# Expressive and Interpretable User Engagement Prediction using Multivariate Survival Processes

Akshay Aravamudan
Florida Institute of Technology
April 1, 2025

# Introduction

Understanding *Information diffusion* – the spread of information online – is pivotal to characterize how information evolves across a network of users.

As information spreads across media, they accumulate to form *information cascades*.

Understanding these dynamics can help

- Devising marketing strategies.
- Political campaign messaging.
- Mis/Disinformation mitigation.
- Predicting product adoption.
- Rumor control.


Figure: Example of information diffusion affecting real-world stock prices. Image Credit[a]

[a] https://www.linkedin.com/pulse/11-tweets-turned-stock-market-upside-down-adam-kornblum/

# Characterizing Information Diffusion

One facet of information diffusion is popularity prediction of online content. Popularity prediction has several forms across several media. We are primarily interested in user engagement prediction.

- Predicting number of new users that will engage an information cascade.
- Engagement can manifest as sharing, liking, retweeting and is highly platform dependent.
- A user can engage an information cascade *up to once*.
- By resharing a post, on say Facebook, they have engaged the information cascade created from the inception of the post.

# User Engagement Prediction

We are interested in the number of users that will engage an information cascade within a period of time. For an information cascade where we have observed events up to a censoring time $t_c$, we are interested in the number of new users that will engage the cascade after a forecast horizon of $\Delta t$.

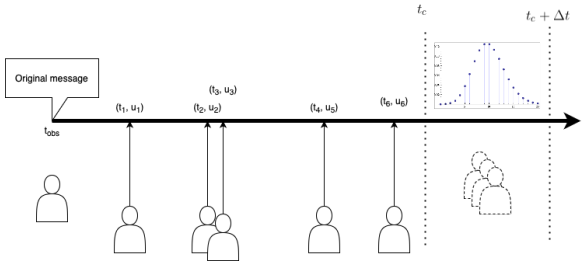



Figure: Illustration of a potential cascade for information diffusion

# Literature Overview

Existing work on popularity prediction can be broadly grouped into two categories – macro-level and user-level popularity prediction.

**Macro-level works**

- Do not model individual behaviors
- Mainly predict final popularity count, unable to attribute to individual users.
- Generative and discriminative approaches via Hawkes process.

**User-level works**

- Model individual user behaviors
- TPP based approaches can be used by modeling individual user processes.
- Generative approaches typically use survival processes, while discriminative approaches use neural techniques.

**Generative models:** often produce lackluster **predictive performance** on account of being trained to describe the data distribution.

**Discriminative works** aim mainly to produce deterministic counts that do not necessarily aim to learn the underlying dynamics.

# Contributions

This dissertation makes the following contributions in the realm of user engagement prediction

- We present a *single prediction model* for anytime user engagement prediction that can be applied across arbitrary observation periods and forecast horizons. This is accomplished via a split-population multi-variate survival process.
- Based on this, we present a user engagement model named DANTE that uses a discriminative loss function.
- We extend DANTE to a neural setting in addition to using a refined discriminative loss. We call this model EXPEDITE.
- We illustrate how to extract meaningful quantities relating to Granger causality that can aid attribution of events sequences.

# Presentation Outline

- Survival Process and additional considerations.
- Discriminative ANyTime user Engagement prediction (DANTE)
- Expressive and interpretable DIscriminative user engagement anyTimE prediction (EXPEDITE)
- Approaching causal attribution via Granger Causality
- Summary

# Contents

Survival processes are a type of random process that model the event time distributions and are capable of producing only one event Aalen et al. (2008).

A Survival process is uniquely characterized by its hazard function which can be formally defined as

$$h_e(t) \triangleq \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}\{T_e \leq t + \Delta t \mid T_e > t\}}{\Delta t} \tag{1}$$

$$= [\![N(t) = 0]\!]\, h_e(t) \tag{2}$$

where $T_e \geq 0$ a.s. is the random variable (RV) representing the process' event time. The hazard rate can be described as the instantaneous rate of an event being yielded by the process at a given time $t$.

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

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

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

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

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

Typically, either the hazard is defined and survival function is derived or vice-versa.

# Incorporating Covariates

Covariates can help enrich the survival process and therefore improve the quality of the learned survival process. In these cases, the hazard function typically is conditioned on these covariates $\mathbf{x}$

$$h_e(t \mid \mathbf{x}) = f(t, \mathbf{x}) \tag{6}$$

The Cox proportional hazard model Cox (1972) is one such simple example where the hazard function is represented as

$$h_e(t) = \lambda_o(t) e^{\beta \mathbf{x}} \tag{7}$$

# Unobserved Heterogeneity

Survival functions will always produce an event as $t \to \infty$. This may not always represent reality for several application. As a correction, we can multiply the hazard function with a RV

$$\tilde{h}(t \mid \mathbf{x}, Z) = Z \cdot h(t \mid \mathbf{x}) \tag{8}$$

Here Z is a random latent term that is assumed to have a non-negative distribution. Under this modification, the unconditional hazard rate $h(t \mid \mathbf{x})$ can be thought of as a special case when $\tilde{h}(t \mid \mathbf{x}, Z = 1)$.

Such a treatment captures unobserved heterogeneity in survival data. The specific form above is referred to as a hazard model with frailty.

In the context of information diffusion, it is undeniable that historical events play a role in how information spreads. In such cases, introducing context can further help enrich the learned model.

$$h_e(t \mid \mathcal{H}_t) = \sum_{t_j \in \mathcal{H}_t} \phi(t - t_j) \tag{9}$$

where $\mathcal{H}_t$ includes the history up until time $t$ and $\phi(\cdot)$ is some non-negative function. Hawkes process Hawkes (1971), another temporal process, that takes as inputs the event history.

# 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.

Therefore we observe $T \triangleq \min\{T_e, T_{\mathrm{RC}}\}$ and $\Delta \triangleq [\![T_e \leq T_{\mathrm{RC}}]\!]$, instead of $T_e$. It is usually the case that $T_e$ and $T_{\mathrm{RC}}$ are independent RVs (*independent right-censoring assumption*) and, moreover, that $T_{\mathrm{RC}} = t_{\mathrm{RC}} > 0$ a.s. (*fixed time right-censoring assumption*).

$$p_G\left(t, \delta \mid \mathcal{H}_t\right) \triangleq \left[\, f_e(t \mid \mathcal{H}_t)\right]^{\delta} \cdot \left[S_e(t \mid \mathcal{H}_t)\right]^{1-\delta} \cdot \\ \cdot \left[f_{\mathrm{RC}}(t)\right]^{1-\delta} \left[S_{\mathrm{RC}}(t)\right]^{\delta} \tag{10}$$

where, assuming that $T_{\mathrm{RC}}$ has an absolutely continuous distribution,

(11)

NETRATE Gomez-Rodriguez et al. (2011) introduce a multi-variate survival process to model information diffusion across a network. They also additionally infer a network from information cascades. The hazard rate is defined as

$$h_e^i(t \mid \mathcal{H}_t) = \sum_{j:t_e^j \in \mathcal{H}_t} a_{i,j}\phi(t - (t_e^j - t_o)) \tag{12}$$

$a_{i,j}$ is the transmission rate and $\phi(\cdot)$ is a non-negative memory kernel function. We adopt a similar approach to modeling information diffusion with additional machinery that is more reflective of real-world diffusion data.

# On Optimizing these models

Such multivariate survival models are typically optimized via log-likelihood and can be expressed as

$$\mathcal{L} = \sum_{c=1}^{|C|} [ \sum_{t_i < T_{\text{RC}}} \log \left[ \sum_{t_j < t_i} h_e^{i,j} \left( t_j - t_i \mid \mathcal{H}_{t_j} \right) \right] + \left[ \sum_{t_i < T_{\text{RC}}} \sum_{t_j < t_i} S_e^{i,j} \left( t_j - t_i \mid \mathcal{H}_{t_j} \right) \right] + \left[ \sum_{t_i < T_{\text{RC}}} \sum_{t_j > T_{RC}} S_e^{i,j} \left( T_{RC} - t_i \mid \mathcal{H}_{t_j} \right) \right] ]$$

The above likelihood can be easily shown to be a convex function, a proof sketch for the convexity of the function can be found in Gomez-Rodriguez et al. (2011). The only constraint on the parameters are that the parameters $\{a_{i,j}\}_{i,j=1}^{N}$ are all greater than or equal zero.

# Contents

**Discriminative ANyTime user Engagement prediction (DANTE)** uses a split-population based multivariate survival model for producing user engagement count estimates.

- like NETRATE, each user will have their own survival process.
- Since some users may never engage an information cascade. This leads to the notion of a *split-population* survival process, whose realization can be thought of as being drawn as follows: a RV $R \sim \text{Bernoulli}(\pi)$ is sampled, where $\pi \in [0, 1]$ is a *susceptibility probability*.

The susceptibility probability is parametrized using a weight vector $\widetilde{\mathbf{w}}$.

$$\pi\big(\mathbf{x}^i, \widetilde{\mathbf{w}}\big) \triangleq \mathbb{P}\big\{R = 1 \mid \mathbf{x}^i, \widetilde{\mathbf{w}}\big\} = \frac{1}{1 + e^{-\widetilde{\mathbf{w}}^T \widetilde{\mathbf{x}}^i}} \tag{13}$$

The unconditional hazard function is

$$h_e(t \mid \mathcal{H}_t) \triangleq \frac{\pi f_e(t \mid R = 1, \mathcal{H}_t)}{S_e(t \mid \mathcal{H}_t)} \tag{14}$$

where the unconditional survival function is

$$S_e(t \mid \mathcal{H}_t) \triangleq \pi \, S_e(t \mid R = 1, \mathcal{H}_t) + (1 - \pi) \tag{15}$$

The form of the user-level hazard function is

$$h_e^i(t \mid R = 1, \mathcal{H}_t) = \sum_{j : t_e^j \in \mathcal{H}_t} \underbrace{a_{i,j}}_{\text{transmission rate}} \underbrace{\phi(t - (t_e^j - t_o))}_{\text{memory kernel}} \tag{16}$$

**Definition**

For an observation periods of $[0, t_c]$ with, $t_c > 0$ where there is no recorded event and $\Delta \triangleq [\![T_e \leq T_{\text{RC}}]\!]$. For a forecast horizon of $\Delta t > 0$ and given that the right censoring time has not occurred until $t_c$, we define the prediction probability

$$\text{pp}(t_c, \Delta t) \triangleq \mathbb{P}\{T \leq t_c + \Delta t, \Delta = 1 \mid T > t_c, \mathcal{H}_{t_c}\} \tag{17}$$

It can be shown that for $r \triangleq (1 - \pi)/\pi$ and $S_{\text{RC}}(t_c) = 1$,

$$\text{pp}(t_c, \Delta t) = \frac{\mathbb{E}\left[S^{+}_{\text{RC}}(T_e)\,[\![0 < T_e - t_c \leq \Delta t]\!] \mid R = 1, \mathcal{H}_{t_c}\right]}{[S_e(t_c \mid R = 1, \mathcal{H}_{t_c}) + r]\, S_{\text{RC}}(t_c)} \tag{18}$$

$$= \frac{S_e(t_c | R = 1, \mathcal{H}_{t_c}) - S_e(t_c + \Delta t | R = 1, \mathcal{H}_{t_c})}{[S_e(t_c \mid R = 1, \mathcal{H}_{t_c}) + r]} \tag{19}$$

We want to optimize for a discriminative, *i.e.,* predictive likelihood. For an RV $L \triangleq [\![t_c < T_e \leq t_c + \Delta t]\!]$, we can aim to maximize

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

However, the dependence on $t_c$ and $\Delta t$ is a crutch in our attempt to formulate an *anytime* prediction model.

We show that prediction probability can be approximately lower-bounded as follows

$$\mathbb{P}\{L = \ell \mid t_c, \Delta t\} \gtrapprox \left[h_e(t \mid \mathcal{H}_{t_c})\, \Delta t\right]^{\delta} \left[S_e(t \mid \mathcal{H}_t)\right]^{1-\delta} \tag{21}$$

Given this, the relation to the classical log-likelihood is

$$p_D(t, \delta \mid \mathcal{H}_t) \triangleq \frac{p_G(t, \delta \mid \mathcal{H}_t)}{\left[S_e(t \mid \mathcal{H}_t)\right]^{\delta}} \propto \left[h_e(t \mid \mathcal{H}_t)\right]^{\delta} \left[S_e(t \mid \mathcal{H}_t)\right]^{(1-\delta)} \tag{22}$$

The $c^{\text{th}}$ 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 (22) and, finally, reads as

$$E(\mathbf{A}, \widetilde{\mathbf{w}}) \triangleq \sum_{i=1}^{N} E^i(\mathbf{a}^i, \widetilde{\mathbf{w}}) \tag{23}$$

where

$$E^i(\mathbf{a}^i, \widetilde{\mathbf{w}}) \triangleq -\sum_{c=1}^{|\mathcal{C}|} \left[\delta^{i,c} \ln h_e^i\Big(t_e^{i,c} \mid \mathcal{H}_{t_e^{i,c}}\Big) + (1-\delta^{i,c}) \ln S_e^i\big(t_{\text{RC}}^{c} \mid \mathcal{H}_{t_{\text{RC}}^{c}}\big)\right] + \nu \left\|\mathbf{a}^i\right\|_1 \tag{24}$$

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 $i^{\text{th}}$ 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 $i^{\text{th}}$ 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.

# Experimental Setting

We conducted experiments on four datasets

- **LastFM** Celma (2010) is a music streaming platform with 1,000 users (no features) and 13,998 cascades (songs).
- **Irvine** Opsahl (2013)is a social media dataset with 893 users (no features) and 13,228 cascades.
- **Digg** Hogg and Lerman (2012) is news aggregator with 200 users and 3,554 cascades. We have friendship network of users converted to graph embeddings.
- **Memes** Leskovec et al. (2009) is a dataset from meme tracking efforts. There are 200 "users" with 10,460 cascades.

We compared DANTE against four other models

- Features-Linear and Features-deep
- **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.
- **CasFlow** Xu et al. (2023) 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.

# Estimating Engaging Users

In each case, we picked either varying $t_c$ or varying start sizes to predict the final size of the cascade.

$$\text{SLE} = \left[\ln m[t_o, t_c + \Delta t] - \widehat{\ln m}[t_o, t_c + \Delta t]\right]^2 \tag{25}$$

where $m[t_0, t_c + \Delta t]$ is the actual number of users that have engaged the cascade in that interval while the latter stands for the estimated log-count..

We generate the PMF by convolving individual Bernoulli distributions to produce a Poisson Binomial.

# Results: Varying $t_c$

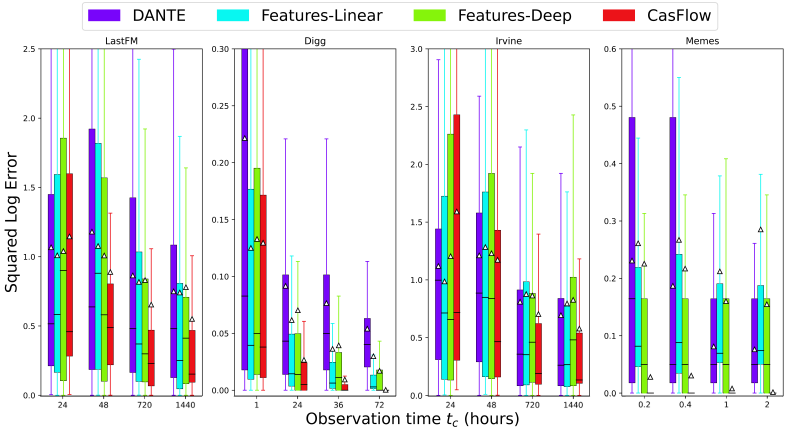



Figure: SLE results for final size prediction with varying values of observation time $t_c$.

# Results: Varying Start Size

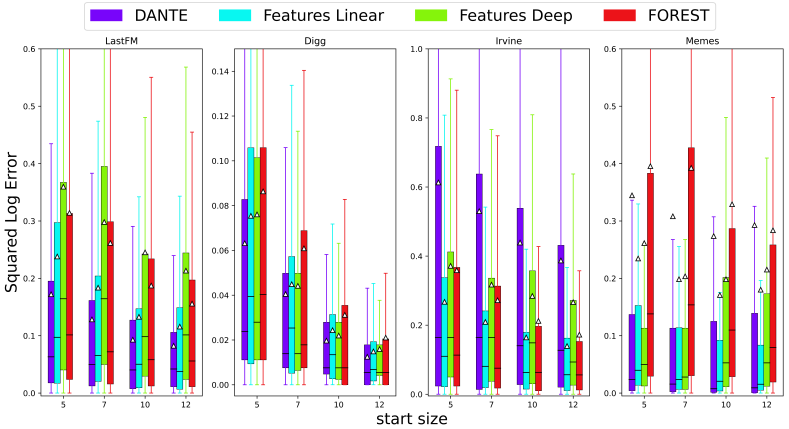
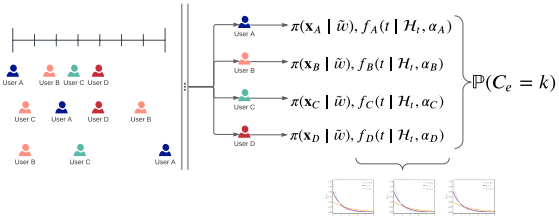



Figure: SLE results for final size prediction with varying start sizes of the cascades. The white triangle indicates the mean while the black line is the median SLE.

# DANTE Summary




For the task of user engagement prediction for arbitrary forecast horizons and observation periods, we propose DANTE that

1. is a split-population based multi-variate survival process model
2. utilizes a discriminative loss function that performs better in lower data settings.
3. performs competitively when compared to models that are trained per $t_c$ and/or $\Delta t$ pair

# Contents

# EXPEDITE

FLORIDA TECH

While DANTE produced competitive results, we faced the following shortcomings

- Having to choose the kernel function through hyperparameter searches limited it's ability.
- An approximate upper bound does not guarantee superior performance in across all $t_c$ and $\Delta t$.
- The implementation of the optimization algorithm was somewhat prohibitive.

To remedy this, we propose Expressive and interpretable DIscriminative user engagement anyTimE prediction (EXPEDITE) that use a data-driven neural approach to learning the memory kernel function. Additionally, we propose a new discriminative loss function that is an exact upper bound to the binary-cross entropy loss.

# Architecture

$$h_e^i(t \mid R = 1, \mathcal{H}_t) = \sum_{j:t_e^j \in \mathcal{H}_t} a_{i,j} \phi(t - (t_e^j - t_o)) \tag{26}$$

$$H_e^i(t \mid R = 1, \mathcal{H}_t) = \sum_{j:t_e^j \in \mathcal{H}_t} a_{i,j} \psi(t - (t_e^j - t_o)) \tag{27}$$

Inspired by point process works that use neural networks to represent underlying quantities Danks and Yau (2022); Omi et al. (2019); Wu et al. (2023), we use a neural architecture to represent key quantities

- The integrated memory kernel $\psi(t \mid \mathbf{w}_\theta)$
- The probability of susceptibility $\pi(\mathbf{x}_i \mid \mathbf{w}_\theta)$
- The matrix of alphas $\{\alpha_{i,j}\}$ uses an embedding matrix.

Optimized via AdamW with learning rates and number of layers adjusted via hyper-parameter search.

# Max-Hinge Loss

$$L_{\mathrm{BCE}}(\mid t_c, \Delta t, \mathcal{H}_{t_c}) \triangleq -\log_2\big(\mathbb{P}\big\{L^i = \ell^i \mid t_c, \Delta t, \mathcal{H}_{t_c}\big\}\big) \tag{28}$$

where

$$\mathbb{P}\big\{L^i = \ell^i \mid t_c, \Delta t, \mathcal{H}_{t_c}\big\} \triangleq \mathrm{pp}^i(t_c, \Delta t \mid \mathcal{H}_{t_c})^{\ell^i} \cdot \cdot \big[1 - \mathrm{pp}^i(t_c, \Delta t \mid \mathcal{H}_{t_c})\big]^{1-\ell^i} \tag{29}$$

and where $\ell^i \triangleq [\![t_c < t^i \leq t_c + \Delta t]\!]$. Defining the binary classification problem's *discriminant function* as

$$d^i(t_c, \Delta_t \mid \mathcal{H}_{t_c}) \triangleq \ln\left(\frac{\mathrm{pp}^i(t_c, \Delta t \mid \mathcal{H}_{t_c})}{1 - \mathrm{pp}^i(t_c, \Delta t \mid \mathcal{H}_{t_c})}\right) \tag{30}$$

one can re-express the binary cross-entropy loss term as

$$L_{\text{BCE}}(u^i) = \log_2\left(1 + e^{-u^i}\right) \tag{31}$$

where $y^i \triangleq 2\ell^i - 1 \in \{-1, 1\}$ and $u^i \triangleq y^i \cdot d^i(t_c, \Delta_t \mid \mathcal{H}_{t_c})$. It is well known that $L_{\text{BCE}}(\cdot)$ is an upper bound to the $0/1$ loss defined as

$$L_{0/1}(u^i) \triangleq [\![u^i < 0]\!] \tag{32}$$

A tighter upper bound is provided by the hinge loss, which, for $\rho > 0$, is defined as

$$L_{\rho}(u^i) \triangleq \max\left\{0, 1 - \frac{u^i}{\rho}\right\} \tag{33}$$

# Max-Hinge Loss

$$L_\rho^{\max}(t_{\text{obs}}^i, \delta^i, \Delta t_{\min}) \triangleq \sup_{(t_c, \Delta t) \in \mathcal{S}_{y^i}} L_\rho(y^i \cdot d^i(t_c, \Delta_t \mid \mathcal{H}_{t_c})) \tag{34}$$

$$\begin{aligned} L_\rho^{\max}(t_{\text{obs}}^i, \delta^i, \Delta t_{\min}) = &\left[L_\rho(d^i(t^i, \Delta t_{\min} \mid \mathcal{H}_{t^i}))\right]^{\delta^i} \cdot \\ &\cdot \left[L_\rho(-d^i(t_o, t_{\text{RC}} \mid \mathcal{H}_{t_o}))\right]^{1-\delta^i} \end{aligned} \tag{35}$$

When the model correctly predicts that a user engages a given cascade in the interval $(t_c, t_c + \Delta t]$, then it will also correctly predict the same event for a larger forecast window $(t_c', t_c' + \Delta t']$, such that $(t_c, t_c + \Delta t] \subseteq (t_c', t_c' + \Delta t']$. Hence, learning to accurately predict user engagement in the smallest possible forecast window guarantees accurate predictions for larger windows.

# Experiments

- For both EXPEDITE and DANTE, we use the same trained model for all $t_c$ and start sizes (as indicated by bold).
- For all other models, we have to train per $(t_c, \Delta t)$ pair.
- The following tables shows the final size predictions results for varying $t_c$ values as well as varying start sizes.

# Results: Varying $t_c$

| Irvine | CasFlow | | **DANTE** | | Features-linear | | Features-deep | | **EXPEDITE** | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t_c$ (hours) | mean | median | mean | median | mean | median | mean | median | mean | median |
| 24 | 1.60 | 0.71 | 1.70 | 0.66 | **0.99** | **0.71** | 1.21 | 0.65 | 1.30 | 0.99 |
| 48 | **1.17** | **0.46** | 1.21 | 0.89 | 1.28 | 0.85 | 1.23 | 0.84 | 1.18 | 0.48 |
| 720 | 0.70 | **0.19** | 0.81 | 0.36 | 0.88 | 0.35 | 0.86 | 0.46 | **0.54** | 0.22 |
| 1440 | 0.58 | **0.13** | 0.69 | 0.26 | 0.80 | 0.27 | 0.83 | 0.48 | **0.47** | 0.16 |

| LastFM | CasFlow | | **DANTE** | | Features-linear | | Features-deep | | **EXPEDITE** | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t_c$ (hours) | mean | median | mean | median | mean | median | mean | median | mean | median |
| 24 | **1.15** | **0.46** | 1.07 | 0.52 | 1.01 | 0.58 | 1.04 | 0.90 | 1.18 | 0.79 |
| 48 | **0.89** | **0.49** | 1.18 | 0.64 | 1.08 | 0.99 | 1.01 | 0.58 | 0.93 | 0.48 |
| 720 | 0.65 | 0.23 | 0.86 | 0.48 | 0.82 | 0.37 | 0.83 | 0.30 | **0.54** | **0.22** |
| 1440 | 0.55 | **0.15** | 0.75 | 0.48 | 0.74 | 0.25 | 0.78 | 0.41 | **0.47** | 0.22 |

Bold indicates that it is the best metric on that column and an underline indicates that its the second best. A star next to the metric indicates the result is statistically significant (p ¡ 0.01) when compared to the preceding ranked metric (via the Wilcoxon signed rank test. Conover (1999)).

# Results: Varying Start Size

| Irvine | FOREST | | **DANTE** | | Features-linear | | Features-deep | | **EXPEDITE** | |
|---|---|---|---|---|---|---|---|---|---|---|
| start size | mean | median | mean | median | mean | median | mean | median | mean | median |
| 5 | 0.37 | 0.11 | 0.60 | 0.16 | 0.26 | 0.11 | 0.36 | 0.17 | **0.26** | 0.13 |
| 7 | 0.27 | 0.07 | 0.52 | 0.16 | **0.21** | 0.08 | 0.32 | 0.19 | 0.22 | 0.12 |
| 10 | 0.22 | **0.06** | 0.45 | 0.14 | **0.16*** | 0.18 | 0.29 | 0.15 | 0.19 | 0.09 |
| 12 | 0.18 | 0.06 | 0.40 | 0.14 | **0.14*** | **0.05** | 0.28 | 0.10 | 0.17 | 0.09 |

| LastFM | FOREST | | **DANTE** | | Features-linear | | Features-deep | | **EXPEDITE** | |
|---|---|---|---|---|---|---|---|---|---|---|
| start size | mean | median | mean | median | mean | median | mean | median | mean | median |
| 5 | 0.31 | 0.10 | **0.17*** | **0.06*** | 0.24 | 0.10 | 0.36 | 0.16 | 0.24 | 0.13 |
| 7 | 0.25 | 0.07 | **0.12*** | **0.05*** | 0.18* | 0.06* | 0.29 | 0.16 | 0.2 | 0.12 |
| 10 | 0.18 | 0.06 | **0.09*** | **0.04*** | 0.13* | 0.05* | 0.24 | 0.09 | 0.16 | 0.11 |
| 12 | 0.15 | 0.06 | **0.08*** | **0.04*** | 0.12* | 0.04* | 0.21 | 0.10 | 0.15 | 0.12 |

Bold indicates that it is the best metric on that column and an underline indicates that its the second best. A star next to the metric indicates the result is statistically significant (p ¡ 0.01) when compared to the preceding ranked metric (via the Wilcoxon signed rank test. Conover (1999)).

# Synthetic Experiments

To validate the loss function we propose, we conduct a synthetic experiment.

- We simulated data via Ogata's thinning algorithm Ogata (1981) for 20 users with 85% of them being susceptible.
- We train three the EXPEDITE architecture with three different loss functions – log-likelihood, DANTE loss and Max-Hinge loss.
- Each model is trained with different sets of cascades - 30, 50, 500 and 1000.
- Each model is trained until the validation loss stops improving.

For each model, we evaluate the Squared Log Error (SLE) over fixed pairs of $(t_c, \Delta t)$.

# Synthetic Experiments Results

| **30 training cascades** | (0.2, 0.2) | | (0.2, 0.5) | | (0.3, 0.3) | | (0.3, 0.5) | | (0.4, 0.2) | | (0.4, 0.5) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model trained on | mean | median | mean | median | mean | median | mean | median | mean | median | mean | median |
| Log-likelihood | 0.46 | 0.26 | 0.87 | 0.48 | 0.37 | 0.11 | 0.47 | 0.11 | 0.15 | 0.02 | 0.26 | 0.01 |
| DANTE loss | **0.35*** | **0.16*** | 0.64* | 0.38* | 0.27* | 0.063* | 0.36* | 0.038* | **0.12*** | **0.01*** | 0.19 | 0.003 |
| Max-Hinge loss | 0.63 | 0.21 | **0.08*** | **0.011*** | **0.14*** | **0.026*** | **0.048*** | **0.011*** | 0.14 | 0.03 | **0.03** | **0.01** |

| **50 training cascades** | (0.2, 0.2) | | (0.2, 0.5) | | (0.3, 0.3) | | (0.3, 0.5) | | (0.4, 0.2) | | (0.4, 0.5) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model trained on | mean | median | mean | median | mean | median | mean | median | mean | median | mean | median |
| Log-likelihood | 0.42 | 0.22 | 0.83 | 0.48 | 0.35 | 0.11 | 0.47 | 0.11 | 0.15 | 0.01 | 0.25 | 0.01 |
| DANTE loss | **0.33*** | **0.16*** | 0.59* | 0.26* | 0.26* | 0.05* | 0.33* | 0.03* | **0.11*** | **0.01*** | 0.19 | **0.003** |
| Max-Hinge loss | 0.72 | 0.21 | **0.1*** | **0.01*** | **0.16*** | **0.03*** | **0.05*** | **0.01*** | 0.15 | 0.03* | **0.03*** | 0.01 |

| **500 training cascades** | (0.2, 0.2) | | (0.2, 0.5) | | (0.3, 0.3) | | (0.3, 0.5) | | (0.4, 0.2) | | (0.4, 0.5) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model trained on | mean | median | mean | median | mean | median | mean | median | mean | median | mean | median |
| Log-likelihood | 0.4 | 0.22 | 0.77 | 0.4 | 0.33 | 0.1 | 0.44 | 0.07 | 0.13 | 0.015 | 0.24 | 0.009 |
| DANTE loss | **0.31*** | **0.14*** | 0.54* | 0.24* | 0.25* | 0.043* | 0.3* | 0.03* | **0.1*** | **0.01*** | 0.19* | 0.003* |
| Max-Hinge loss | 0.65 | 0.21 | **0.08*** | **0.003*** | **0.14** | **0.012** | **0.04*** | **0.003*** | 0.13 | 0.01 | **0.022*** | **0.003*** |

| **1000 training cascades** | (0.2, 0.2) | | (0.2, 0.5) | | (0.3, 0.3) | | (0.3, 0.5) | | (0.4, 0.2) | | (0.4, 0.5) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model trained on | mean | median | mean | median | mean | median | mean | median | mean | median | mean | median |
| Log-likelihood | 0.39 | 0.22 | 0.73 | 0.40 | 0.33 | 0.09 | 0.42 | 0.063 | 0.13 | 0.015 | 0.18 | 0.014 |
| DANTE loss | **0.30*** | **0.14*** | 0.53* | 0.23* | 0.23* | 0.038* | 0.29* | 0.033* | **0.098*** | **0.006*** | 0.14* | 0.004* |
| Max-Hinge loss | 0.69 | 0.21* | **0.086*** | **0.003*** | **0.15*** | **0.012*** | **0.042*** | **0.003*** | 0.14 | 0.012 | **0.08*** | **0.003*** |

# Results and Summary

FLORIDA TECH

**Synthetic Experiments**

1. For higher values of $\Delta t$, Max-Hing loss significantly outperforms.
2. For smaller $\Delta t$, DANTE loss outperforms.
3. All models improve in Mean Squared Log Error (MSLE) as number of training cascade increase.

**Summary**

1. EXPEDITE is a neural and split-population based multi-variate survival models used to produce estimates of the integrated memory kernel.
2. EXPEDITE also utilizes a discriminative loss function that is an exact upper bound to the binary cross-entropy loss.
3. Synthetic and Real-world experiments indicate EXPEDITE is a promising model for user engagement prediction while retaining all interpretable aspects of survival processes.

# Contents

# Causal Attribution

FLORIDA TECH

Point processes are a rich statistical tool set that allows us to derive several quantities of interest.

In the context of multi-variate point processes, Eichler *et. al.* Eichler et al. (2017) shows that if $\alpha_{i,j} > 0$ then the $j^{\text{th}}$ process Granger-causes the $i^{\text{th}}$ process.

**A Detour: Hawkes process** We illustrate the benefit of inferring Granger-causal estimates from real-world X (formerly Twitter) data from the Venezuelan Presidential Crisis.

- Each tweet was annotated with multiple narrative labels via topic modeling.
- These topics were refined by Subject Matter Experts (SME).
- A subset was manually labelled and then used to train a BERT-based multilingual model to label rest of the tweets.

# Data Description

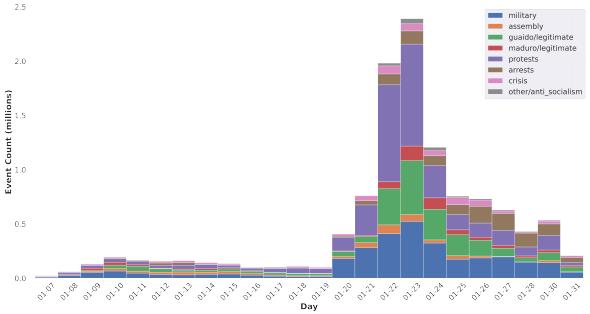



Figure: Histogram of Twitter event (tweet) counts per narrative in 2019. One observes a burst of activity between January 19$^{\text{th}}$ and 21$^{\text{st}}$, during which there was a small-scale coup initiated by 27 soldiers. The peak activity occurring on January 23$^{\text{rd}}$ appears strongly related to massive protests, which demanded Maduro to step down. During the same day, we also witness a significant increase in anti-Maduro tweets.

# Model Description

To model inter-narrative influences, we used a multi-variate Hawkes process that has an intensity of the form

$$\lambda_i(t \mid \mathcal{H}_{t^-}) = b_i(t) + a_{i,i} \sum_{t_k^i \in \mathcal{H}_{t-}^i} \phi_{i,i}(t - t_k^i) + \sum_{\substack{j \in \mathcal{P} \\ j \neq i}} \alpha_{i,j} \sum_{t_k^j \in \mathcal{H}_{t-}^j} \phi_{i,j}(t - t_k^j) \tag{36}$$

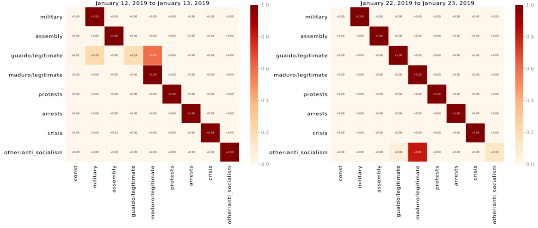
While it is sufficient to use $a_{i,j}$ to represent Granger-causal influences, they don't consider the influence of the memory kernels. To circumvent this, we propose Process Influence Measuress (PIMs).

$$\mathbb{P}\left\{E_k^i \text{ was caused by any earlier event in } \mathcal{E}^j\right\} = \frac{a_{i,j} \sum_{t_\ell^j \in \mathcal{H}^j_{t_k^i -}} \phi_{i,j}(t_k^i - t_\ell^j)}{\lambda_i(t_k^i \mid \mathcal{H}^i_{t_k^i -})} \tag{37}$$

$$\mathbb{P}\left\{E_k^i \text{ was caused by } b_i(t)\right\} = \frac{b_i(t_k^i)}{\lambda_i(t_k^i \mid \mathcal{H}^i_{t_k^i -})} \tag{38}$$

# Experimental Results

We fitted a multi-variate Hawkes process over overlapping time frames. These processes were trained using traditional log-likelihood.




(a) 12–13 January 2019 (b) 22–23 January 2019

Figure: PIM heat maps for two time frames. Columns show influence sources, while rows depict the narratives we studied. These maps illustrate self-driving narratives (prominent diagonal entries), as well as inter-narrative influences (sizeable off-diagonal entries).

# Experimental Results

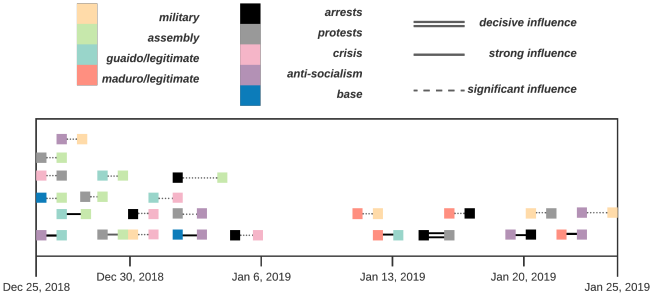
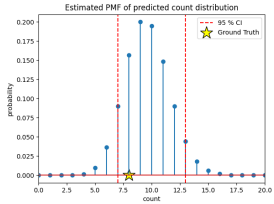


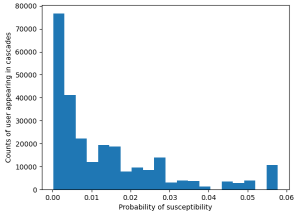

Figure: Timeline indicating noteworthy influences based on our estimated . Weaker cross-narrative influences with values $\pi_{i,j} \in [0.0, 0.2]$ and self-influences of any kind have been omitted for clarity. Note that, during the period of January $6^{\text{th}}$ through the $11^{\text{th}}$, narratives only influence themselves ($\pi_{i,i} \approx 1$).

# Building Intepretable models

As demonstrated earlier, we forward the notion of interpretability by

1. Learning the memory kernel function $\phi(\cdot)$ and integrated memory kernel $\psi(\cdot)$ to derive hazard/intensity function dependent quantities.
2. For prediction, we learn a prediction probability rather than directly predicting next user and next time via black box models.
3. We do not need to obfuscate history via difficult to decipher history vectors while still using highly expressive neural-based models.



# Contents

# Overview

The contributions of this work are as listed

1. **A Split-population based multivariate survival model for predicting user engagement**.
   — We incorporated frailty to accound for unobserved heterogeneity in user engagement.
2. **A deep learning extension** to allow for more evidently expressive capabilities.
   — We parametrized the negative log-survival function and the susceptibility probability using a neural network.
3. **Discriminative loss functions** that aims to learn across arbitrary observation periods and forecast horizons.
   — The loss in DANTE was an approximate upper-bound to the binary-cross entropy loss while the loss in EXPEDITE was an exact upper bound.
4. An illustration of how these models can be easily used to extract **Granger-causal** estimates.
   — We showed how these estimates can be used to extract insights on the Venezuelan Presidential crisis.

# Related Publications

- **Akshay Aravamudan**, Xi Zhang, and Georgios C. Anagnostopoulos. An Expressive yet Interpretable Approach to Anytime User Engagement Predictions. Submitted for consideration to Association for Computing Machinery Knowledge Special Interest Group on Data Discovery and Data Mining ACM SIGKDD 2025 conference. Estimated acceptance rate: 20.1%. Under review.
- **Akshay Aravamudan**, Xi Zhang, G.C. Anagnostopoulos. *Anytime user engagement prediction in information cascades.* AAAI 2023. Acceptance rate: 19.6%. DOI.
- **Akshay Aravamudan**, Xi Zhang, J. Song, S.M. Fiore, G.C. Anagnostopoulos. *Influence dynamics among narratives. Social, Cultural, and Behavioral Modeling*, Springer, 2021. Acceptance rate: 57%. DOI.
- Xi Zhang, **Akshay Aravamudan**, and G C. Anagnostopoulos. A Generalized Time Rescaling Theorem for Temporal Point Processes. Neural Computation, pages 1-15, March, 2025, ISSN 0899-7667. Impact Factor: 3.28. DOI.
- Xi Zhang, **Akshay Aravamudan**, G.C. Anagnostopoulos. *Anytime information cascade popularity prediction via self-exciting processes.* ICML 2022. Acceptance rate: 21.9%. URL.

# Other Publications

- **Akshay Aravamudan**, Zimeena Rasheed, Xi Zhang, Kira E. Scarpignato, Efthymios I. Nikolopoulos, Witold F. Krajewski, and Georgios C. Anagnostopoulos. 'Data-Driven Super-Resolution of Flood Inundation Maps Using Synthetic Simulations'. arXiv [Cs.CV], 2025. arXiv. URL .
- Zimeena Rasheed, **Akshay Aravamudan**, Xi Zhang, Georgios C. Anagnostopoulos, Efthymios I. Nikolopoulos, Combining global precipitation data and machine learning to predict flood peaks in ungauged areas with similar climate, Advances in Water Resources,Volume 192,2024,104781,ISSN 0309-1708, URL
- Ruksana Kabealo, Steven Wyatt, **Akshay Aravamudan**, Xi Zhang, David N. Acaron, Mawaba P. Dao, David Elliott, Anthony O. Smith, Carlos E. Otero, Luis D. Otero, Georgios C. Anagnostopoulos, Adrian M. Peter, Wesley Jones, Eric Lam, A multi-firearm, multi-orientation audio dataset of gunshots, Data in Brief, Volume 48, 2023, 109091, ISSN 2352-3409, URL
- Zimeena Rasheed, **Akshay Aravamudan**, Ali Gorji Sefidmazgi, Georgios C. Anagnostopoulos, Efthymios I. Nikolopoulos, Advancing flood warning procedures in ungauged basins with machine learning, Journal of Hydrology, Volume 609, 2022, 127736, ISSN 0022-1694, URL.

# Acknowledgment

Akshay Aravamudan acknowledges partial support from National Aeronautics and Space Administration Grant No. 80NSSC23K0500, Defense Threat Reduction Agency Grant No. HDTRA1-22-C-0005, 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.

# Thank You!

# 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.

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

Conover, W. (1999). *Practical nonparametric statistics*. Wiley series in probability and statistics. Wiley, New York, NY [u.a.], 3. ed edition.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

# References


Danks, D. and Yau, C. (2022). Derivative-based neural modelling of cumulative distribution functions for survival analysis. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7240–7256. PMLR.

Eichler, M., Dahlhaus, R., and Dueck, J. (2017). Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242.

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, page 561–568, Madison, WI, USA. Omnipress.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

# References


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

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, page 497–506, New York, NY, USA. Association for Computing Machinery.

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

Omi, T., Ueda, N., and Aihara, K. (2019). Fully neural network based model for general temporal point processes. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

# References


Wu, R., Qiao, J., Wu, M., Yu, W., Zheng, M., Liu, T., Zhang, T., and Wang, W. (2023). Neural frailty machine: beyond proportional hazard assumption in neural survival regressions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xu, X., Zhou, F., Zhang, K., Liu, S., and Trajcevski, G. (2023). Casflow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3484–3499.

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*, pages 4033–4039. International Joint Conferences on Artificial Intelligence Organization.

# Appendix - Frailty

When incorporating frailty we need to account for **identifiability**, *i.e.*, these frailty variables are not observed. So typically, this random latent variable $Z$ in marginalized out when trying to construct the likelihood. The unconditional (on latent Z variable) hazard can be re-written as

$$h(t \mid \mathbf{x}) = \mathbb{E}_{Z \sim f_\theta} \left[ Z \cdot h(t \mid \mathbf{x}) \right] \tag{39}$$

$f_\theta$ is some non-negative probability distribution parametrized by $\theta$. The unconditional (on latent Z variable) survival function can be re-written as

$$S(t \mid \mathbf{x}) = \mathbb{E}_{Z \sim f_\theta} \left[ \tilde{S}(t \mid \mathbf{x}) \right] \tag{40}$$

$$= \mathbb{E}_{Z \sim f_\theta} \left[ e^{-Z \cdot H(t \mid \mathbf{x})} \right] \tag{41}$$

# Appendix - Frailty

On a similar thread, the unconditional (on latent Z variable) survival function can be re-written as

$$S(t \mid \mathbf{x}) = \mathbb{E}_{Z \sim f_\theta}\left[\tilde{S}(t \mid \mathbf{x})\right] \tag{42}$$

$$= \mathbb{E}_{Z \sim f_\theta}\left[e^{-Z \cdot H(t \mid \mathbf{x})}\right] \tag{43}$$

An important tool that can be introduced now is the Laplace transform of a probability distribution. The distribution of a non-negative distribution can be uniquely specified by its Laplace transform

$$\mathcal{L}(c) = \mathbb{E}_{Z \sim f_\theta}\left[e^{-cZ}\right] \tag{44}$$

Based on this Equation 42 can we rewritten as

$$S(t \mid \mathbf{x}) = \mathbb{E}_{Z \sim f_\theta}\left[e^{-Z \cdot H(t \mid \mathbf{x})}\right] \tag{45}$$

$$= \mathcal{L}(H(t \mid \mathbf{x})) \tag{46}$$

This is useful to computing the log-likelihood. Additionally, this can help drive the choice of the frailty distribution. We would prefer to choose distributions for whom the Laplace transform exists in closed-form.

# Appendix - Estimating log-counts

In the case of DANTE, we used the point estimate

$$\begin{aligned}\widehat{\ln m}[t_o, t_c + \Delta t] &= \arg\min_g \mathbb{E}\{(\ln M[t_0, t_c + \Delta t] - g)^2\} = \\ &= \mathbb{E}[\ln M[t_o, t_c + \Delta_t]] = \\ &= \sum_{k=0}^{N - m[t_o, t_c]} \ln(k + m[t_o, t_c])\, \mathbb{P}\{M(t_c, t_c + \Delta t] = k\} \end{aligned} \tag{47}$$

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 (25).

# Appendix - DANTE algorithm

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**: $\mathbf{w}$

$\rho \leftarrow \rho_{\text{init}}$
$\{\mathbf{y_i}\}_{i=1}^{N} \leftarrow \{\mathbf{y_i^{init}}\}_{i=1}^{N}$
$\mathbf{w_i} \leftarrow w_i^{\text{init}}$
$\mathbf{w} \leftarrow \frac{1}{N}\sum_{i=1}^{N} w_i$
**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
    $\mathbf{w_i} \leftarrow \arg\min_{\alpha_i \geqslant \mathbf{0}, \mathbf{w}} E^i(\mathbf{a}^i, \mathbf{w}) + \frac{\rho}{2}\|\mathbf{w_i} - \mathbf{q}\|_2^2$
  **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**