UTS

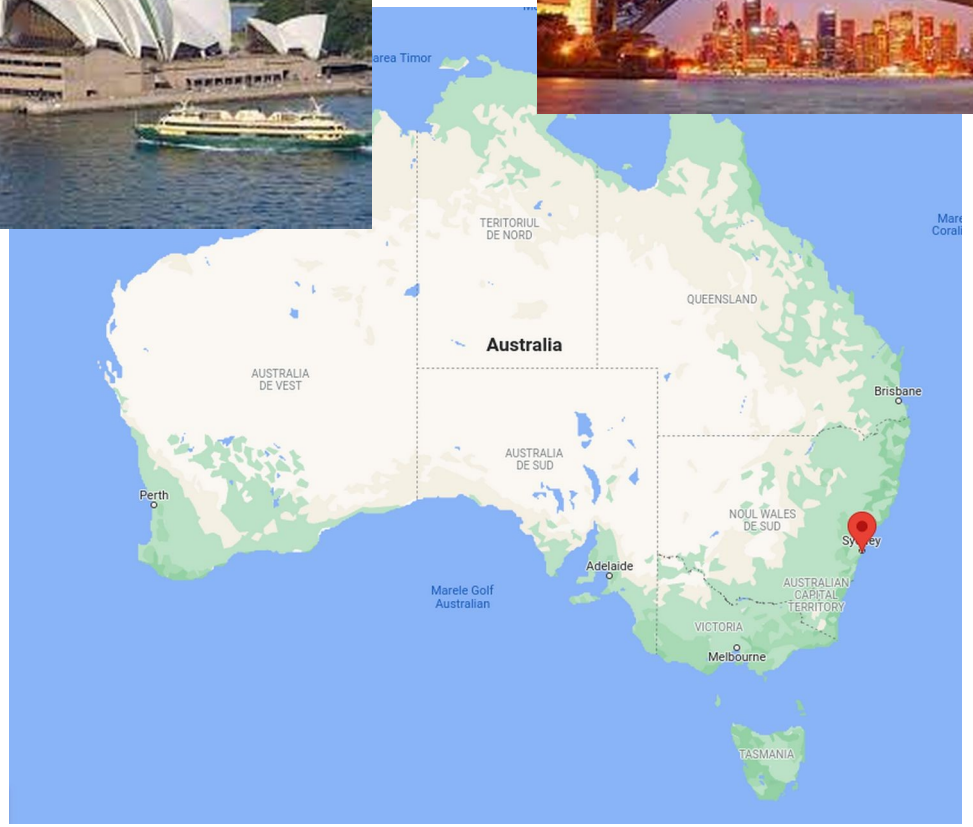
Behavioral Data Science

Interval-censored Transformer Hawkes

Detecting Information Operations using the Reaction of Social Systems

Data Science Institute

Dr Marian-Andrei Rizoiu | Behavioral Data Science
Marian-Andrei.Rizoiu@uts.edu.au
https://www.behavioral-ds.science



Located in Sydney, Australia

A city campus, iconic brutalist style blended with modern buildings

The research group

Behavioral Data Science

2 PostDocs, 7 PhD, 1 Masters, 1 assistant prof.







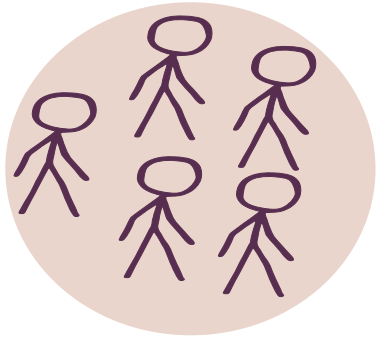




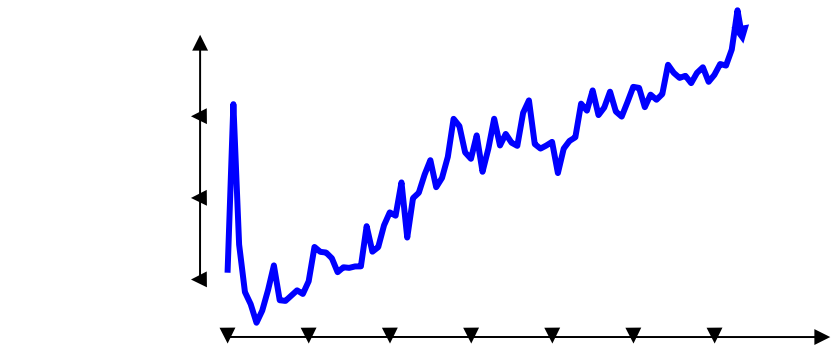
The Behavioral Data Science

Behavioral Data Science

1.



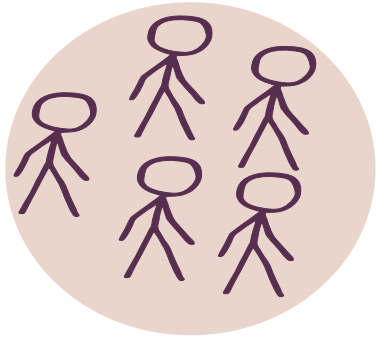
information diffusion
epidemics spreading
behavioral modeling



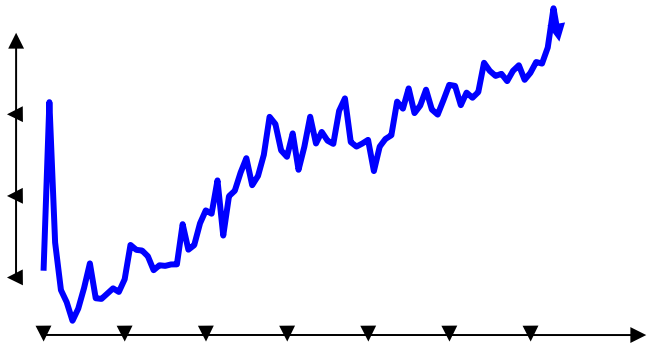
The Behavioral Data Science

Behavioral Data Science

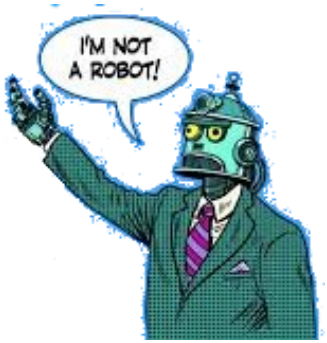
1.



information diffusion
epidemics spreading
behavioral modeling



2.



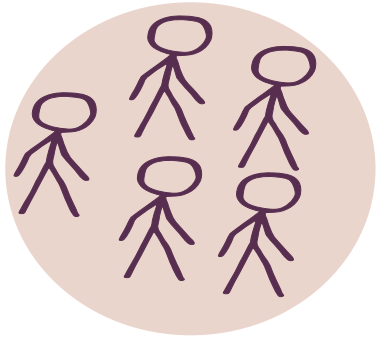
[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]

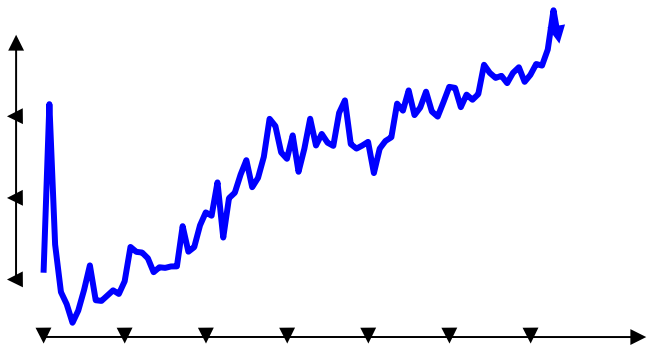
The Behavioral Data Science

Behavioral Data Science

1.

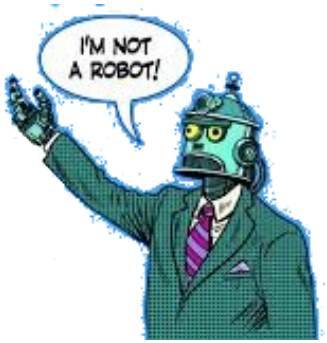


information diffusion
epidemics spreading
behavioral modeling



3.

2.



[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]

Our founders in the mis-, dis-, IO and IW spaces

Australian Government
Department of Defence
Defence Science and Technology Group

Real-time detection of disinformation campaigns

Australian Government
Department of Home Affairs

Information integrity initiative: fighting misinformation in Australia

DEFENCE INNOVATION NETWORK
To go far, go together

INFORMATION WARFARE DIVISION
AUSTRALIAN DEFENCE FORCE

Effectiveness of Information Operations in the Pacific

Australian Government
Department of Defence
Defence Science and Technology Group

Information Warfare STaR Shot "Developing Situational Awareness"

Hate Speech propagation on Social Media

Interval-censored Transformer Hawkes: Detecting Information Operations using the Reaction of Social Systems

Quyu Kong
Alibaba Group &
University of Technology Sydney
Hangzhou, China
kongquyu.kqy@alibaba-inc.com

Pio Calderon
University of Technology Sydney
Sydney, Australia
pio.calderon@student.uts.edu.au

Rohit Ram
University of Technology Sydney
Sydney, Australia
rohit.ram@student.uts.edu.au

Olga Boichak
University of Sydney
Sydney, Australia
olga.boichak@sydney.edu.au

Marian-Andrei Rizoiu
University of Technology Sydney
Sydney, Australia
marian-andrei.rizoiu@uts.edu.au

ABSTRACT

Social media is being increasingly weaponized by state-backed actors to elicit reactions, push narratives and sway public opinion. These are known as Information Operations (IO). The covert nature of IO makes their detection difficult. This is further amplified by missing data due to the user and content removal and privacy requirements. This work advances the hypothesis that the very reactions that Information Operations seek to elicit within the target social systems can be used to detect them. We propose an Interval-censored Transformer Hawkes (IC-TH) architecture and a novel data encoding scheme to account for both observed and missing data. We derive a novel log-likelihood function that we deploy together with a contrastive learning procedure. We showcase the performance of IC-TH on three real-world Twitter datasets and two learning tasks: future popularity prediction and item category prediction. The latter is particularly significant. Using the retweeting timing and patterns solely, we can predict the category of YouTube videos, guess whether news publishers are reputable or controversial and, most importantly, identify state-backed IO agent and behavior (inauthentic coordination) to elicit reactions and sway public opinion [29]. Content- and user-based detection tools are notoriously difficult to build due to language nuances and suffer from language drift and adversarial attacks. Here, we start from the hypothesis that the actions of IO actors are designed to elicit particular reactions from the target audience. This makes analyzing the posting behavior of users and the reaction of the social systems surrounding them a helpful starting point to map actors and agendas behind those posts. This paper presents a novel approach to model the reactions of social systems, to disentangle the structural and functional types of users from the content posted by them. Based solely on the timing of retweet cascades, we can separate YouTube videos' categories, quantify news publishers' trustworthiness, and even identify users involved in state-linked coordinated operations. The ubiquitous diffusion cascades generated in the online environment, such as reshare events on Twitter and replies on Reddit, unveil the temporal dynamics of online items [20, 38] and have been shown to be a proxy for characterizing them [15]. A typical difficulty when analyzing reshare cascades is partial data loss —

Kong, Q., Calderon, P., Ram, R., Boichak, O., & Rizoiu, M.-A. (2023). Interval-censored Transformer Hawkes: Detecting Information Operations using the Reaction of Social Systems. In ACM Web Conference (WWW'23) (pp. 1–9). https://doi.org/10.1145/3543507.3583481

Who are our online opinion leaders?

Common traits:

- Pro-republican;
- Highly influential, highly followed and retweeted;
- Opinion leaders;
- …

Who are our online opinion leaders?

Common traits:

- Pro-republican;
- Highly influential, highly followed and retweeted;
- Opinion leaders;
- …

Russian-controlled trolls operated by the Internet Research Agency in St. Petersburg

Information Operations

"computational propaganda [..] use of algorithms, automation, and human curation to purposefully distribute misleading information over social media networks"
[Woolley & Howard, 2018]

(defense env.) Information operations includes [..] the dissemination of propaganda in pursuit of a competitive advantage over an opponent.

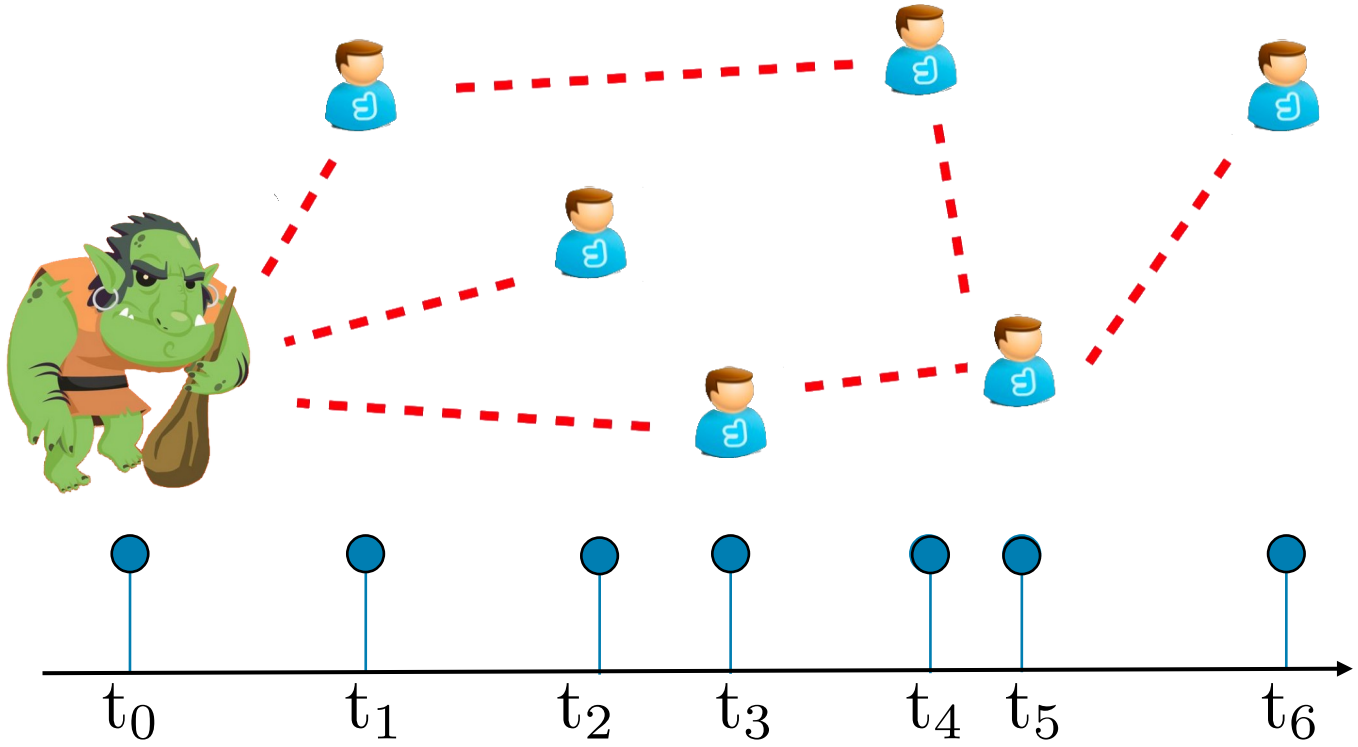
vs.

Challenge: beyond content-based detectors

Content- and user-based detection tools:
language nuances, language drift and adversarial attacks

IO are designed to elicit particular reactions from the target audience

RQ1:
Can we distinguish users and content types based on on the reaction of online social systems? **no content**



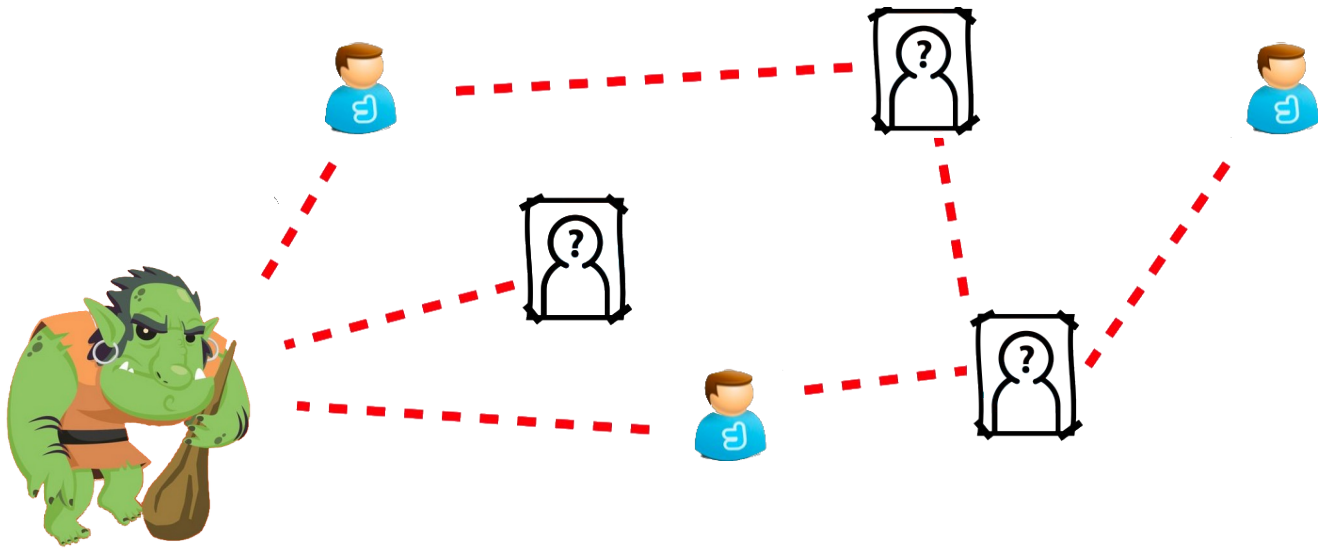
Challenge: partial missing data

Missing tweets

Restrictions from Twitter API [Wu et al, 2020]

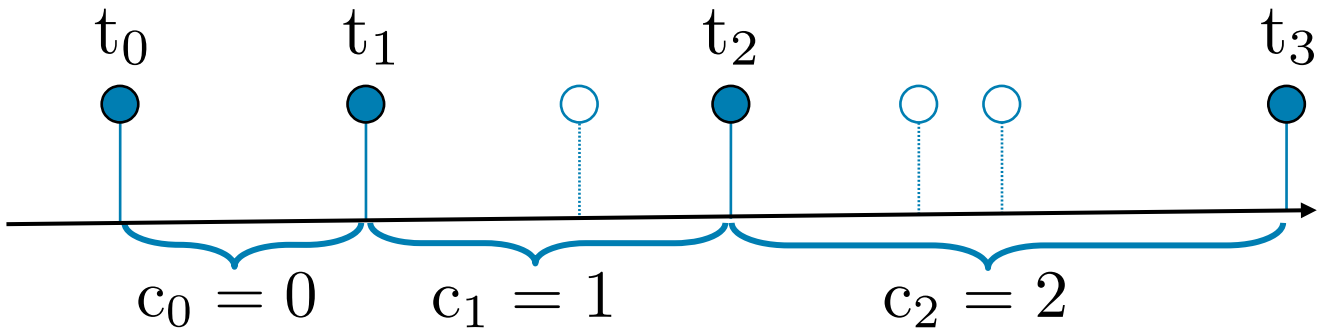
User moderation, content removal

Only event counts are available between events (via the retweet_count property)



RQ2:

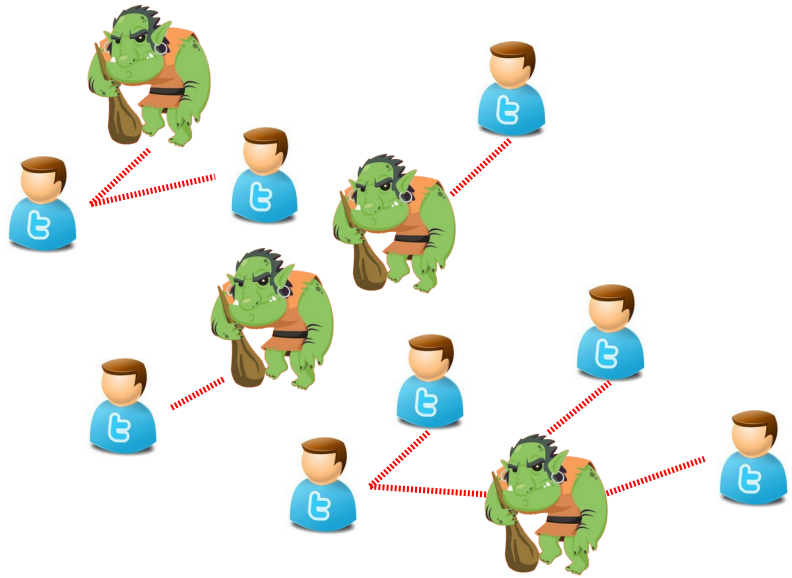
Can we model reshare cascades containing both event times and missing event counts?



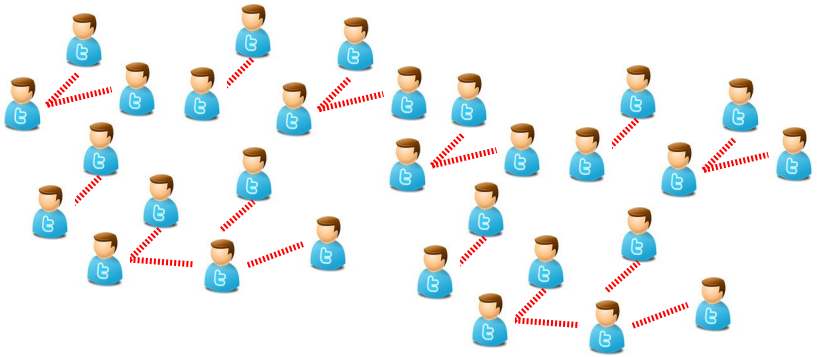
Challenge: (very) limited labelled data

Very limited training data
- Covert nature of IO
- Costly human labelling

Abundant amount of unlabelled data
- Public datasets
- APIs



Labelled Cascades

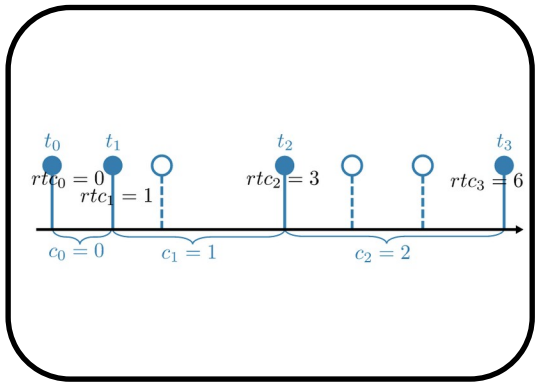


Unlabelled Cascades

RQ3:
Can we use (large amounts of) unlabelled data to pretrain representations?

Presentation plan

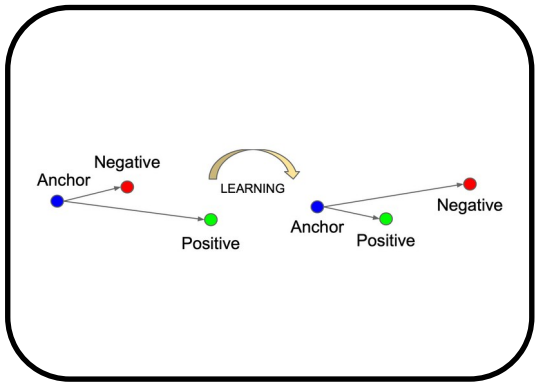
- Motivation and Challenges
- Interval-Censored Transformer Hawkes (IC-TH)



Unified representation times & event counts

$$\mathscr{L}_{\text{IC-TH-LL}}(\theta) = \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1}}_{\text{missing event counts}}$$
$$= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i$$

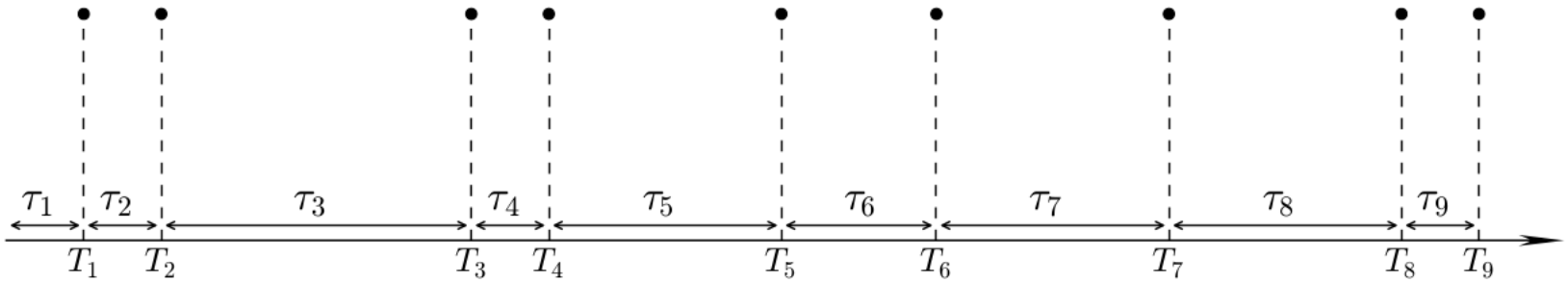
Novel log-likehood function



Model Pre-training via Contrastive learning

- Experiments and findings

Point process



A random process – a collection of random variables – the event times

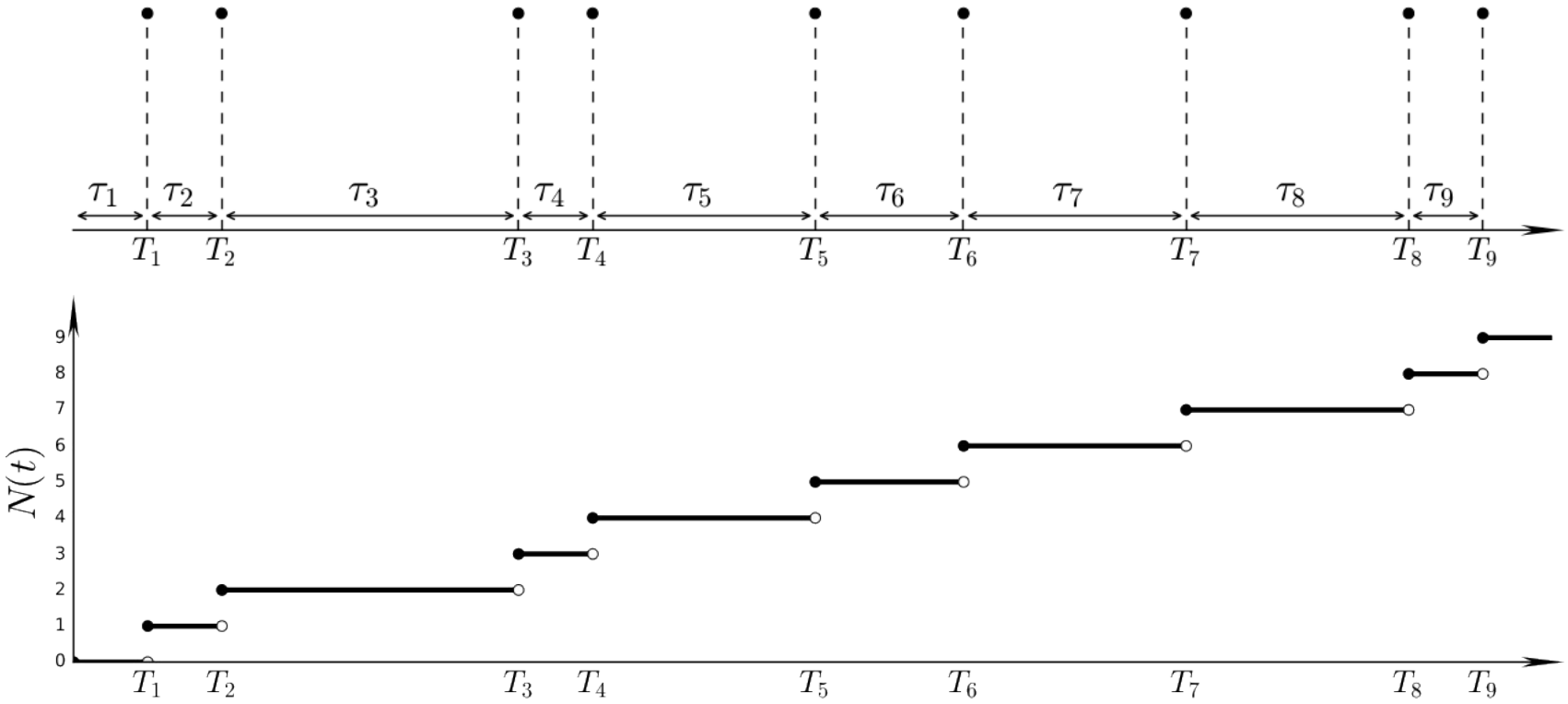
T_i event times

τ_i inter-arrival times

$$T_i := \sum_{j=1}^{i} \tau_j$$

Point process

$$N(t) := \sum_{i \geq 1} 1_{\{t \geq T_i\}}$$



Equivalent counting process. A random function defined on time t ≥ 0, takes integer values 1, 2, .. .
The number of events of the point process by time t

Homogeneous Poisson process

The inter-arrival times are exponentially distributed $\tau_i \sim Exp(\lambda)$

λ is the event intensity of the homogeneous Poisson process

$$f(\tau = t) = \lambda e^{-\lambda t}, \text{ for } t \geq 0$$

Memorylessness property: the probability of having to wait an additional t time units after already having waited m time units is the same as the probability of having to wait t time units when starting at time o

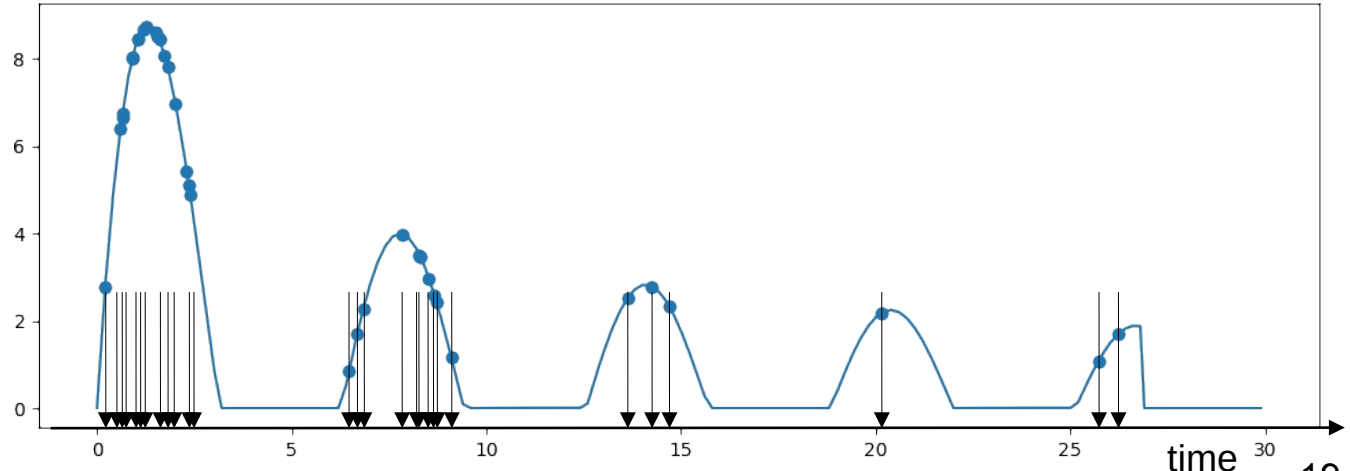
$$\mathbb{P}(\tau > t + m | \tau > m) = \mathbb{P}(\tau > t)$$

Inhomogeneous Poisson process

The rate of event arrivals is a function of time, i.e. $\lambda = \lambda(t)$

The conditional intensity function:

$$\lambda(t|\mathcal{H}_t) = \lim_{h \to 0} \frac{\{N_{t+h} - N_t = 1 | \mathcal{H}_t\}}{h} \qquad \mathcal{H}_t = \{T_1, T_2, ..., T_{N_t}\}$$



Inhomogeneous Poisson process

The rate of event arrivals is a function of time, i.e. $\lambda = \lambda(t)$

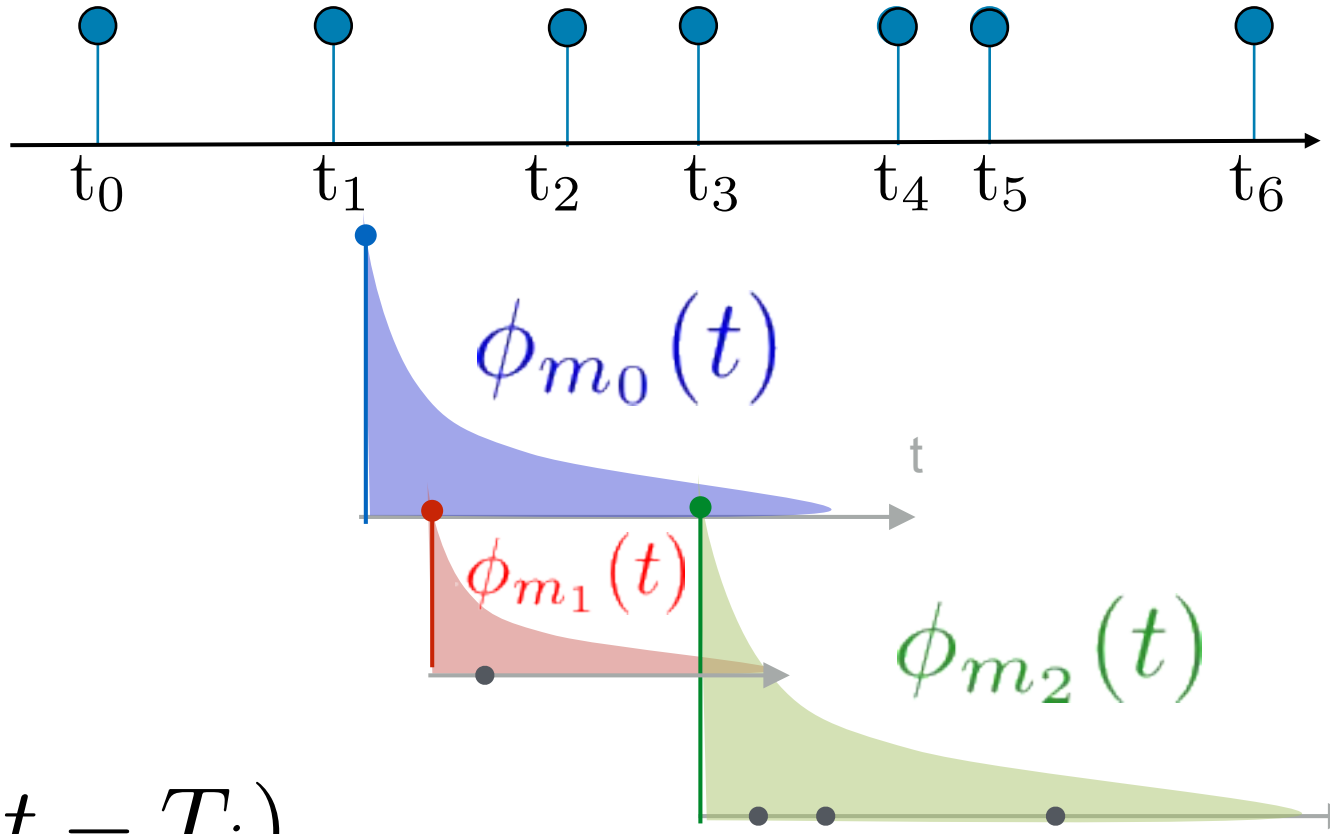
The conditional intensity function:

$$\lambda(t|\mathcal{H}_t) = \lim_{h \to 0} \frac{\{N_{t+h} - N_t = 1 | \mathcal{H}_t\}}{h} \qquad \mathcal{H}_t = \{T_1, T_2, ..., T_{N_t}\}$$

Proportional to the probability of observing an event at time t

$$\begin{aligned}
&\mathbb{P}(N_{t+h} = n + m \,|\, N_t = n) = \lambda(t)h + o(h) && \text{if} \quad m = 1 \\
&\mathbb{P}(N_{t+h} = n + m \,|\, N_t = n) = o(h) && \text{if} \quad m > 1 \\
&\mathbb{P}(N_{t+h} = n + m \,|\, N_t = n) = 1 - \lambda(t)h + o(h) && \text{if} \quad m = 0
\end{aligned}$$

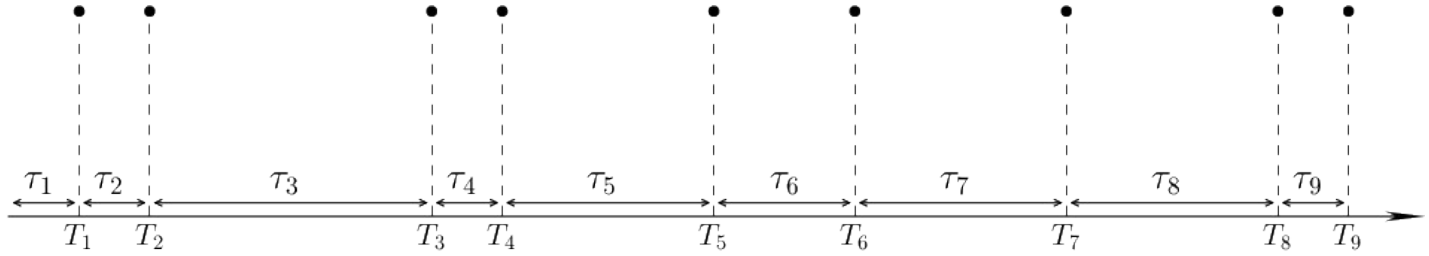
Self-exciting (Hawkes) processes [Hawkes, 1971]



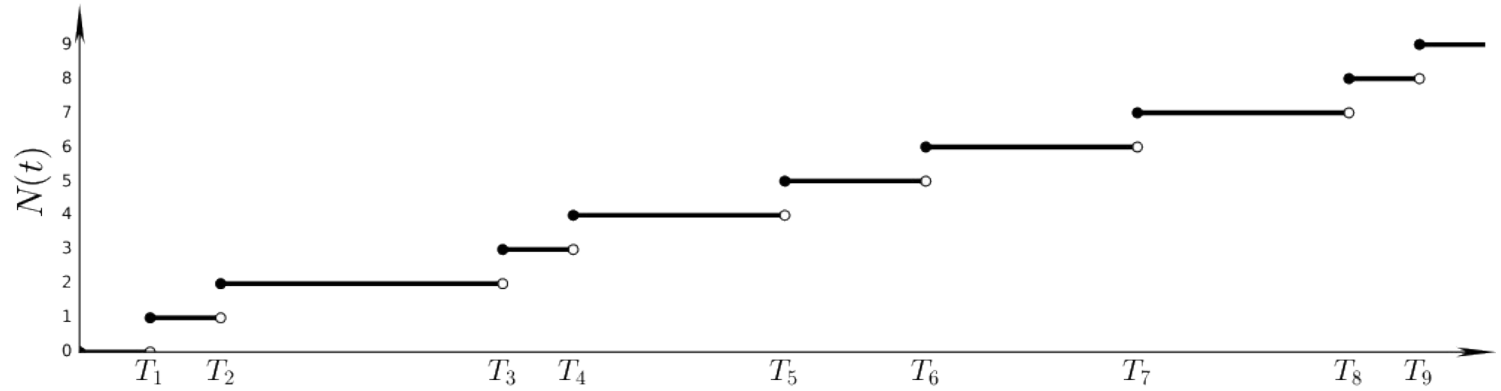
Event intensity

$$\lambda(t|\mathcal{H}_t) = \underbrace{\mu(t)}_{\text{base intensity (exogenous)}} + \underbrace{\sum_{i:t>T_i} \phi(t - T_i)}_{\text{self-excitation (endogenous)}}$$

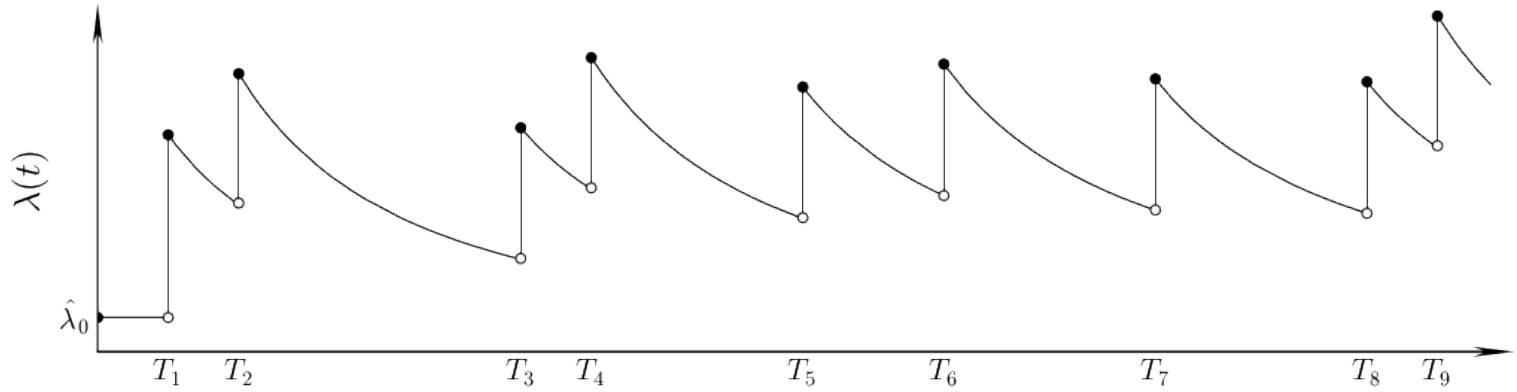
Hawkes process – definition



event times T_i



counting process $N(t)$

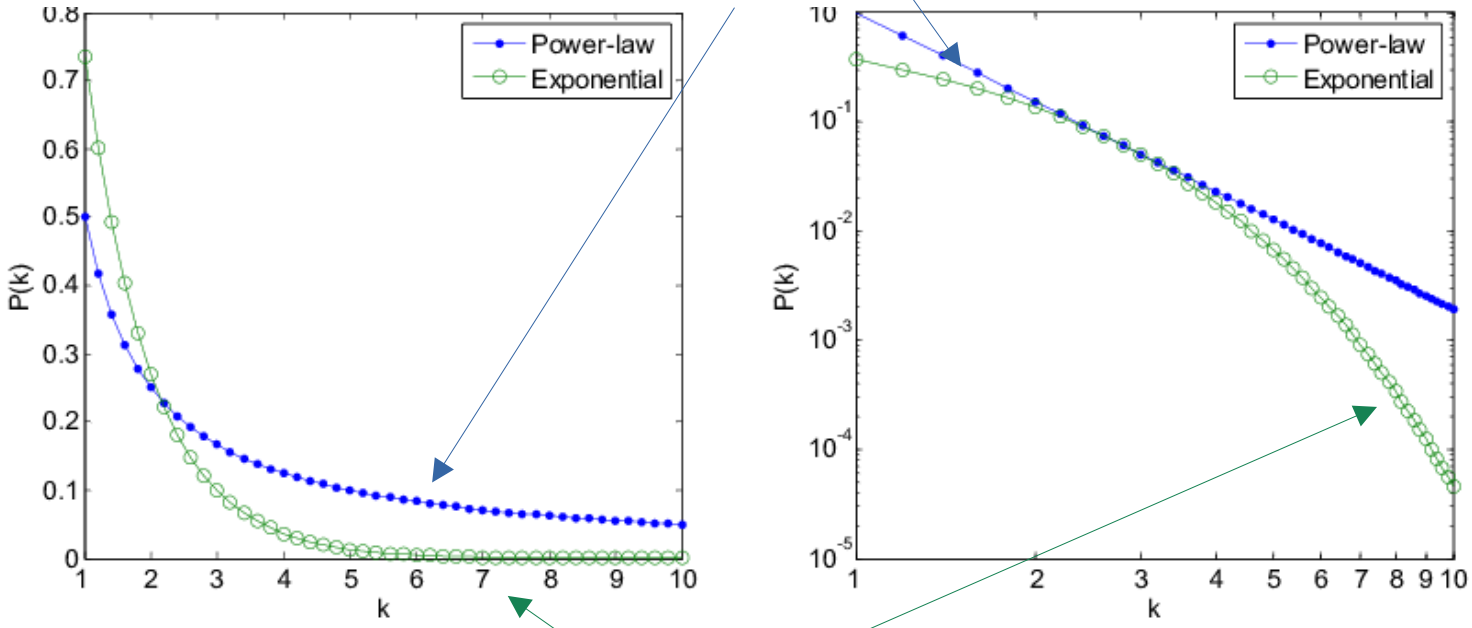


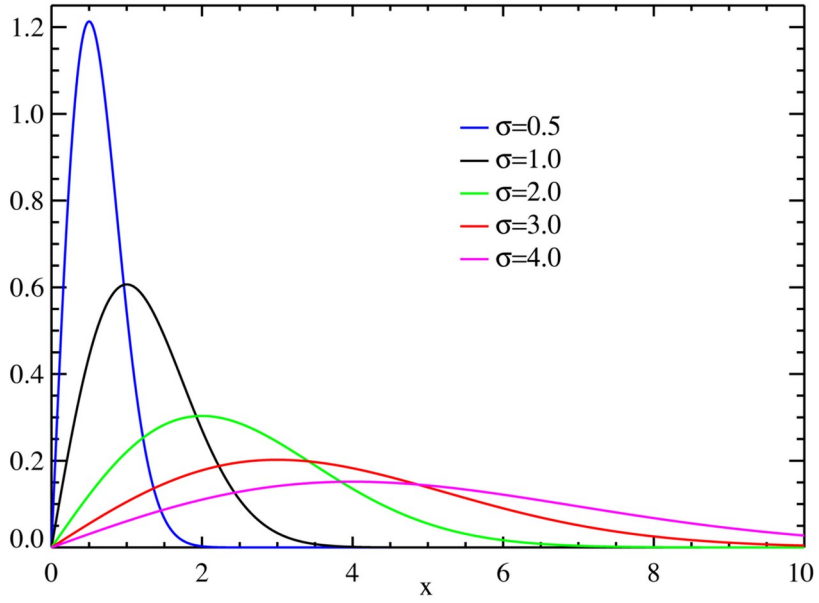
conditional intensity function $\lambda(t|\mathcal{H}_t)$

Hawkes process – choice of kernel

Common choices: Exponential, Power-Law, Rayleigh

Power-law has
a long tail.





Exponential
decays very fast.

The “Twitter” kernel

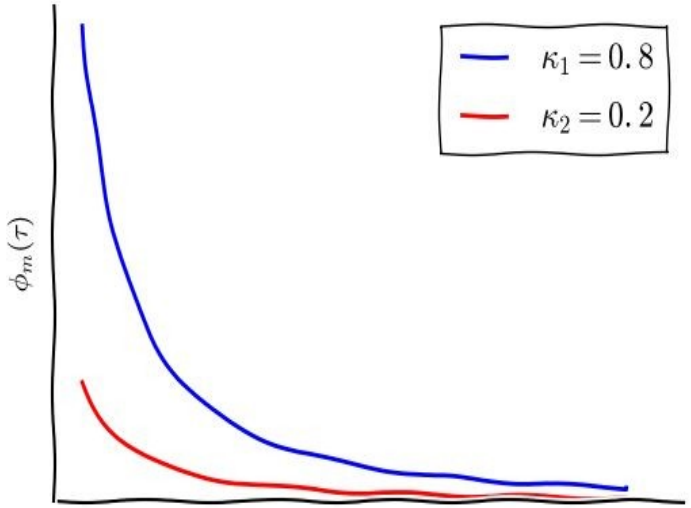
$$\phi_m(t - T_i) = \kappa m^{\beta} (t - T_i)^{-(1+\theta)}$$

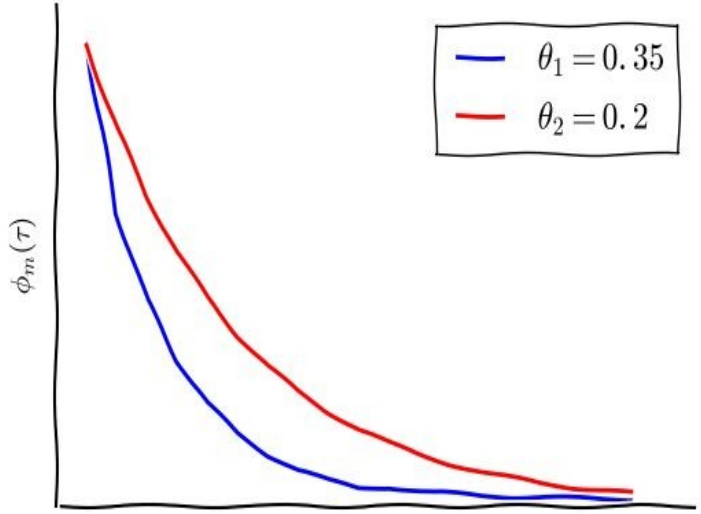
the rate of ‘daughter’ events

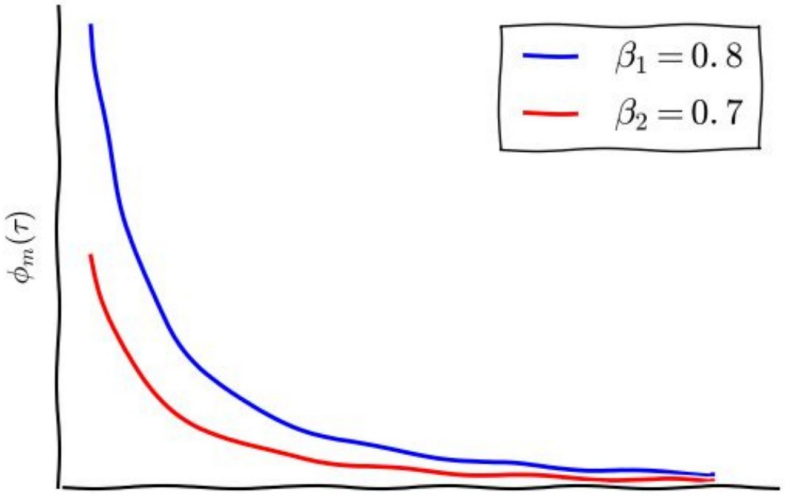
content virality

user influence

memory







S. Mishra, M.-A. Rizoiu, and L. Xie, "Feature Driven and Point Process Approaches for Popularity Prediction, " in Proc. International Conference on Information and Knowledge Management (CIKM ’16), Indianapolis, USA, p. 1069-1078, 2016.

Transformer Hawkes [Zuo et al, 2021]

Event intensity

Softplus function

$$\lambda(t) = f(w^\top h(t))$$

Hidden-state

Multi-head self-attention module

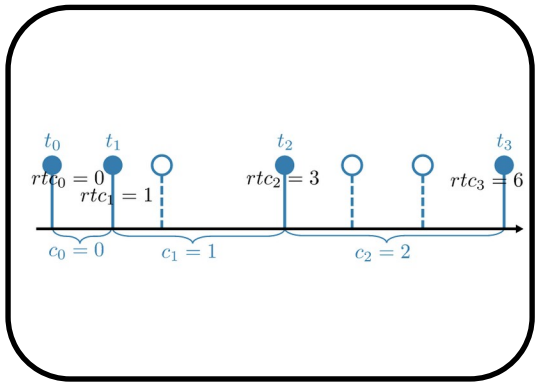
$$h(t_j) = H(j,:)$$

$$H(j,:) = ReLU(SW_1 + b_1)W_2 + b_2$$

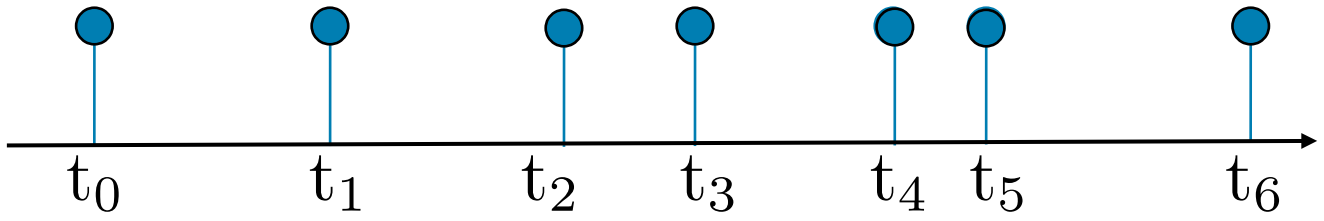
$$S = Concat(head_1, head_2, \cdots)W^O$$

$$head_i = Softmax\left(\frac{XW_i^Q(XW_i^K)^\top}{\sqrt{d_k}}\right)XW_i^V$$

IC-TH: a mixed data format



Hawkes & Transformer Hawkes



$$\mathcal{H} = \{t_0, t_1, \ldots\}$$

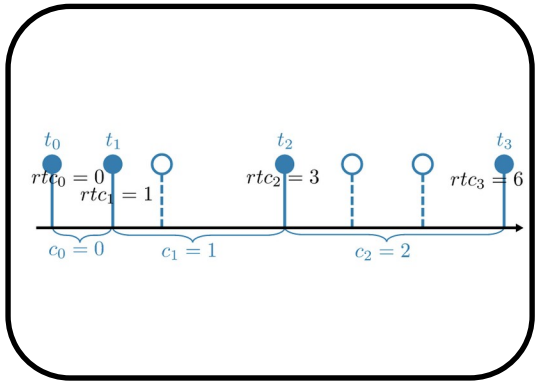
duration

$$\mathcal{H}^* = \{(t_i, d_i, c_i)\}$$

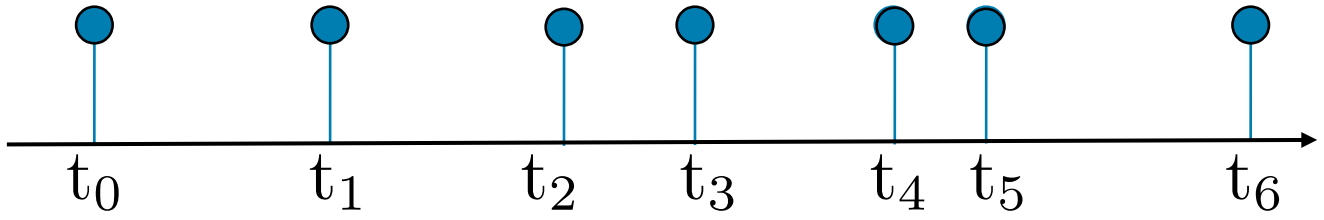
interval start time

event count

IC-TH: a mixed data format



Hawkes & Transformer Hawkes



$$\mathcal{H} = \{t_0, t_1, \ldots\}$$

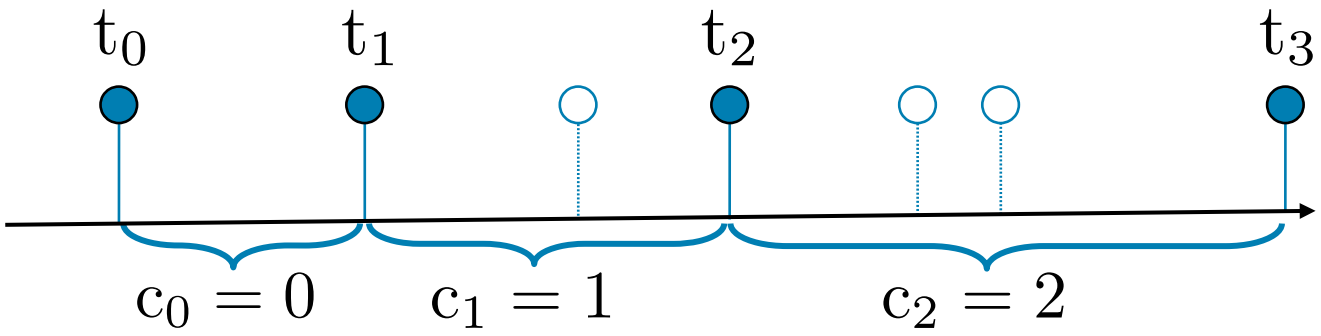
duration

$$\mathcal{H}^* = \{(t_i, d_i, c_i)\}$$

interval start time

event count

Interval-Censored Tranformer Hawkes



$$\mathcal{H}^* = \{(t_0 - dt, 2dt, 1), (t_0, t_1 - t_0, 0), (t_1 - dt, 2dt, 1), (t_1, t_2 - t_1, 1), \cdots\}$$

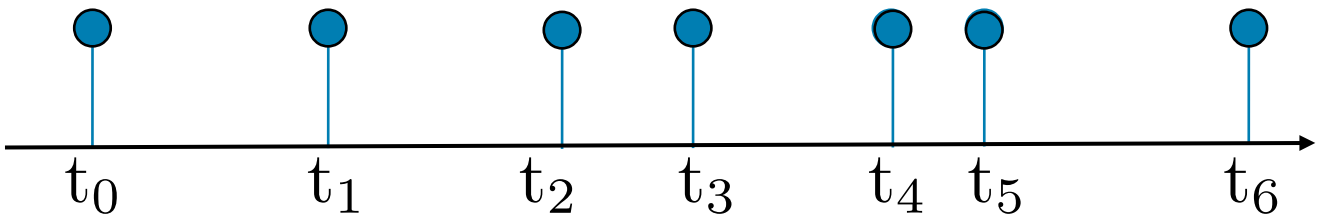
IC-TH: Novel log-likehood function

$$\mathscr{L}_{\text{IC-TH-LL}}(\theta) = \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1}}_{\text{missing event counts}}$$

$$= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i$$

Hawkes process

$$\log L(\theta) = \sum_{i=1}^{N(t)} \log \lambda(t_i) - \int_0^t \lambda(\tau) d\tau$$

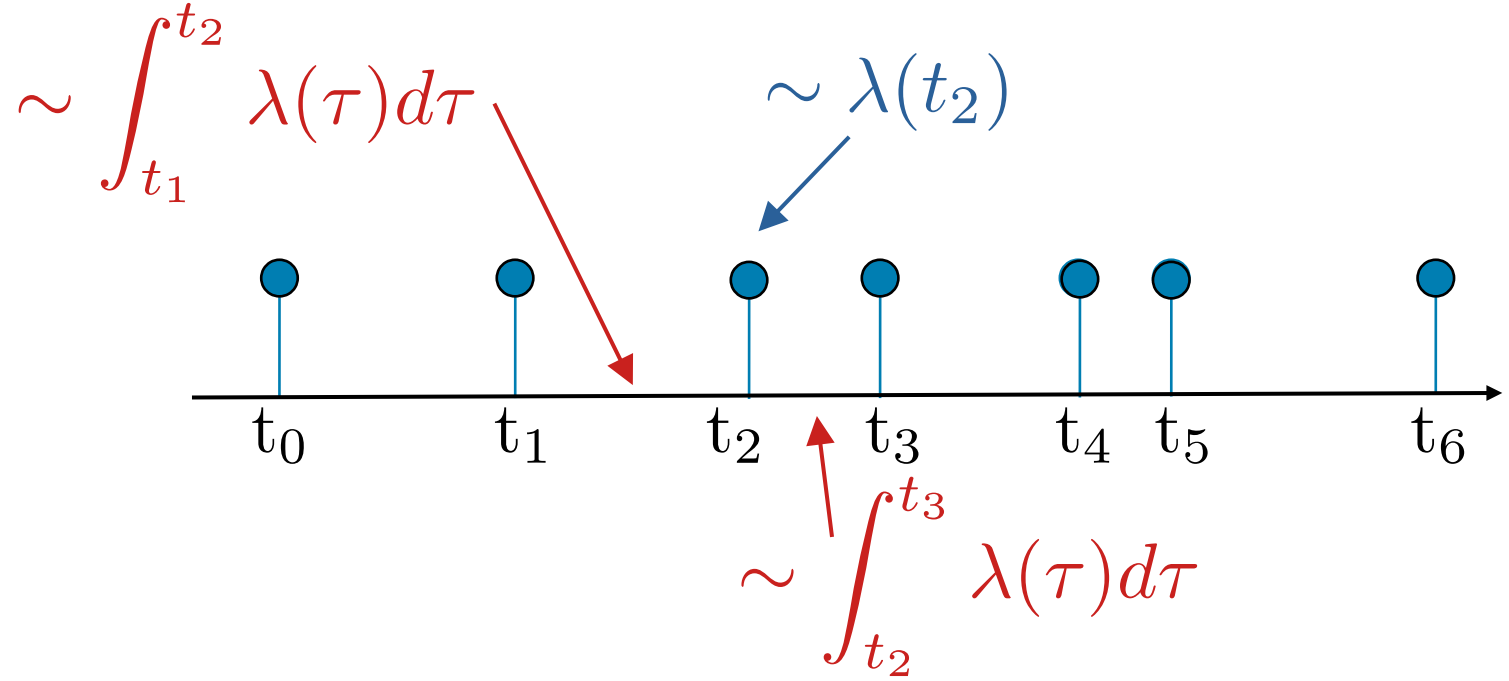


IC-TH: Novel log-likehood function

$$\mathscr{L}_{\text{IC-TH-LL}}(\theta) = \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1}}_{\text{missing event counts}}$$
$$= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i$$

Hawkes process

$$\log L(\theta) = \sum_{i=1}^{N(t)} \log \lambda(t_i) - \int_0^t \lambda(\tau) d\tau$$

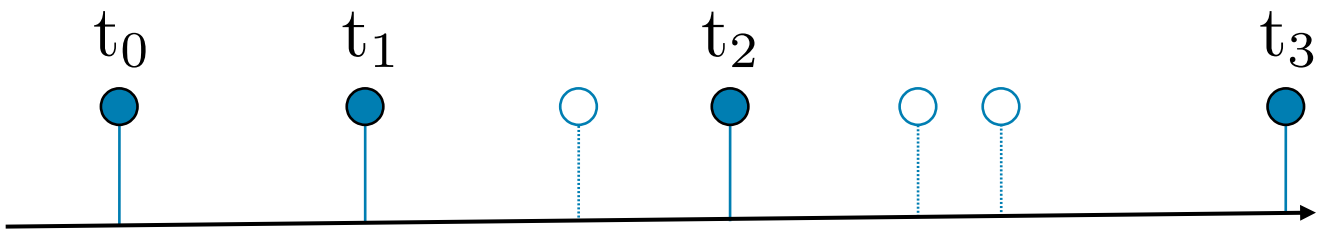


IC-TH: Novel log-likehood function

$$\mathscr{L}_{\text{IC-TH-LL}}(\theta) = \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1}}_{\text{missing event counts}}$$
$$= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i$$

Interval-Censored Transformer Hawkes

$$\log L(\theta) = \sum_{i \in \mathcal{H}_u^*} c_i \log \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau + \sum_{i \in \mathcal{H}_c^*} \log \lambda(t_i) - \sum_{i \in \mathcal{H}^*} \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau$$

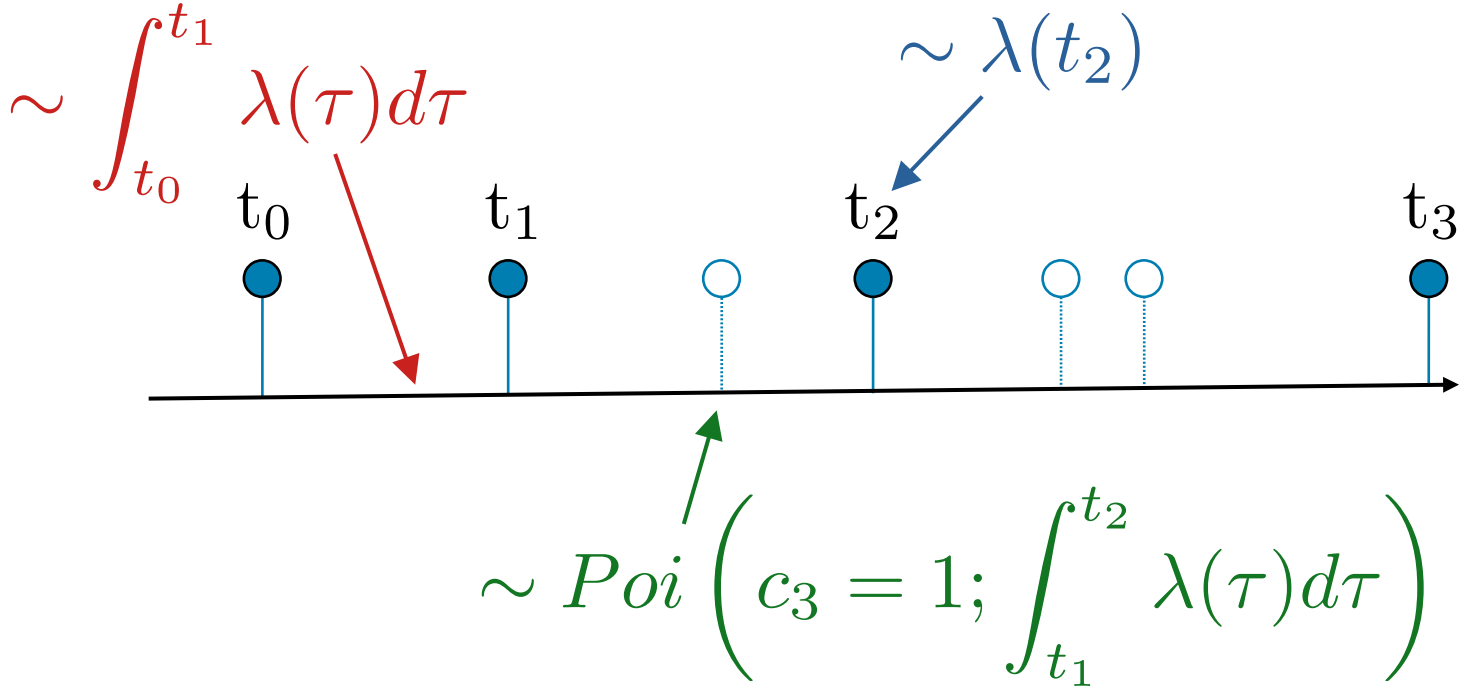


IC-TH: Novel log-likehood function

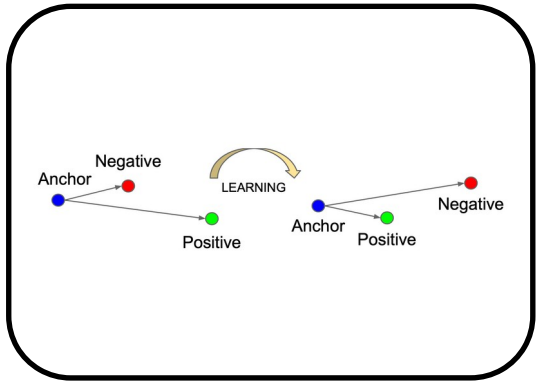
$$\mathscr{L}_{\text{IC-TH-LL}}(\theta) = \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1}}_{\text{missing event counts}}$$
$$= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i$$

Interval-Censored Transformer Hawkes

$$\log L(\theta) = \sum_{i \in \mathcal{H}_u^*} c_i \log \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau + \sum_{i \in \mathcal{H}_c^*} \log \lambda(t_i) - \sum_{i \in \mathcal{H}^*} \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau$$



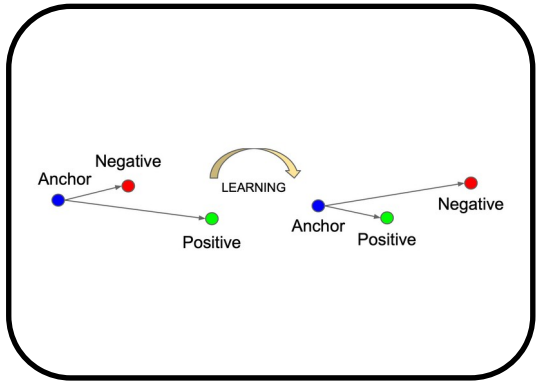
IC-TH: contrastive learning



On a large, unlabeled dataset "contrast samples against each other to learn attributes that are common between data classes and attributes that set apart a data class from another."

Build representations that distinguish users based on the cascades they appear in

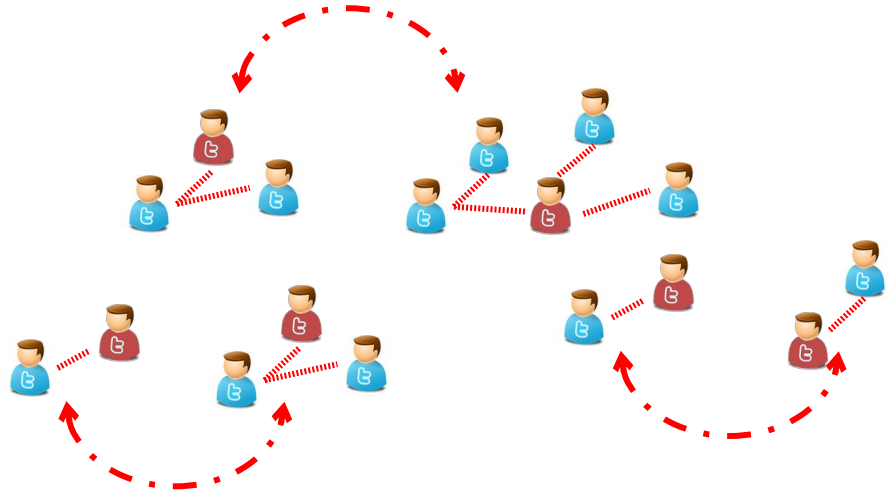
IC-TH: contrastive learning



On a large, unlabeled dataset "contrast samples against each other to learn attributes that are common between data classes and attributes that set apart a data class from another."

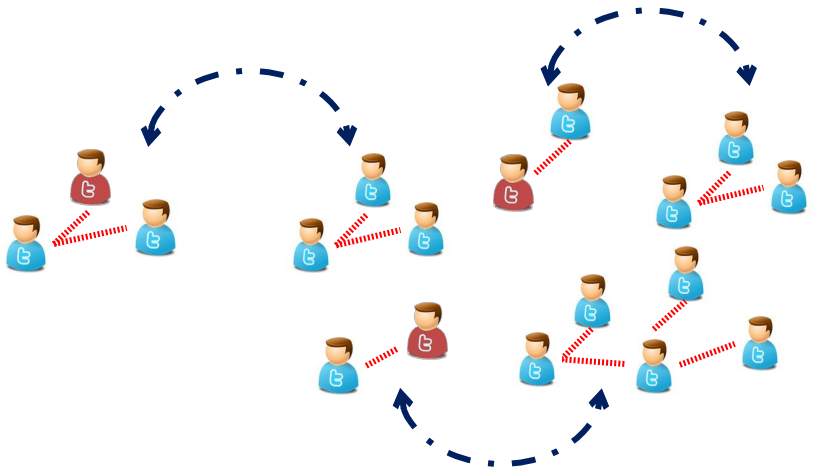
Positive pairs

cascades in which a given user participates



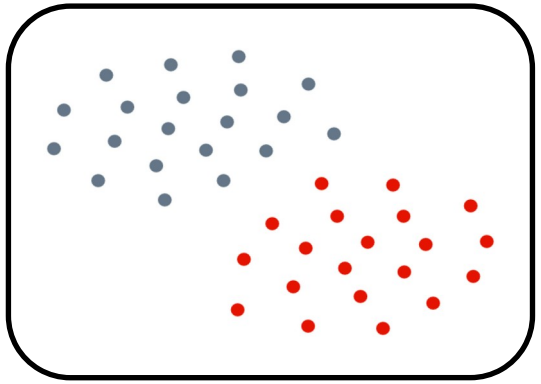
Negative pairs

given user only appears in one cascade



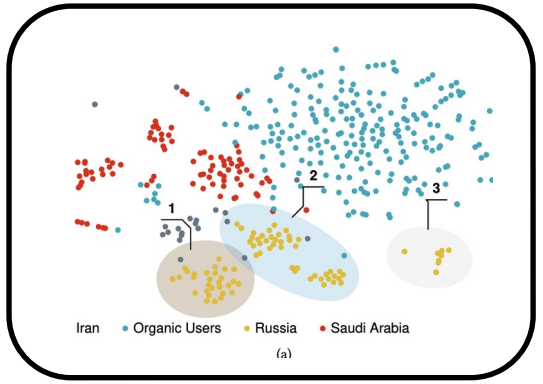
Presentation plan

- Motivation and Challenges
- Interval-Censored Transformer Hawkes (IC-TH)
- Experiments and findings



Effects of data loss

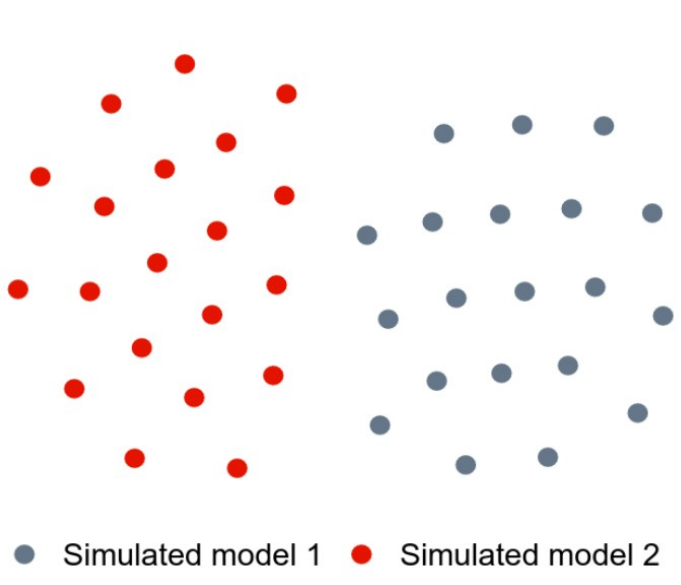
Category prediction

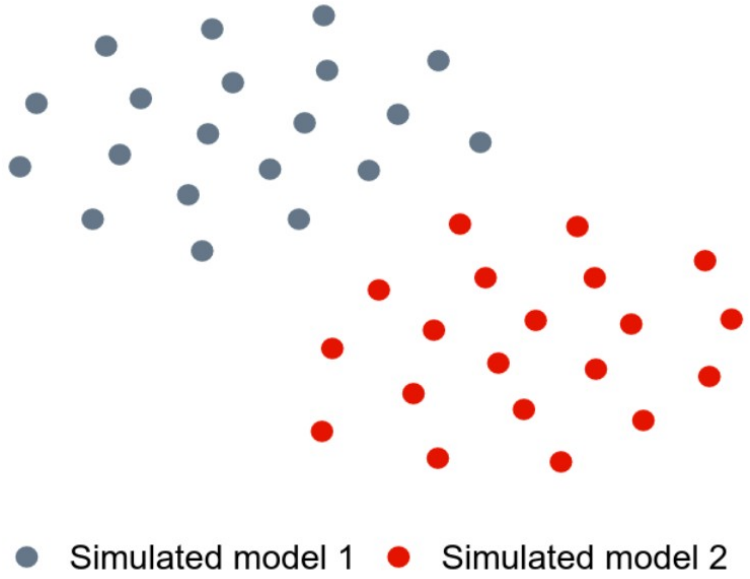


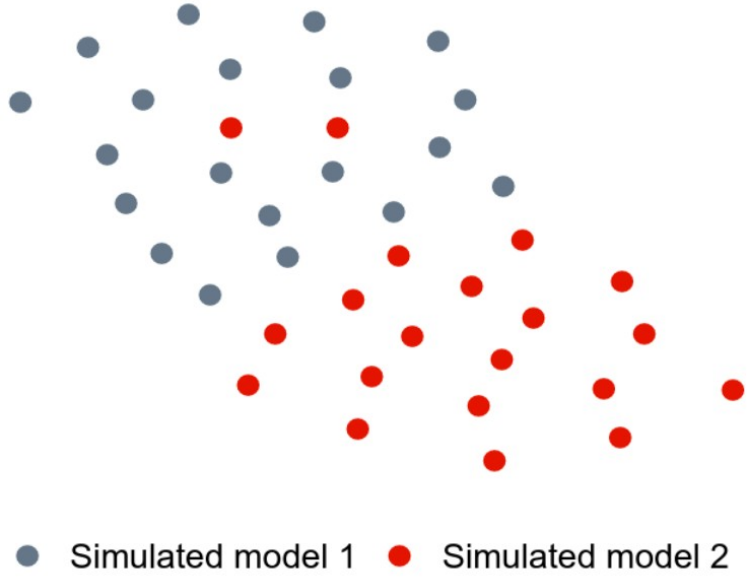
Analysis of Information Operations Dataset

Synthetic XP: effect of data loss

2 models (PowerLaw and Exponential), 10,000 cascades per model, 20 groups of 500 cascades







Missing events: 0% 50% 90%

IC-TH is robust with data loss; achieves near-perfect separability even at very large data sampling rates (90%).

Dataset: Twitter Moderation Research Consortium (TMRC) Information Operations dataset

Manipulation that Twitter can reliably attribute to a government or state linked actor – an information operation. [Twitter, TMRC]

Profiling:

Nov 2010 to Aug 2020
32,486 users
22,845,053 tweets
19,476,766 cascades

Classes – states sponsoring IO:

Russia, Iran, Saudi Arabia, Organic users

Dataset: Reputable&controversial news sources, tweeted YouTube

RNCNIX:
102,429 articles
56,397,252 tweets
8,129,126 cascades

Classes – states sponsoring IO:
Reputable, controversial

ActiveRT2017:
75,717 videos
85, 334, 424 tweets
30,535,891 cascades

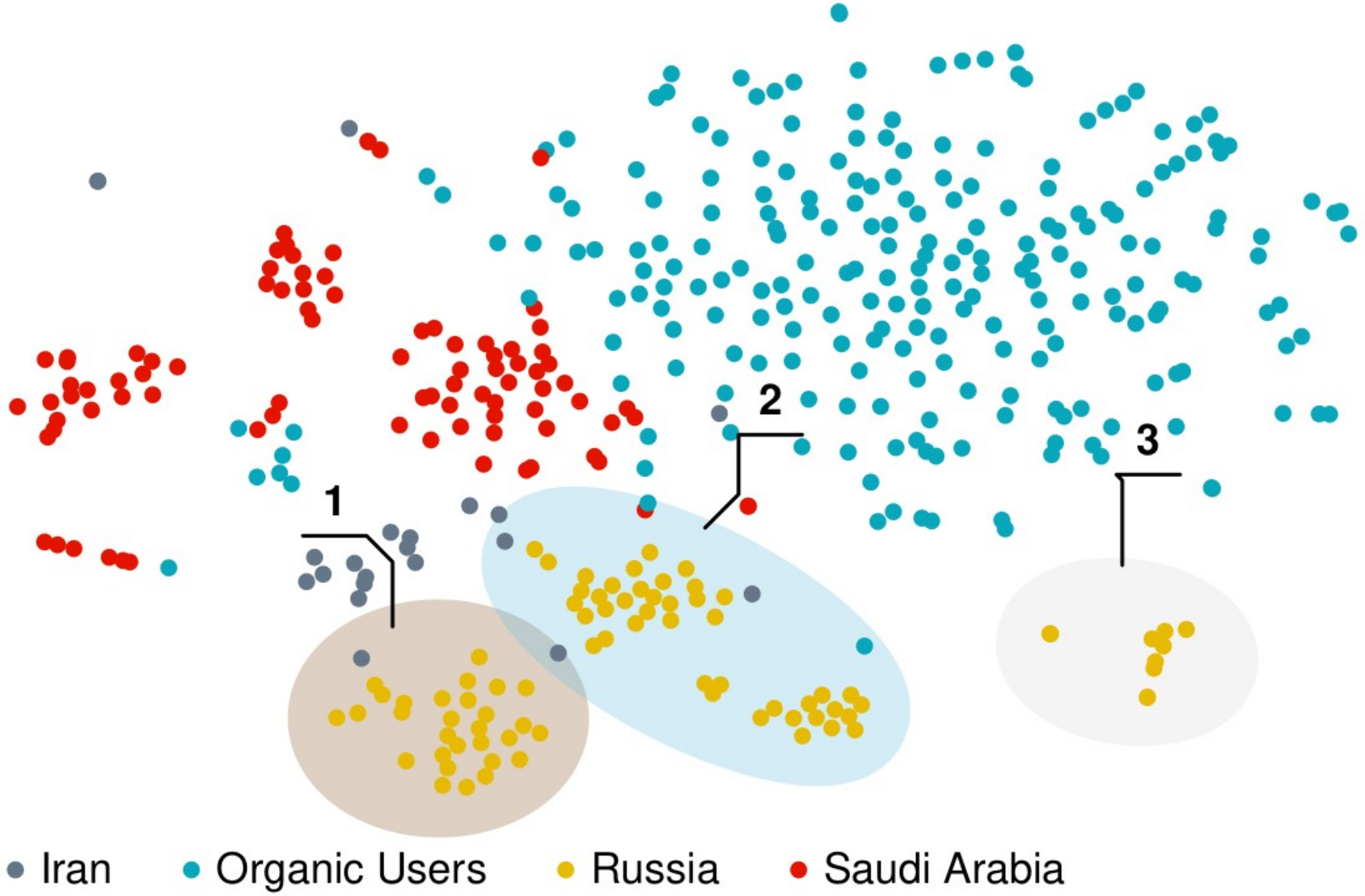
Classes – states sponsoring IO:
Entertainment, Gaming, Music and News&Politics

Predicting the category of content

Models			
Discrete Mixture Models	0.488	0.675	0.968
Transformer Hawkes [Kong et al, 2020]	0.469	0.823	0.983
IC-TH w/o missing counts [Zuo et al, 2021]	0.495	0.840	0.985
IC-TH	0.499	-	**0.987**
Pre-trained IC-TH	**0.503**	**0.853**	**0.987**

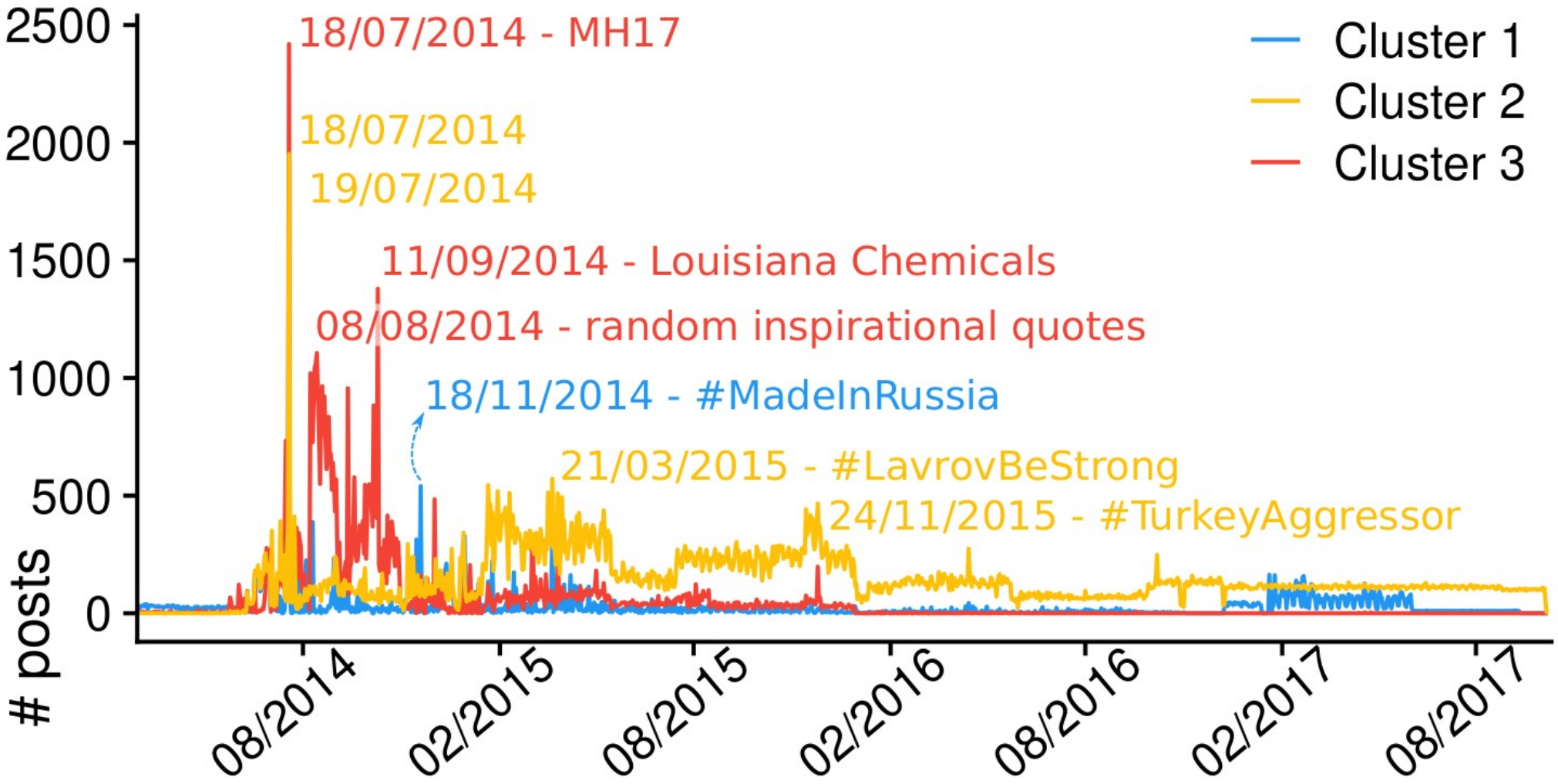
IC-TH outperforms all baselines;
Mixed data format + loglikelihood contribute most to performance
Missing counts and pre-training lead to moderate performance increases.

Identifying agent types and coordinated behavior



IC-TH clusters IO agents from specific countries based solely on the timing of the cascades in which they participate; it identifies even individual “troll farms”.

Coordination among Russia-backed clusters



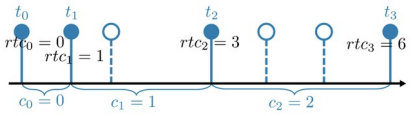
Qualitative investigations uncovers strategies of Russian trolls farms:
C1: Russian news with patriotic framing;
C2: Regional and conservative news;
C3: tweet in English, *#music*, *#usa*, relationship advice

Graphic representation of narratives in clusters C2 and C3

C2

C3

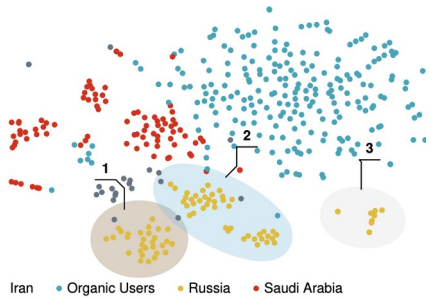
Conclusion



A unified representation and a novel log-likelihood for event times and missing events for the Transformer Hawkes architecture.

A contrastive learning approach that leverages large amounts of unlabeled data to build representations.

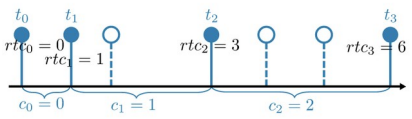
Predict the type of users and online content based solely on how the social systems react to them.



IC-TH reveals even “troll farms” – qualitative analysis reveals their strategies and roles, and the coordinated activity at strategic times.

Conclusion

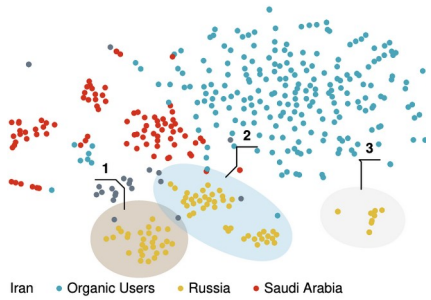
Thank you!



A unified representation and a novel log-likelihood for event times and missing events for the Transformer Hawkes architecture.

A contrastive learning approach that leverages large amounts of unlabeled data to build representations.

Predict the type of users and online content based solely on how the social systems react to them.



IC-TH reveals even “troll farms” – qualitative analysis reveals their strategies and roles, and the coordinated activity at strategic times.