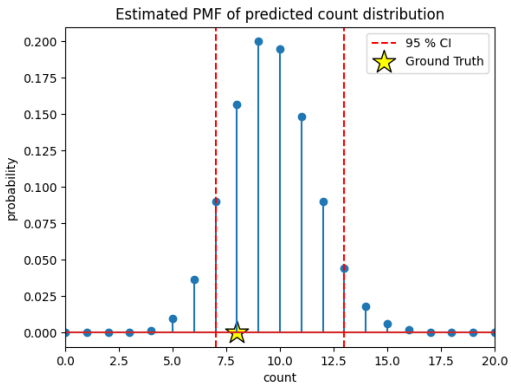

Figure 5: The Probability Mass Function (PMF) of the count distribution of a cascade from the Digg dataset.

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.

8 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

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 Co-evolution. *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.

Ghosh, J. K. 2009. Survival and Event History Analysis: A Process Point of View by Odd O. Aalen, Ørnulf Borgan, Håkon K. Gjessing. *International Statistical Review*, 77(3): 463–464.

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).

Table 1: Table showing the training time per ADMM iterations for the datasets

Dataset	Training cascades	users	Time per ADMM iteration
LastFM	8,398	1,000	∼1,331s
Digg	2,275	200	∼145s
Memes	6,314	200	∼180s
Irvine	7,972	893	∼5030s

A Technical Appendix

Algorithm 1: Consensus ADMM for learning DANTE's parameters

Input: Training set $\mathcal{C}$, $\epsilon_{tol} > 0$, $\tau^{\text{incr}} > 0$, $\tau^{\text{decr}} > 0$, $\mu > 1$, $\rho_{\text{init}} > 0$, $\mathbf{w}^{\text{init}}$

Output: $\mathbf{w}$

$\rho \leftarrow \rho_{\text{init}}$
$\{\mathbf{y_i}\}_{i=1}^{N} \leftarrow \{\mathbf{y_i^{\text{init}}}\}_{i=1}^{N}$
$\mathbf{w}_i \leftarrow w_i^{\text{init}}$
$\mathbf{w} \leftarrow \frac{1}{N}\sum_{i=1}^{N} w_i$
for $t = 1, ..., t_{\max}$ **do**
  for $i = 1, ..., N$ **do**
    $\mathbf{q} \leftarrow \mathbf{w} - \frac{1}{\rho}\mathbf{y_i}$
    ▷Solve sub-problem for user i
    $\mathbf{w}_i \leftarrow \arg\min_{\alpha_i \succcurlyeq \mathbf{0}, \mathbf{w}} E^i(\mathbf{a}^i, \mathbf{w}) + \frac{\rho}{2}\|\mathbf{w}_i - \mathbf{q}\|_2^2$
  end for
  $\mathbf{w}^{\text{new}} \leftarrow \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_i$
  for $i = 1, ..., N$ **do**
    $\mathbf{y_i} \leftarrow \mathbf{y_i} + \rho(\mathbf{w}_i - \mathbf{w}^{\text{new}})$
  end for
  $\|\mathbf{r}\|_2 \leftarrow \sqrt{\sum_{i=1}^{N}\|\mathbf{w_i} - \mathbf{w}^{\text{new}}\|_2}$
  $\|\mathbf{s}\|_2 \leftarrow \rho\|\mathbf{w}^{\text{new}} - \mathbf{w}\|$
  if $\|\mathbf{r}\|_2 > \mu\|\mathbf{s}\|_2$ **then**
    $\rho \leftarrow \tau^{\text{incr}}\rho$
  end if
  if $\|\mathbf{s}\|_2 > \mu\|\mathbf{r}\|_2$ **then**
    $\rho \leftarrow \frac{\rho}{\tau^{\text{decr}}}$
  end if
  $\mathbf{w} \leftarrow \mathbf{w}^{\text{new}}$
end for

A.1 Experiments

This section includes additional details in relation to both the synthetic and real-world experiments.

A.1.0.0.1 Synthetic Experiments As mentioned in the main paper, we conduct experiments on data simulated via Ogata's thinning algorithm for multivariate processes in order to show the benefits of our approach over the generative approach by using the ground truth parameters of the process. There is an additional step involved in the simulation to incorporate the split population setting. For a set of fixed users, we generate features for each user by sampling from two isotropic Gaussian distributions. For simulating each cascade, we first obtain the set of susceptible users by using their respective susceptibility probabilities $\pi(\mathbf{x}^i|\mathbf{w})$ to fetch from a binomial distribution. Having obtained the susceptible users for a cascade, we then simulate the process via Ogata's thinning algorithm with the hazard function for the non-susceptible users effectively set to zero for non-susceptible users.

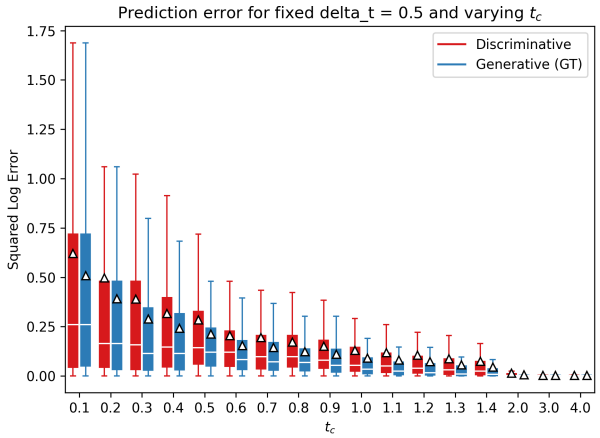

Figure 6: SLE results on the synthetic dataset for changing t_c with fixed Δt

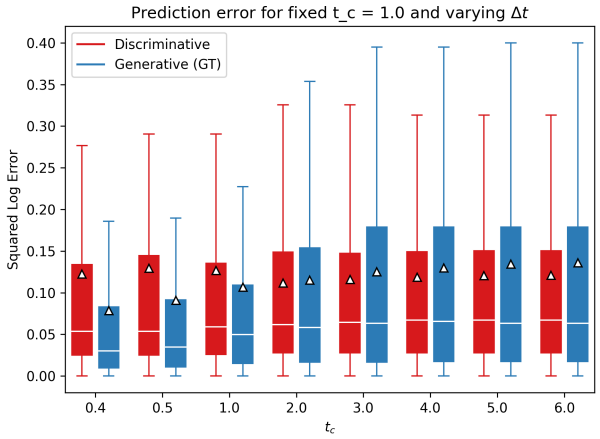

Figure 7: SLE results on the synthetic dataset for changing Δt with fixed t_c

In general, we notice that the performance of the discriminative model can be custom oriented towards certain $(t_c, \Delta t)$ pairs of interest, whereas the ground truth generative model tends to show behaviors that cannot be easily controlled.

A.1.0.0.2 Statistical Significance Test Table 2 and Table 3 show the p-values obtained by performing a Wilcoxon comparison test between DANTE's prediction and the comparison baselines. Note that for values $< 1e-6$ we set the value to be ~ 0. For p-values less than a singificance level

Table 2: P-values for the Wilcoxon signed rank test conducted against DANTE's output for final size prediction with varying start sizes.

Dataset	Irvine				Lastfm				Digg				Memes			
	5	7	10	12	5	7	10	12	5	7	10	12	5	7	10	12
Features-Linear	∼0	∼0	∼0	∼0	∼0	∼0	2e-5	5e-3	.18	.43	.05	.96	.05	5e-4	.06	0.05
Features-Deep	∼0	∼0	∼0	∼0	∼0	∼0	∼0	∼0	.03	.26	.72	.77	7e-3	.01	.42	.58
CasFlow	∼0	∼0	∼0	∼0	∼0	∼0	∼0	4e-5	.9	.07	.01	.36	∼0	∼0	2e-4	.03

Table 3: P-values for the Wilcoxon signed rank test conducted against DANTE's output for final size prediction with varying observation times t_c.

Dataset	Irvine				Lastfm				Digg				Memes			
	1d	2d	30d	60d	1d	2d	30d	60d	1h	24h	36h	72h	.2h	.4h	1h	2h
Features-Linear	.39	.63	.1	8e-3	.85	.67	.08	0.1	∼0	0.29	.21	∼0	.07	∼0	∼0	∼0
Features-Deep	.94	.48	.09	2e-5	.98	.19	.10	.13	∼0	.22	.02	∼0	∼0	∼0	∼0	∼0
CasFlow	.07	.79	1e-3	∼0	.5	.19	∼0	∼0	∼0	∼0	∼0	∼0	0	0	0	0

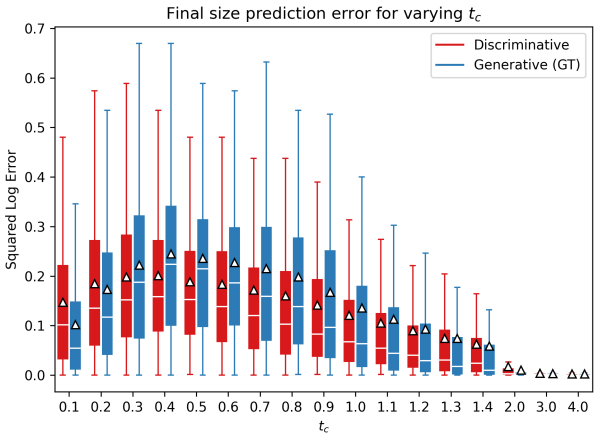

Figure 8: SLE results on the synthetic dataset for final size prediction with changing t_c

of 0.05, we can say that there is enough evidence to reject the null hypothesis that both the SLE errors are from the same population. On the other hand, for p-values greater than 0.05, there is not enough evidence to reject the null hypothesis. For start size performance in Table 2, we notice that the model outputs are similar Irvine and LastFm data since the p-values ~ 0. Digg on the other hand presents enough evidence to say that the model outputs are different with DANTE showing better performance in terms of median. Memes on the other hand shows a significant difference for start sizes 5 and 7 with DANTE producing better medians, while for the other start sizes the null hypothesis cannot be rejected. In terms of performance for final size prediction with varying t_c in Table 3, the model outputs are significantly different for Memes. Digg, Irvine and LastFM show mixed results with CasFlow taking the lead in both mean and median.

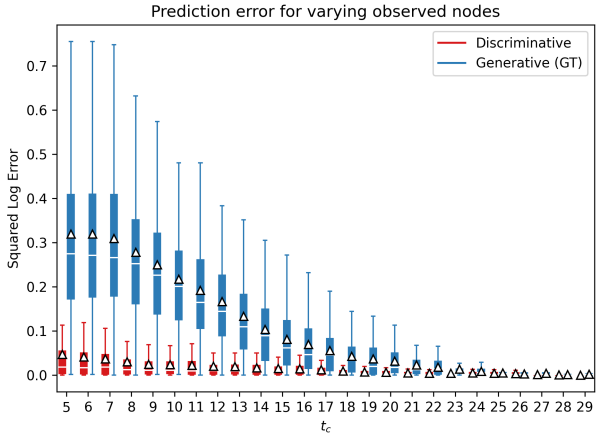

Figure 9: SLE results on the synthetic dataset for final size prediction with changing start size

B Code & Data Appendix

B.1 Dataset

All the datasets apart from the synthetic dataset we used are public. We list the locations where each of them can be found along with additional details that may have been omitted in the main paper. The synthetic dataset is published and can be found in the GitHub repository where we host our code.

- **Digg** and **Memes** can be found in TopoLSTM's GitHub repository[3].
- **LastFM** and **Irvine** were found from Neural Diffusion Model's GitHub repository [4].

B.2 Code and Reproduction information

The GitHub repository contains a README file that has all the information for reproducibilty. A user simply has to edit the MVSPP_config.py file to create a custom run. Other details about data format, locations of datasets and instruction to train and evaluate the model can also be found in the README.

B.2.0.0.1 Hyper-parameter tuning Memory kernel choices: Table 4 includes some popular choices for memory kernels that we tried in our experimentation. We found that exponential pseudo-kernel and power-law kernel produced the best results in the context of social media related datasets. For the parameters in the kernels, we searched over the set $\beta = [0.005, 1.0]$ for both the kernels.

Consensus ADMM hyper-parameters: This included ρ which was set to either 0.1,1.0 or 10. μ was set tp values between 1 and 2. τ^{incr} and τ^{decr} were typically set to values between 1 and 2 with $\tau^{\text{incr}} < \tau^{\text{decr}}$ usually.

Sub-problem optimization parameters: This include the initial alpha values. We usually set this to be a relatively small value between 1e-5 to 1 and searching through them in exponents of 10. Inner iterations refers to the maximum number of backtracking steps in gradient descent. This was typically set to somewhere between 10 and 20. The number of subproblem iterations refers to the number of step in gradient descent. This was set to values in between 10 and 30 for the different datasets. Finally ν was the regularization hyper-parameters and was typically set between 0.0 and 0.5. β^{u} was the scaling factor for the learning rate and was set to values between 1 and 2. $\frac{1}{L_0}$ was initial learning rate and L_0 was set to 1, 10 or 100.

[3] https://github.com/vwz/topolstm/tree/master/datasets/

[4] https://github.com/albertyang33/NeuralDiffusionModel/tree/master/data/

Table 4: Memory kernel choices

Kernel type	Memory Kernel MK $\phi(t)$	Integrated Memory Kernel IMK $\psi(t)$
Constant	1	t
Power-law	$\frac{[\![t \geq \beta]\!]}{t}$	$\ln \frac{t}{\beta} [\![t \geq \beta]\!]$
Weibull	$\gamma t^{\gamma-1}$	t^{γ}
Exponential Pseudo kernel	$[\![t \geq \beta]\!] \exp -\beta t$	$[\![t \geq \beta]\!] \frac{(1-\exp -\beta t)}{\beta}$