Anytime User Engagement Prediction in Information Cascades for Arbitrary Observation Periods

Akshay Aravamudan, Xi Zhang & Georgios C. Anagnostopoulos

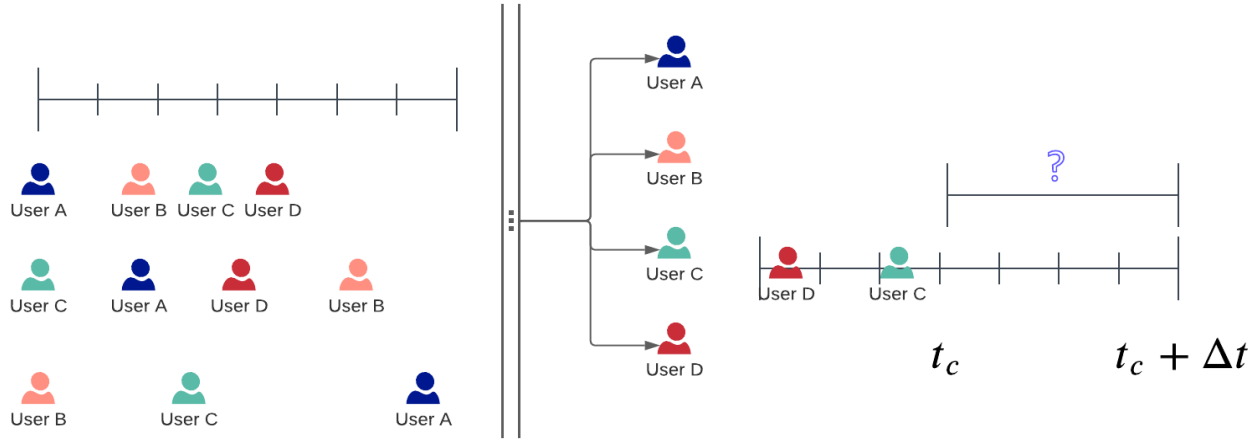
Department of Computer Engineering & Sciences

Florida Institute of Technology

Introduction

- The study and modeling of information diffusion has been a very involved field of late.
- Popularity prediction of messages in social media is of interest to several stakeholders.
- One such facet of popularity prediction is user engagement prediction:

"Predicting the number of new users engaging a specified information cascade"



Motivation

- User engagement prediction has several benefits
  - Product adoption
    - Evaluating how much more popular a new product will become.
  - Political campaigns
    - Determining and framing spread of political messaging.
  - Content moderation and rumor control
    - Determining which messages to send for evaluation by moderators to reduce their workload.
- An additional need is that of explainable predictions.

Goal: For any given observation time t_c we want to predict how many users will engage the information cascade before $t_c + \Delta\, t$

Preliminaries: Temporal Point Processes

- Point processes are used to model ordered events and are uniquely characterized by their intensity function defined as follows

$$\lambda\ (t \mid \mathcal{H}_t) \triangleq \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}\{N(t+\Delta t) - N(t) \mid \mathcal{H}_t\}}{\Delta t}$$

  - $\mathcal{H}_t$ is the history of the process up to time t and is an ordered subset of the information cascade.
  - This dependence on history facilitates complex relationships with previous events.

Related works

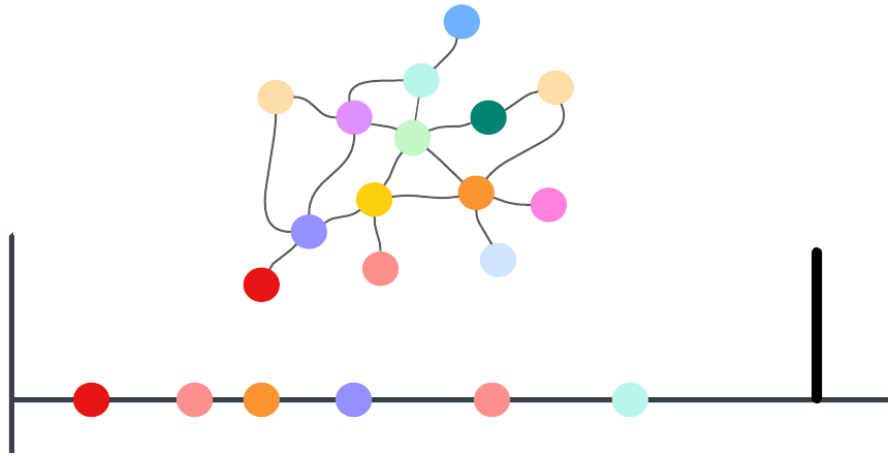
User engagement prediction can be cast as a cascade size prediction problem.

- Given this, existing works primarily fall under two categories
  - Macro level works predicts overall popularity/count.
    - Generative models learn a point process and use it to predict counts via simulations or closed form expressions.
    - Some other discriminative models aim to directly predict the cascade size by using handcrafted features or features learned from deep network models.
  - User level works predict overall popularity/count by aggregating user level predictions.
    - Generative approaches model user level processes and aggregate results for prediction via simulations.
    - Recent methods use deep learning models that aim to predict user level behaviors while targeting macro level count behaviors.

Contributions

- Our approach: Predict user engagement by using a discriminative approach while taking advantage of interpretability offered by point processes.
- In this work, we
  - Provide a single model for all observation times and forecast horizons.
  - Show the benefits of a discriminative split population model over traditional generative modeling.
  - Showcase via real world experiments that our single model performs competitively to the models who require training per t_c and Δt.
  - Provide prediction intervals for the count of users engaging in an information cascade

Assumptions



Modeling assumptions

- Multivariate point process
  - Each user manifests event through their own process.
- Split population
  - Not all users are guaranteed to participate in an information cascade.
- Right censoring
  - The observed data is censored at time independent of the information cascade.
- Discriminative objective
  - We are more interested in modeling sequence of prediction probabilities than occurrence of events.

Preliminaries

- Split Population: users engage with a probability [Schmidt and White 1989]

$$\pi\left(\mathbf{x}^i \,|\mathbf{w}\right) \triangleq \frac{1}{1+e^{-\widetilde{\mathbf{w}}^T\widetilde{\mathbf{x}}^i}}$$

- If users do engage, we employ survival analysis for event time distribution

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

Predictive Likelihood

- Prediction Probabilities

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

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

Where $r \triangleq (1 - \pi)/\pi$

- Building the discriminative likelihood

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

DANTE: Discriminative ANyTime user Engagement Prediction

- Discriminative likelihood

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

- Lower bound of predictive likelihood

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

DANTE training

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

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

We want to train in parallel by effectively decomposing the likelihood function:

Synthetic Experiments and metrics

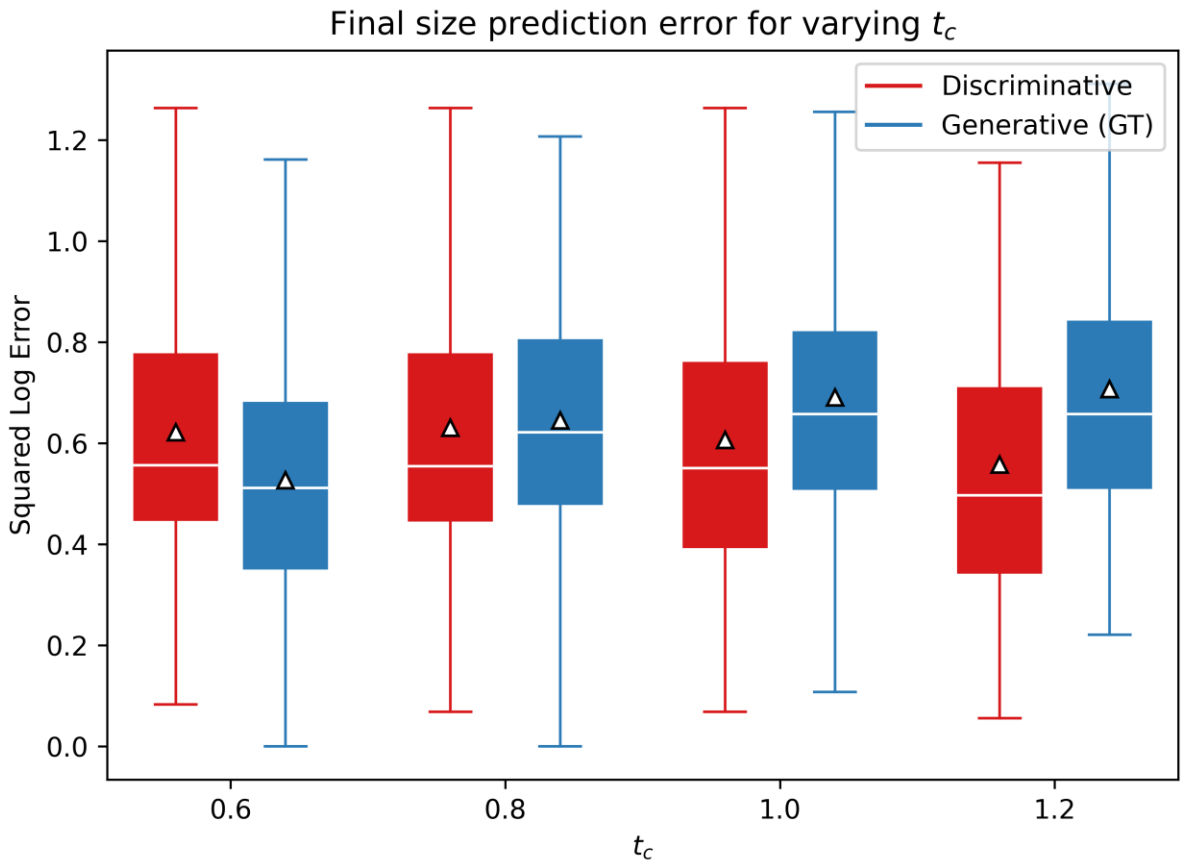
- Squared log error

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

- Synthetic data generation
  - We generate events via a modified Ogata's thinning algorithm that is designed to account for the split population probability.
- Hyperparameter tuning and implementation
  - Kernel hyperparameters: kernel choice and kernel function parameters.
  - Training hyperparameters : consensus ADMM parameters, parameter initialization & sub problem learning rate.

Results synthetic data

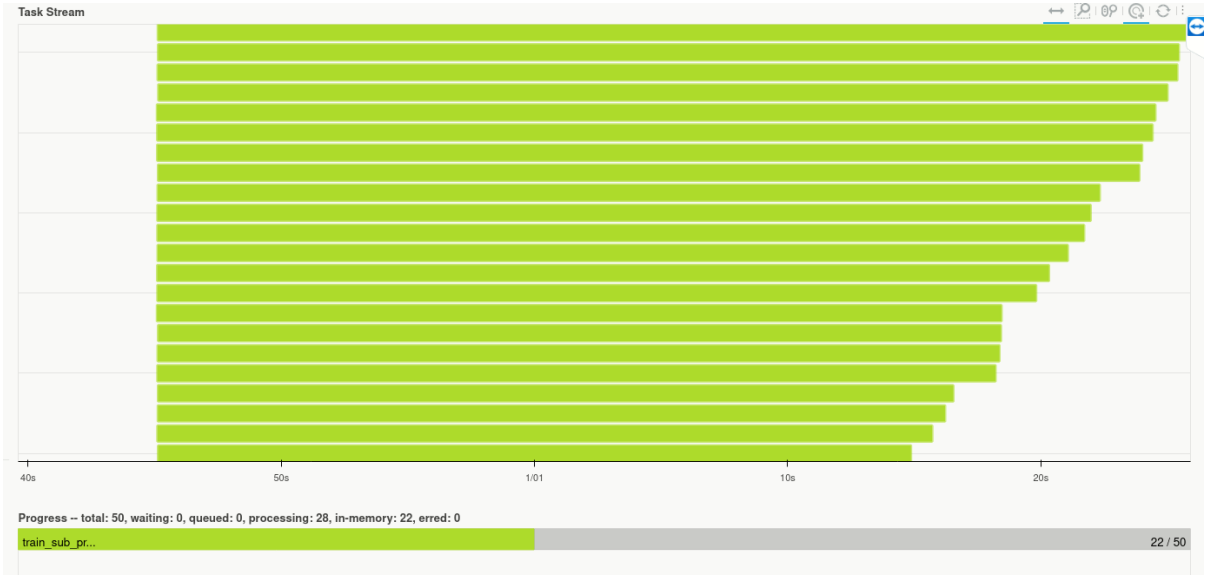
- We evaluated the effect of our discriminative formulation on the predictive performance.
- We noticed that the discriminative model performed better on higher values of observation time.
- Additionally, the discriminative model performed better in instances of smaller training datasets



Real World Experiments

- Datasets
  - LastFm
  - Irvine
  - Digg
  - Memes
- Comparison methods
  - Feature-Linear/deep
  - FOREST (2019)
  - CasFlow (2021)

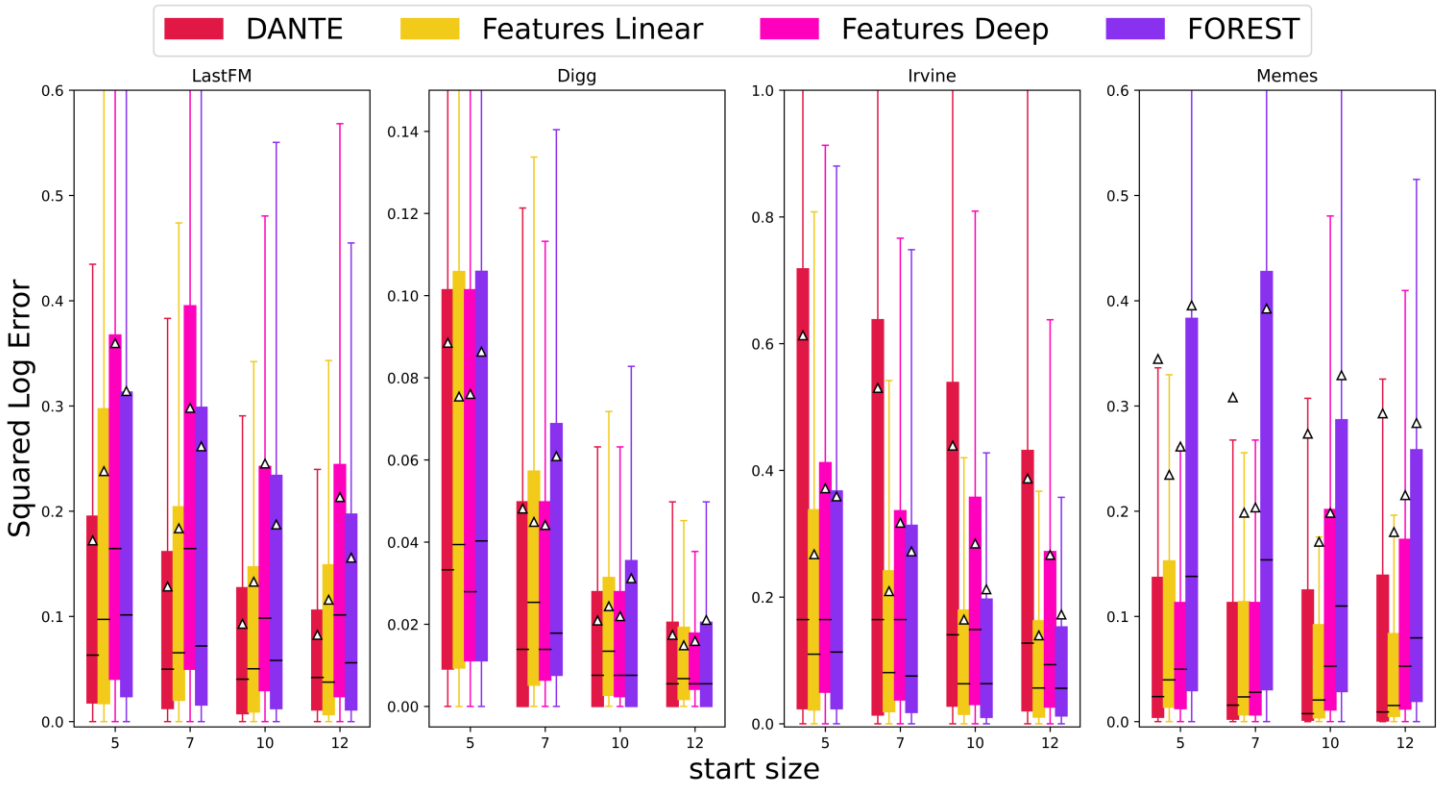
For a given dataset, we trained a single model for all possible t_c and Δt.



A single ADMM iteration involves user-level optimization sub problems running in parallel (using Dask).

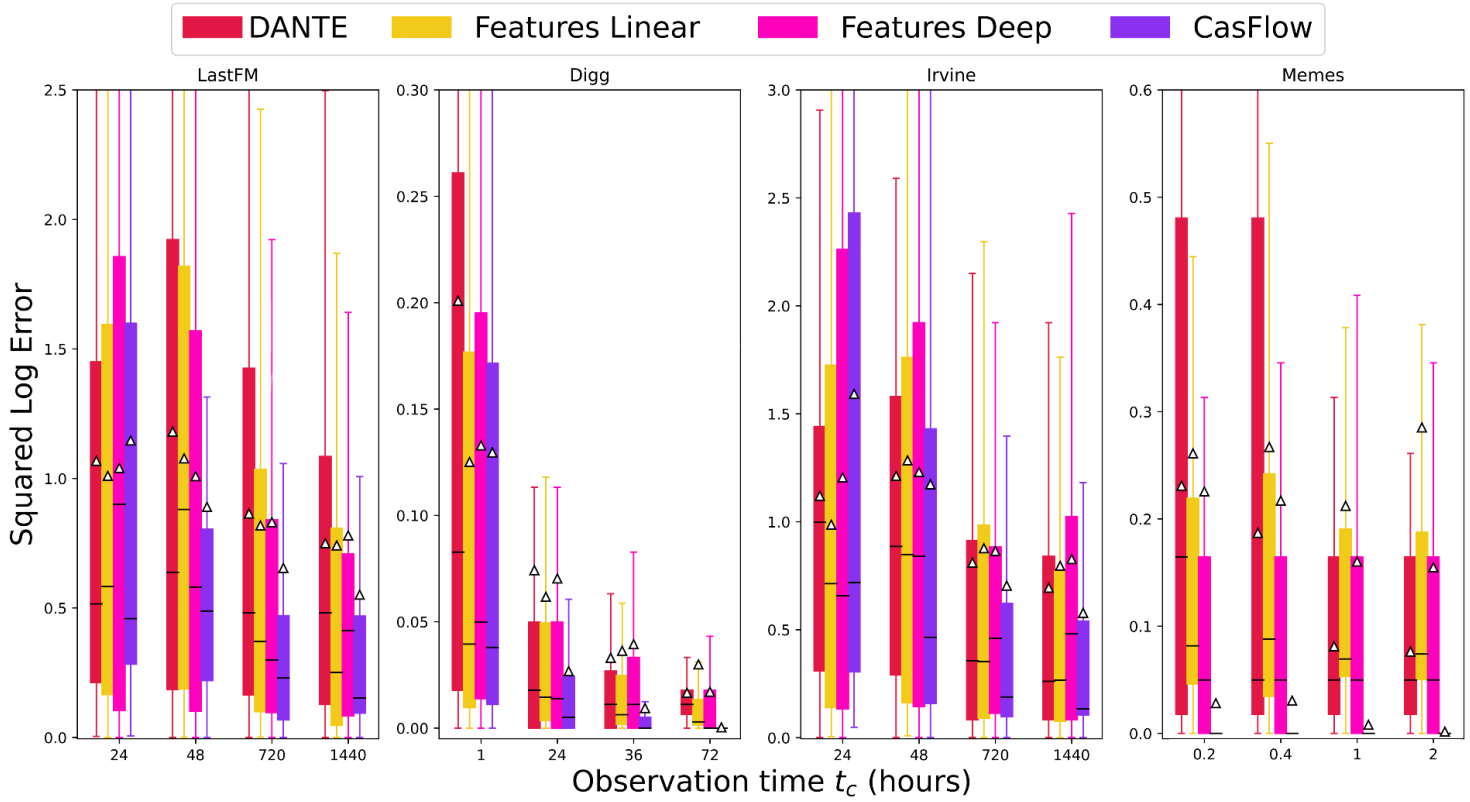
Results: Changing Observation Sizes

- Final size prediction after having observed a fixed number of users.
- For LastFM, Memes and Digg, DANTE outperforms in terms of median and performs comparably in terms of mean.



Results: Changing Observation Times

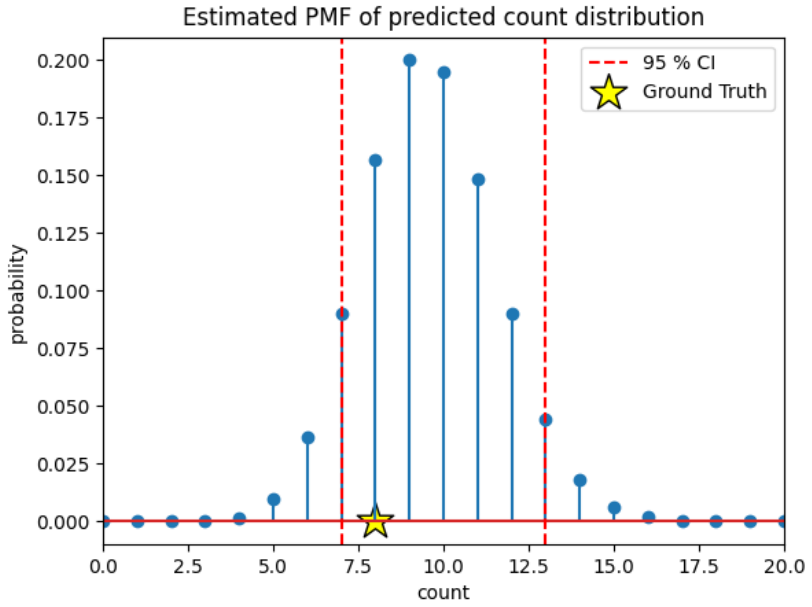
- Here we compare predictive performance against models after having observed events for a fixed amount of time t_c.
- With the exception of CasFlow, DANTE is highly competitive with all other methods.

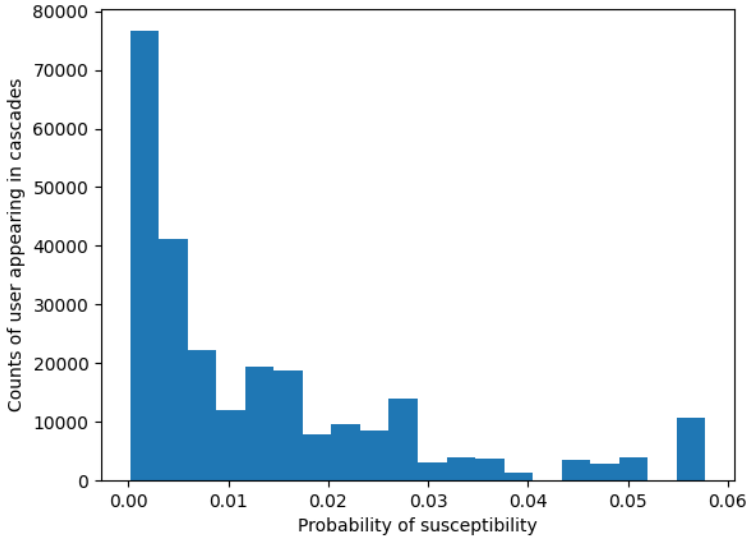


Interpretability

Our probabilistic approach affords us the following benefits.

- We can produce estimated count prediction intervals.
- We can see how user features affect the probability of susceptibility through its weight vector.
- We can answer any specific question about the counts if it is framed as a probabilistic query.





Conclusions

1. We proposed a single discriminative probabilistic model for all observation times t_c and prediction time intervals Δt which performed competitively against state-of-the-art methods that require training per $(t_c, \Delta t)$ pair.
2. Our probabilistic approach renders an interpretable model that allows us to produce estimated count prediction intervals.
3. Future work can seek to model other functional, perhaps non-parametric forms of the hazard function.

Thanks for listening!

[Schmidt and White 1989] Schmidt, Peter, and Ann Dryden Witte. 'Predicting Criminal Recidivism Using "Split Population" Survival Time Models'. *Journal of Econometrics*, vol. 40, no. 1, 1989, pp. 141–159, https://doi.org10.1016/0304-4076(89)90034-1.

[Boyd et al. 2011] Stephen Boyd; Neal Parikh; Eric Chu; Borja Peleato; Jonathan Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* , now, 2011.

- Code and poster can be found in the following GitHub repository: https://github.com/aaravamudan2014/DANTE

- Visit our website for more information on our point process research.

SCAN ME

Personal website: https://aaravamudan2014.github.io/Akshay-Aravamudan/