



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>



Dr Olga Boichak | Digital Cultures  
olga.boichak@sydney.edu.au  
<http://olgarithmic.net/>



Located in Sydney, Australia



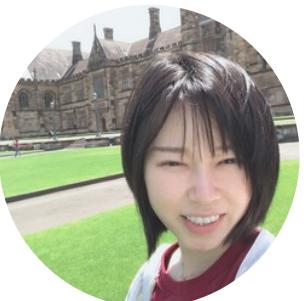
A city campus, iconic brutalist style  
blended with modern buildings

# The research group



# Behavioral Data Science

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

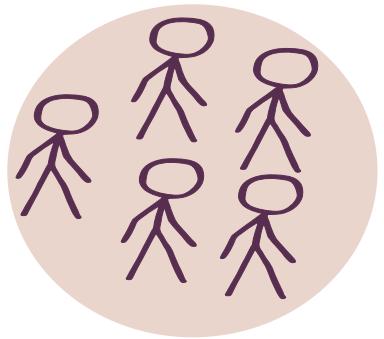


# The Behavioral Data Science

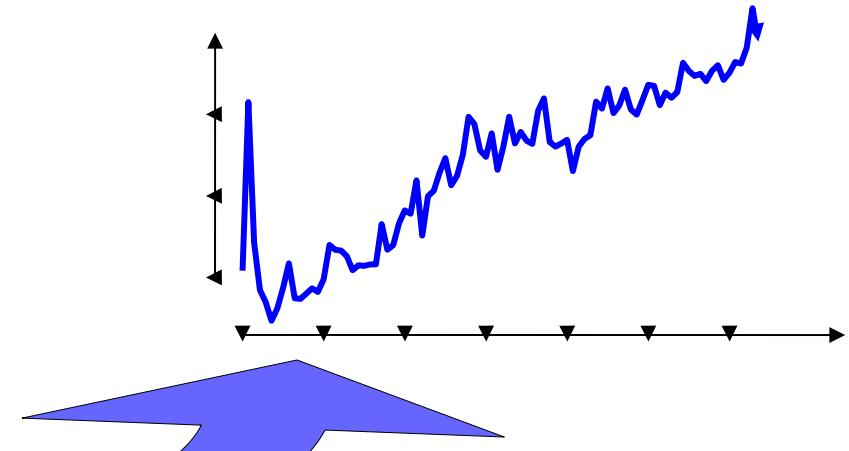


# Behavioral Data Science

1.



information diffusion  
epidemics spreading  
behavioral modeling

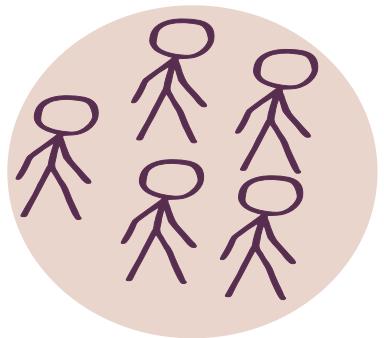


# The Behavioral Data Science

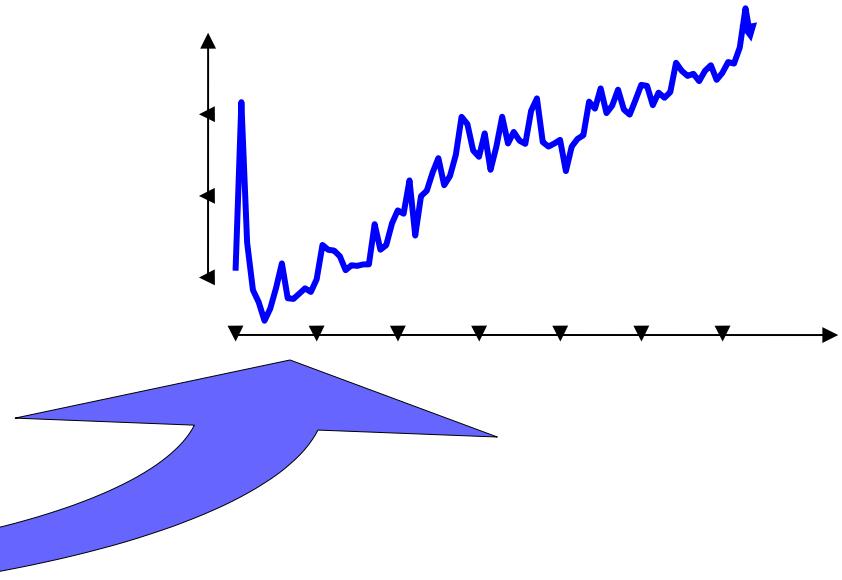


# Behavioral Data Science

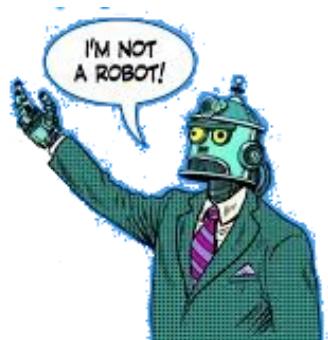
1.



information diffusion  
epidemics spreading  
behavioral modeling



2.



[Rizoiu et al ICWSM'18]

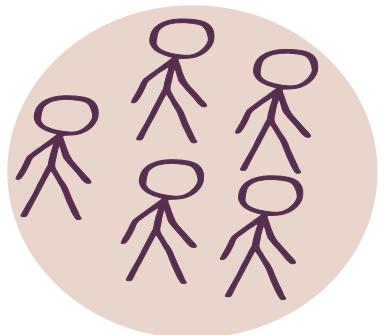
[Kim et al Journ.Comp.SocSci'19]

# The Behavioral Data Science

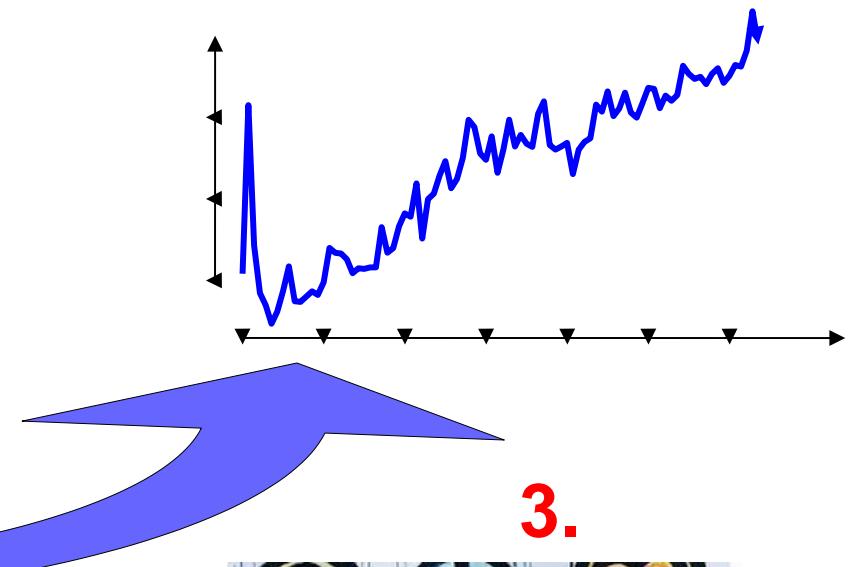


# Behavioral Data Science

1.



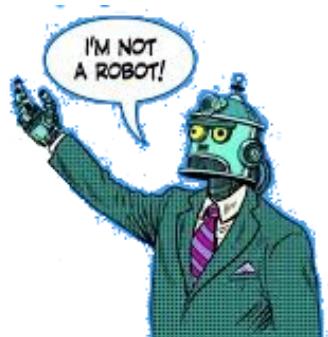
information diffusion  
epidemics spreading  
behavioral modeling



3.



2.



[Rizou 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



Information integrity initiative:  
fighting misinformation in Australia



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

## 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

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>

# 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

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

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

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

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



# Who are our online opinion leaders?



**Jenna Abrams**

@Jenn\_Abrams

Politics is a circus of hypocrisy. I DO care. Any offers/ideas/questions? DM or email me jennnabrams@gmail.com (Yes, there are 3 Ns, this is important)

📍 USA

🔗 [jennabrams.com](http://jennabrams.com)

📅 Joined October 2014

⌚ Born on October 02



**Tennessee GOP**

@TEN\_GOP

I love God, I Love my Country

📍 Tennessee, USA

📅 Joined November 2015

## Common traits:

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



**General Flynn**

@GenFlynn

Believe in #AmericanException  
It's for real. NYT bestselling au  
@fieldoffight (Read it) I stand i  
American principles & values.

📍 GLOBAL

🔗 [mikeflynndefensefund.org](http://mikeflynndefensefund.org)

📅 S-a alăturat în Ianuarie 2014

**Sebastian Gorka DrG**

@SebGorka

@FOXNews Nat. Sy. Strategist Bestseller: DEFEATING JIHAD [amzn.to/2zTuXyl](http://amzn.to/2zTuXyl) Fmr Strategist to Pres. Trump Order my NEW book: [amzn.to/2JkuSGJ](http://amzn.to/2JkuSGJ)

📍 Washington, DC

📅 S-a alăturat în martie 2014

# Who are our online opinion leaders?



**Jenna Abrams**

@Jenn\_Abrams

Politics is a circus of hypocrisy. I DO care. Any offers/ideas/questions? DM or email me jennabrams@gmail.com (Yes, there are 3 Ns, this is important)

📍 USA

🔗 [jennabrams.com](http://jennabrams.com)

📅 Joined October 2014

📍 Born on October 02



**Tennessee GOP**

@TEN\_GOP

I love God, I Love my Country

📍 Tennessee, USA

📅 Joined November 2015

## 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

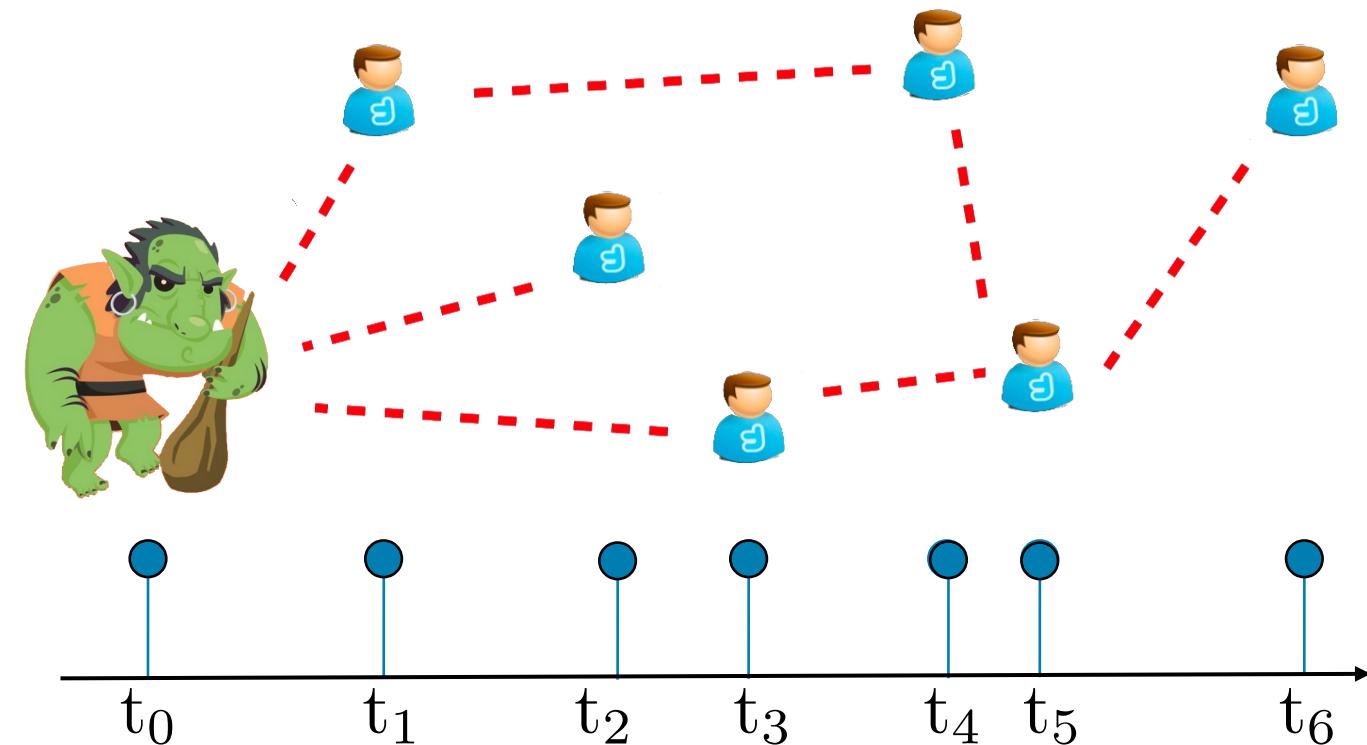
## Content- and user-based detection tools:

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 the reaction of online social systems? no content



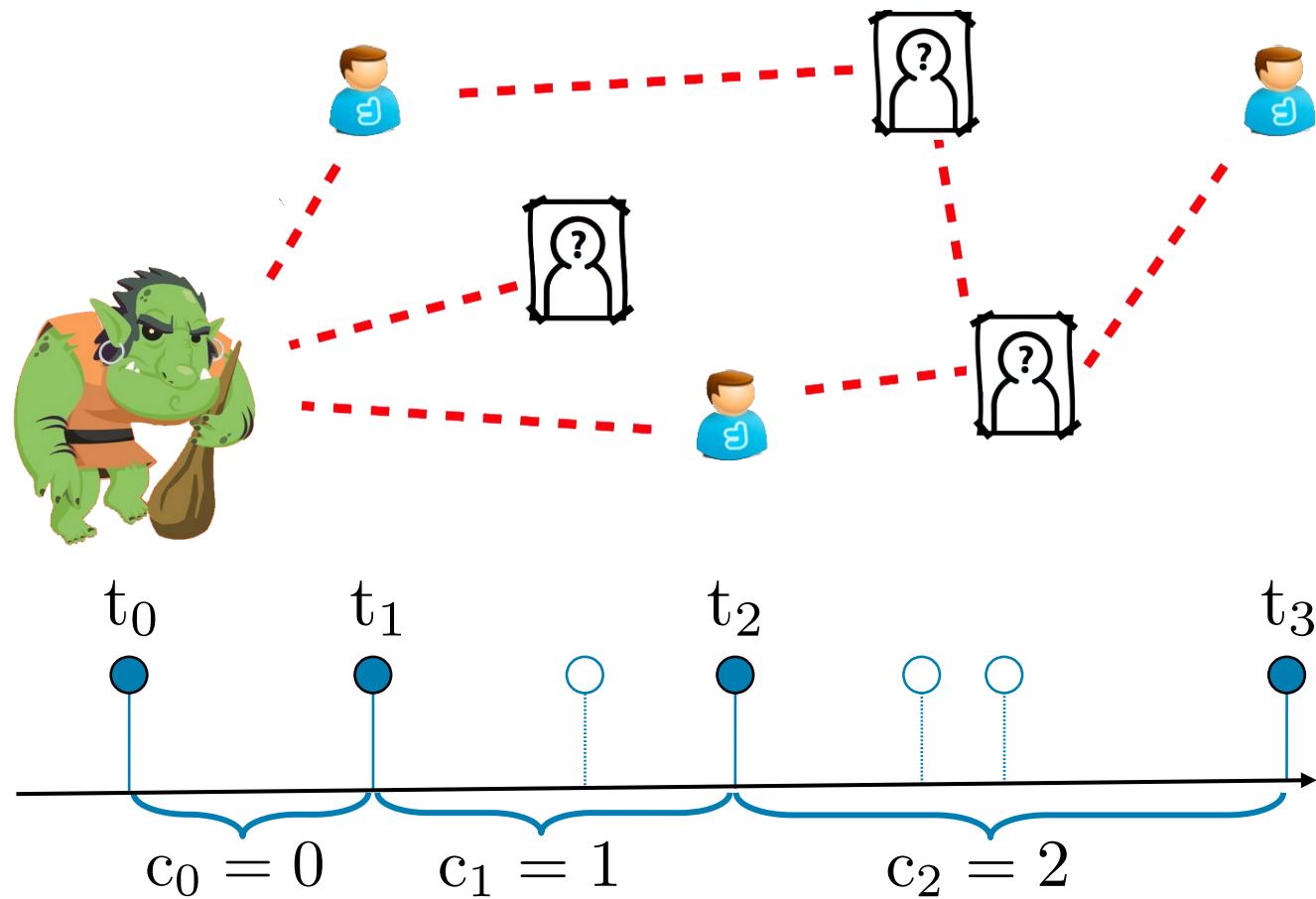
# Challenge: partial missing data

## 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

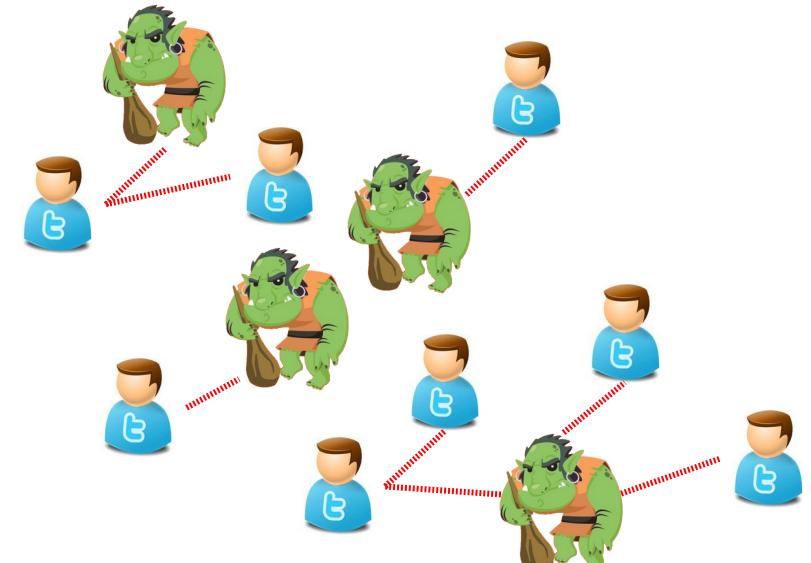
## Abundant amount of unlabelled data

Public datasets

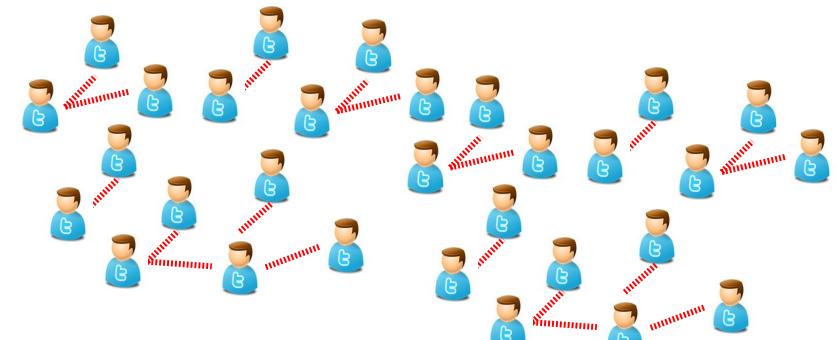
APIs

RQ3:

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



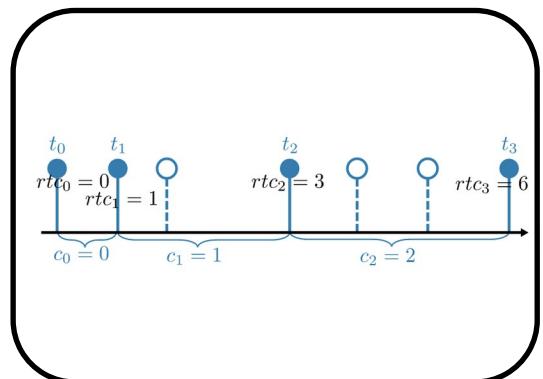
Labelled Cascades



Unlabelled Cascades

# Presentation plan

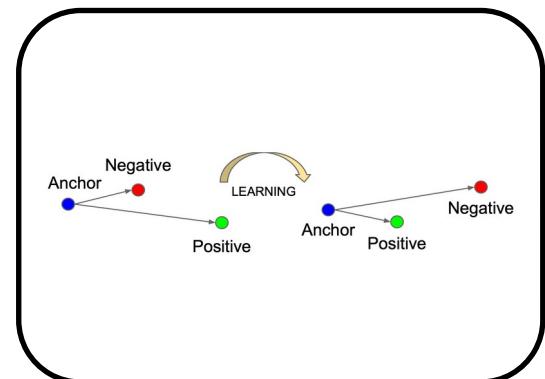
- Motivation and Challenges
- Interval-Censored Transformer Hawkes (IC-TH)



Unified representation times & event counts

$$\begin{aligned}\mathcal{L}_{\text{IC-TH-LL}}(\theta) &= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1}) \\ &\quad \underbrace{\hspace{10em}}_{\text{missing event counts}} \\ &= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i)\end{aligned}$$

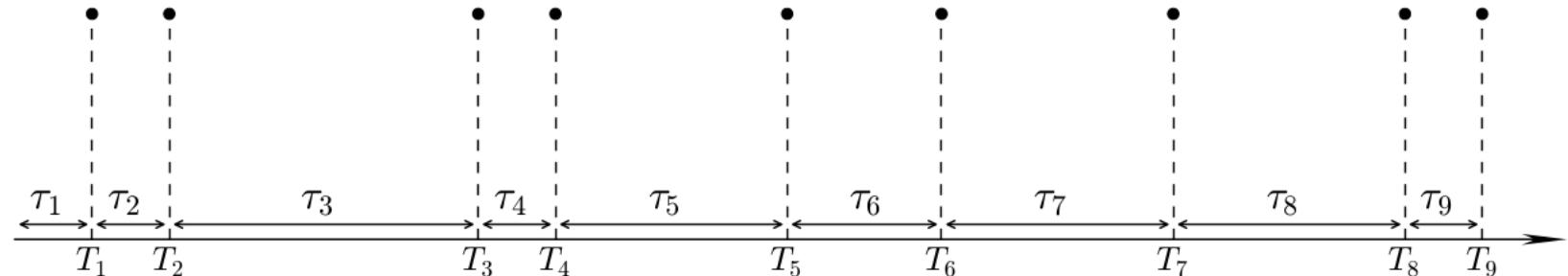
Novel log-likelihood function



Model Pre-training via Contrastive learning

- Experiments and findings

# Point process



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

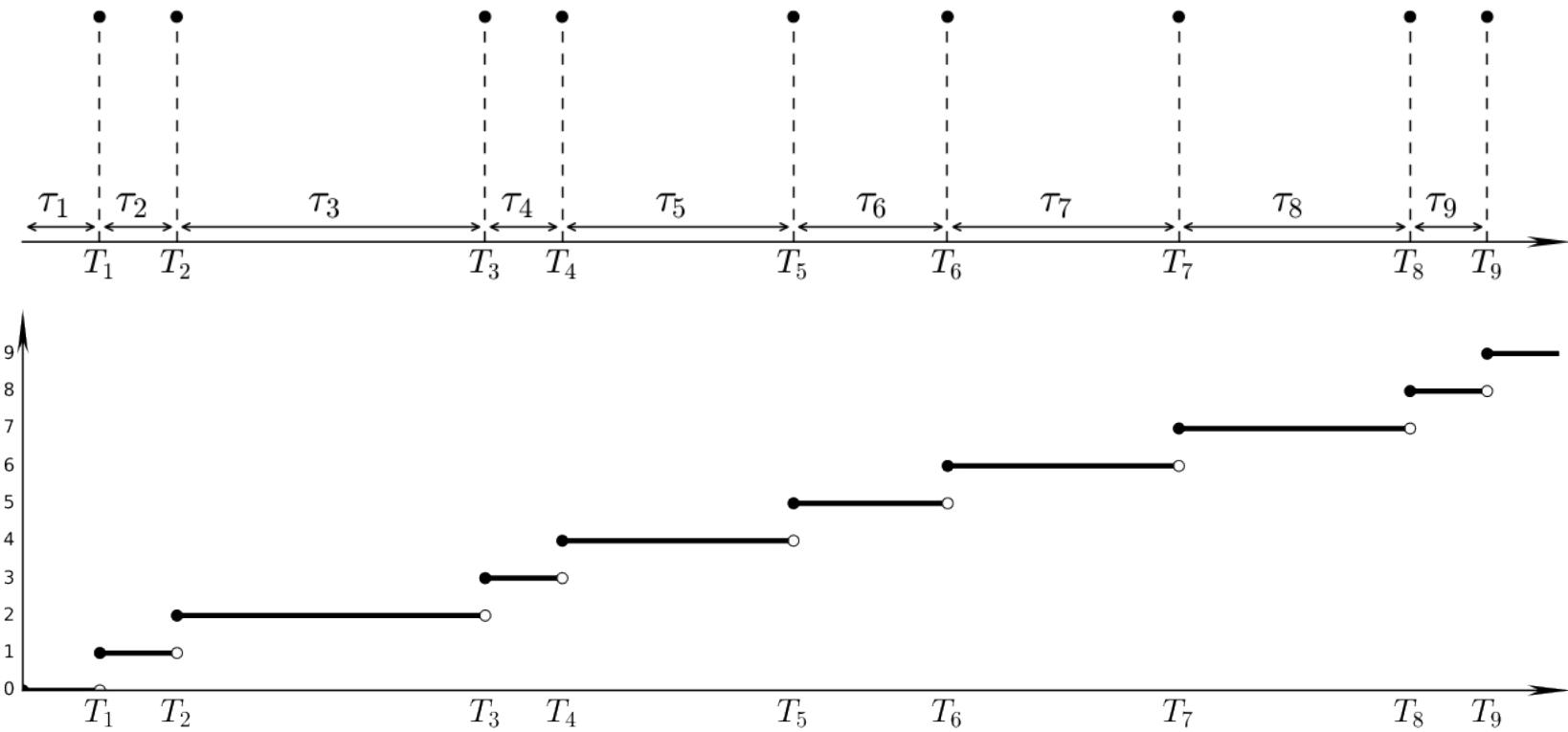
$T_i$  event times

$\tau_i$  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  $t \geq 0$ , takes integer values  $1, 2, \dots$ .  
The number of events of the point process by time  $t$

# Homogeneous Poisson process

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

$\lambda$  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 0

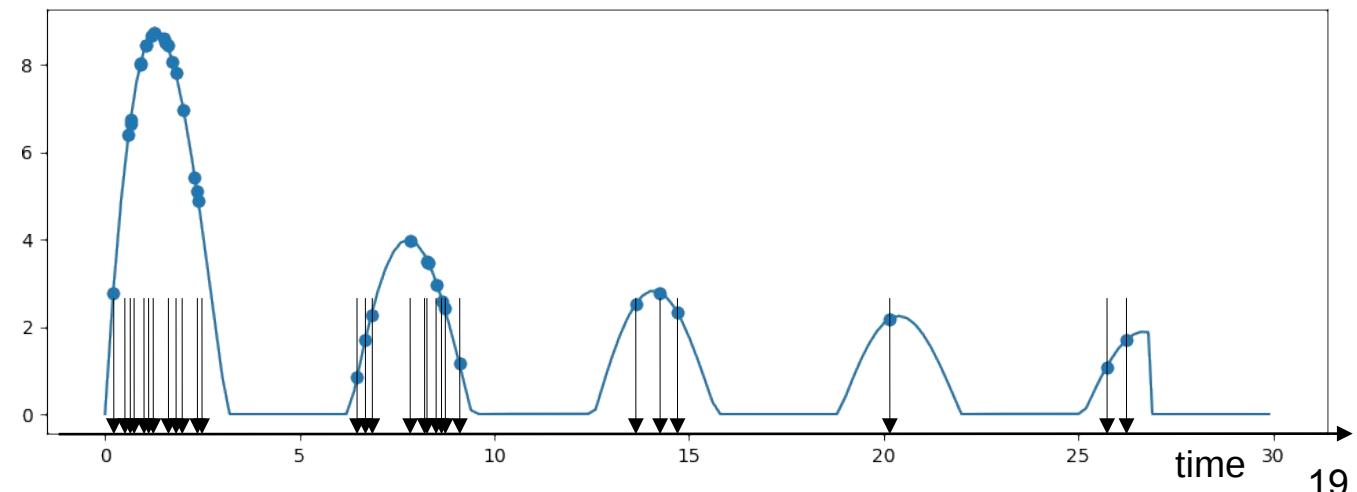
$$\mathbb{P}(\tau > t + m | \tau > m) = \mathbb{P}(\tau > t)$$

# Inhomogeneous Poisson process

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

The conditional intensity function:

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



# Inhomogeneous Poisson process

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

The conditional intensity function:

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

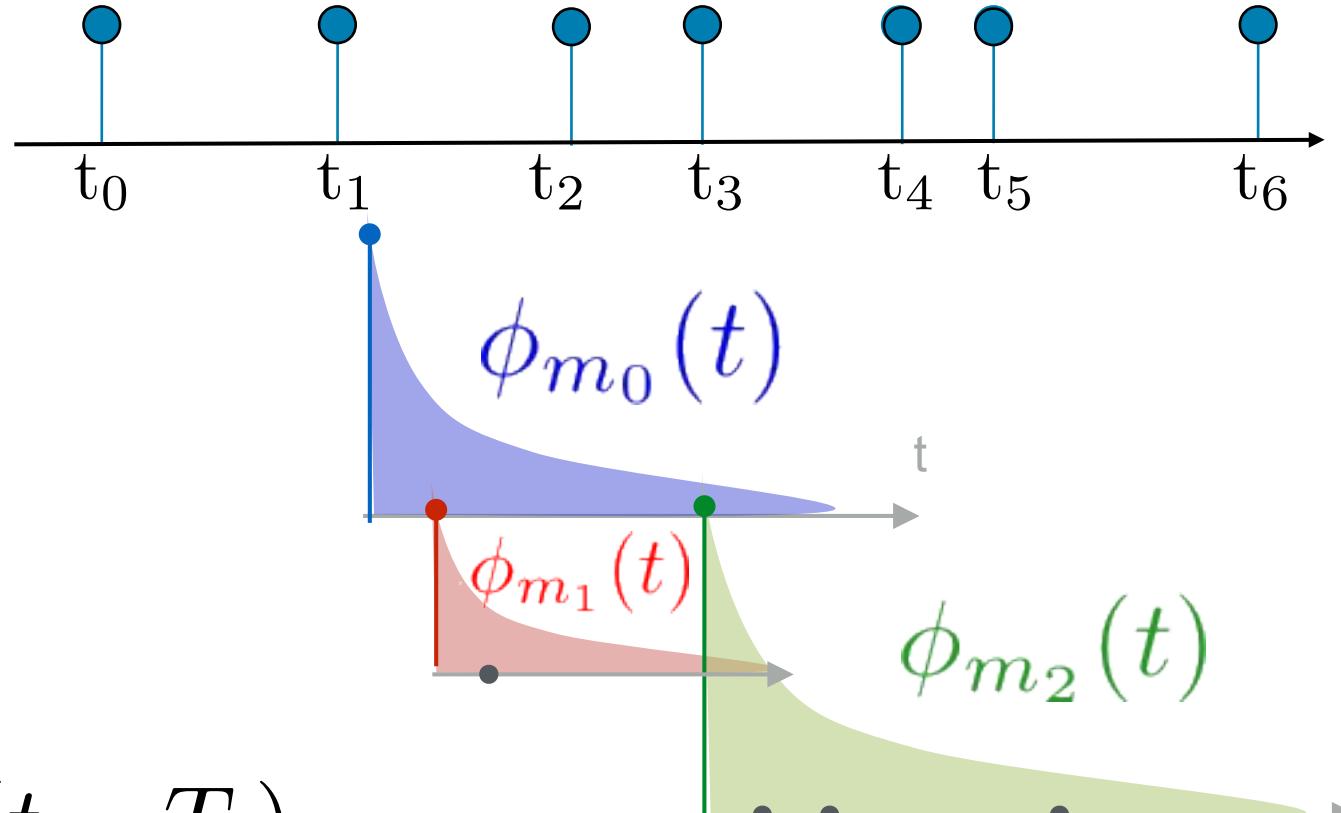
Proportional to the probability of observing an event at time t

$$\mathbb{P}(N_{t+h} = n + m | N_t = n) = \lambda(t)h + o(h) \quad \text{if } m = 1$$

$$\mathbb{P}(N_{t+h} = n + m | N_t = n) = o(h) \quad \text{if } m > 1$$

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

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



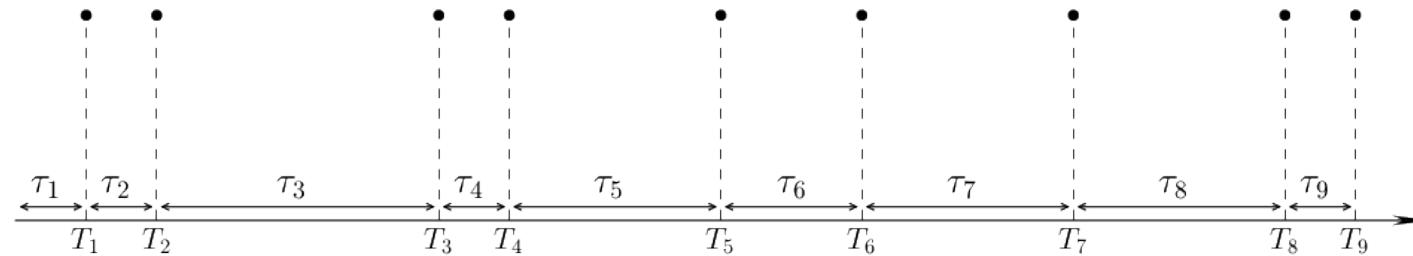
**Event intensity**

$$\lambda(t|\mathcal{H}_t) = \underline{\mu(t)} + \overline{\sum_{i:t>T_i} \phi(t - T_i)}$$

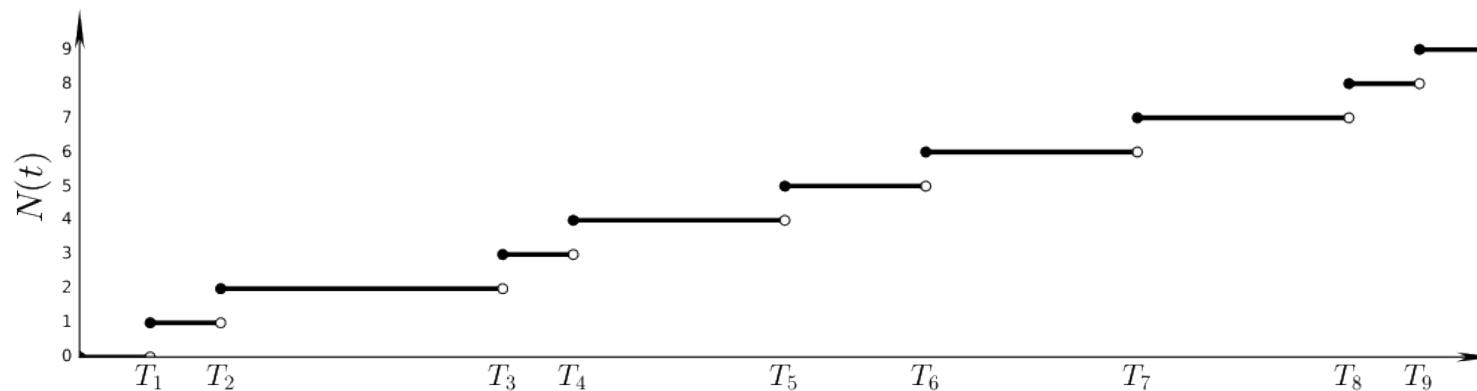
base intensity  
(exogenous)

self-excitation  
(endogenous)

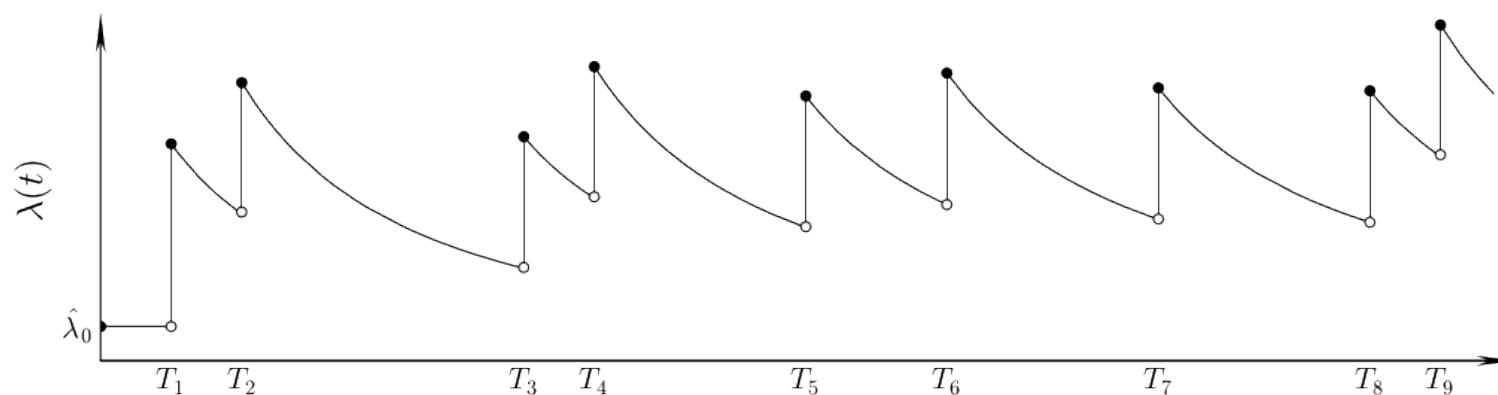
# Hawkes process – definition



event times  $T_i$



counting process  $N(t)$

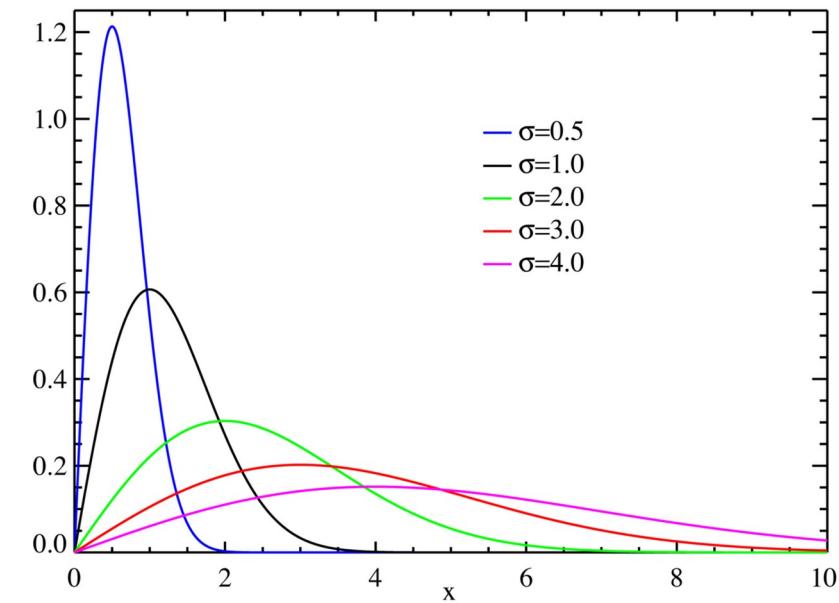
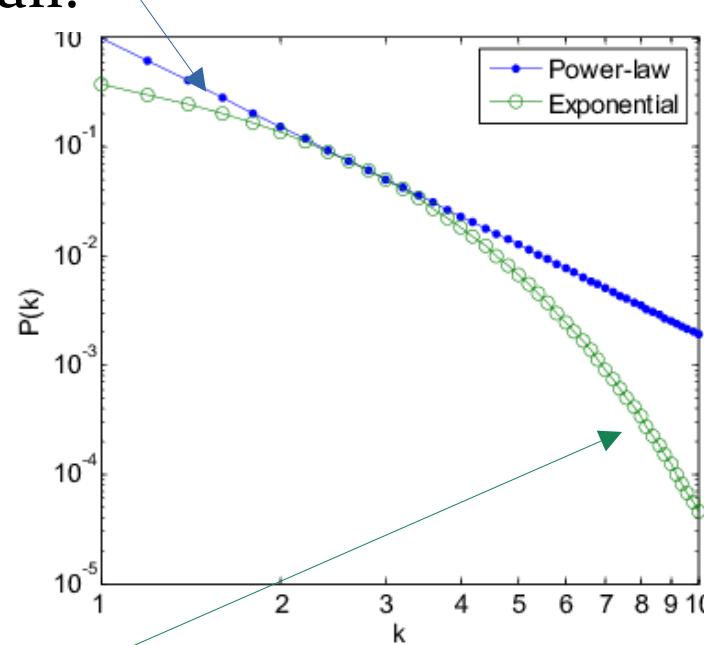
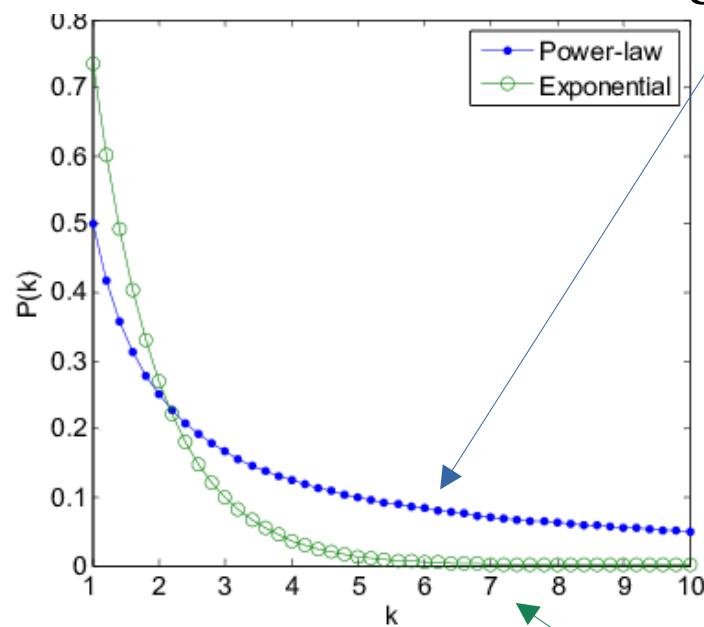


conditional intensity  
function  $\lambda(t|\mathcal{H}_t)$

# Hawkes process – choice of kernel

**Common choices:** Exponential, Power-Law, Rayleigh

Power-law has  
a long tail.

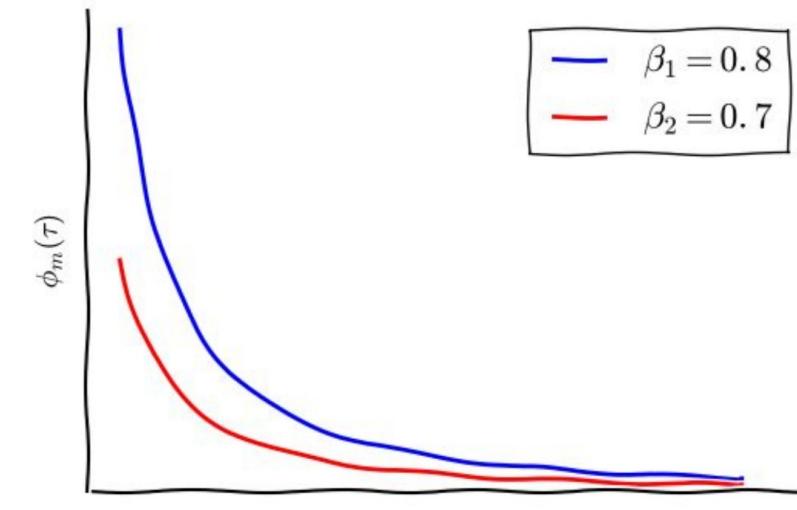
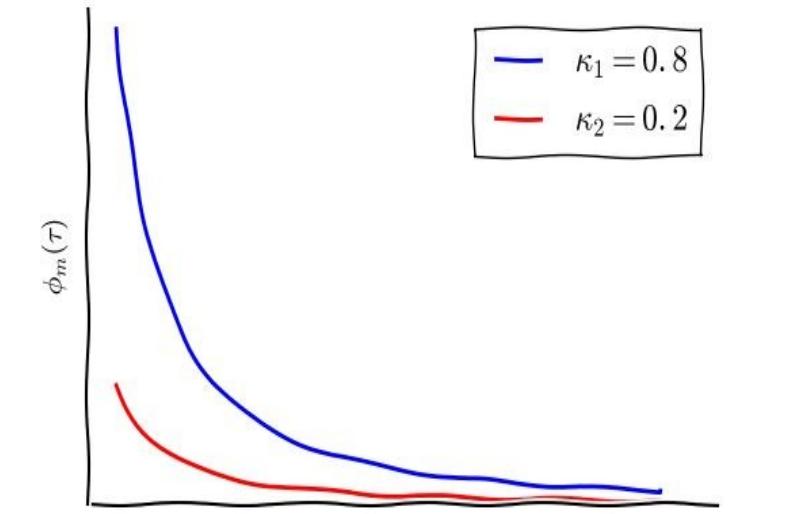
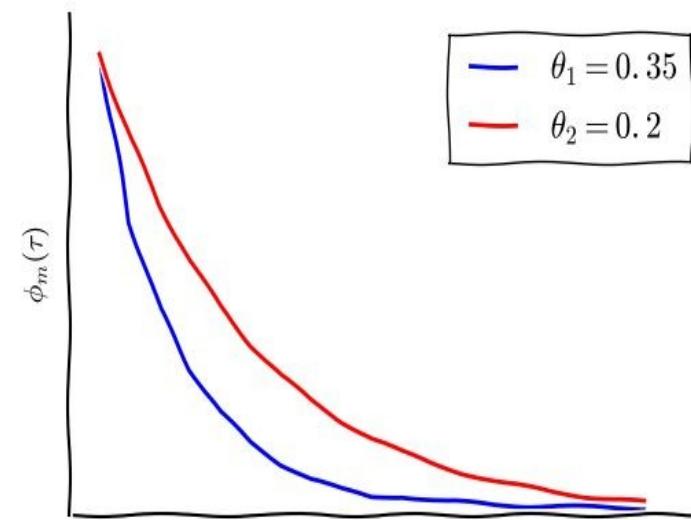


Exponential  
decays very fast.

# The “Twitter” kernel

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

↑  
the rate of  
'daughter' events  
content  
virality  
user  
influence



# Transformer Hawkes [Zuo et al, 2021]

**Event intensity**

Softplus function

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

Hidden-state

**Multi-head self-attention module**

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

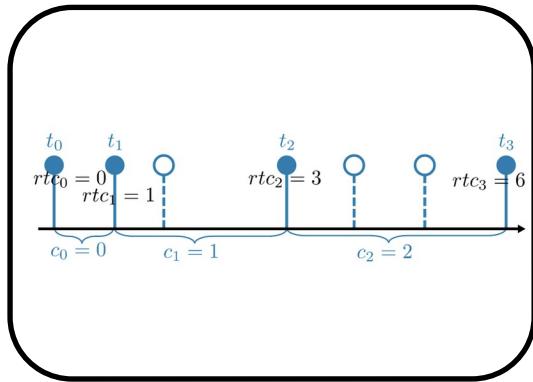
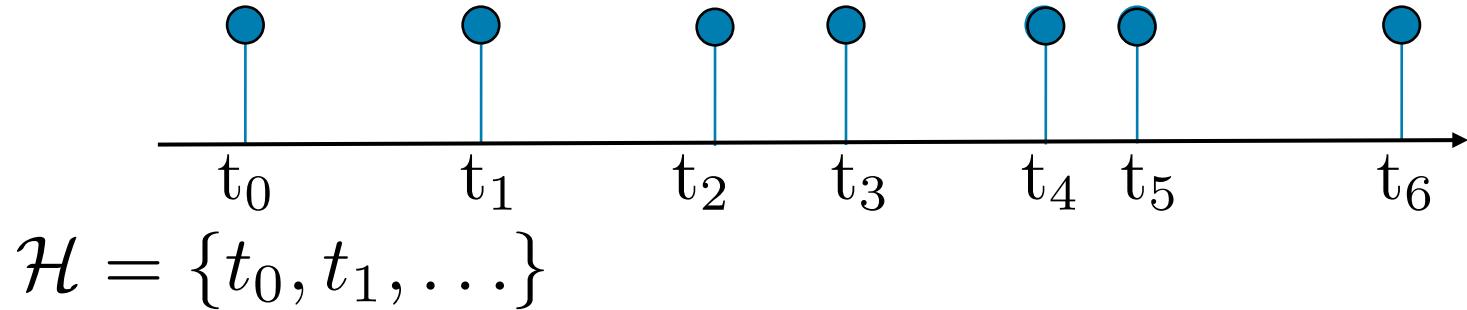
$$H(j, :) = \text{ReLU}(SW_1 + b_1)W_2 + b_2$$

$$S = \text{Concat}(\text{head}_1, \text{head}_2, \dots)W^O$$

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

# IC-TH: a mixed data format

Hawkes & Transformer Hawkes

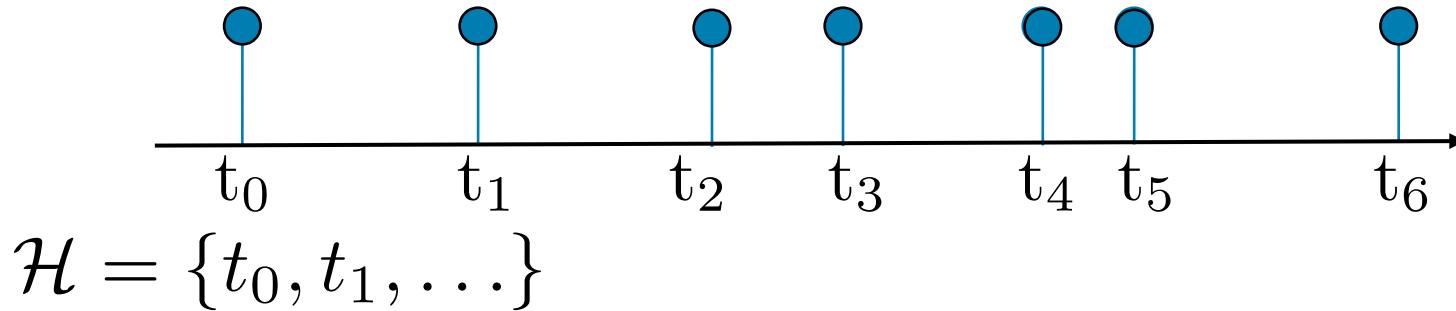


$$\mathcal{H}^* = \{(t_i, d_i, c_i)\}$$

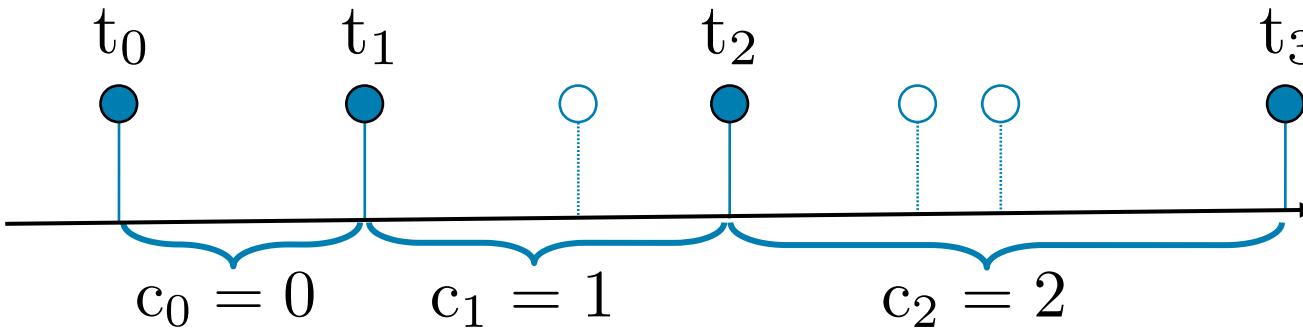
duration  
interval start time  
event count

# IC-TH: a mixed data format

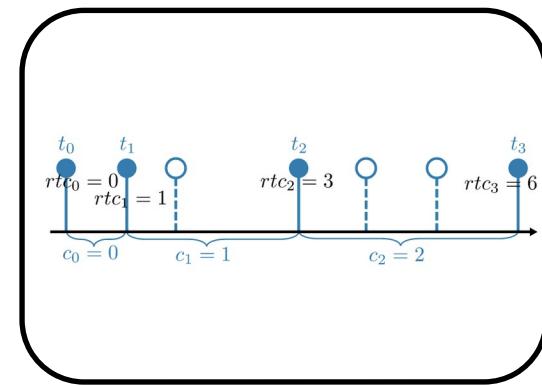
**Hawkes & Transformer Hawkes**



**Interval-Censored Transformer Hawkes**



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



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

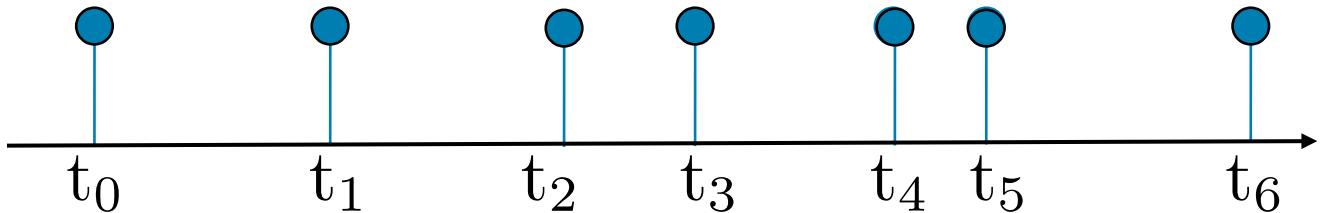
duration  
interval start time  
event count

# IC-TH: Novel log-likelihood function

**Hawkes process**

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

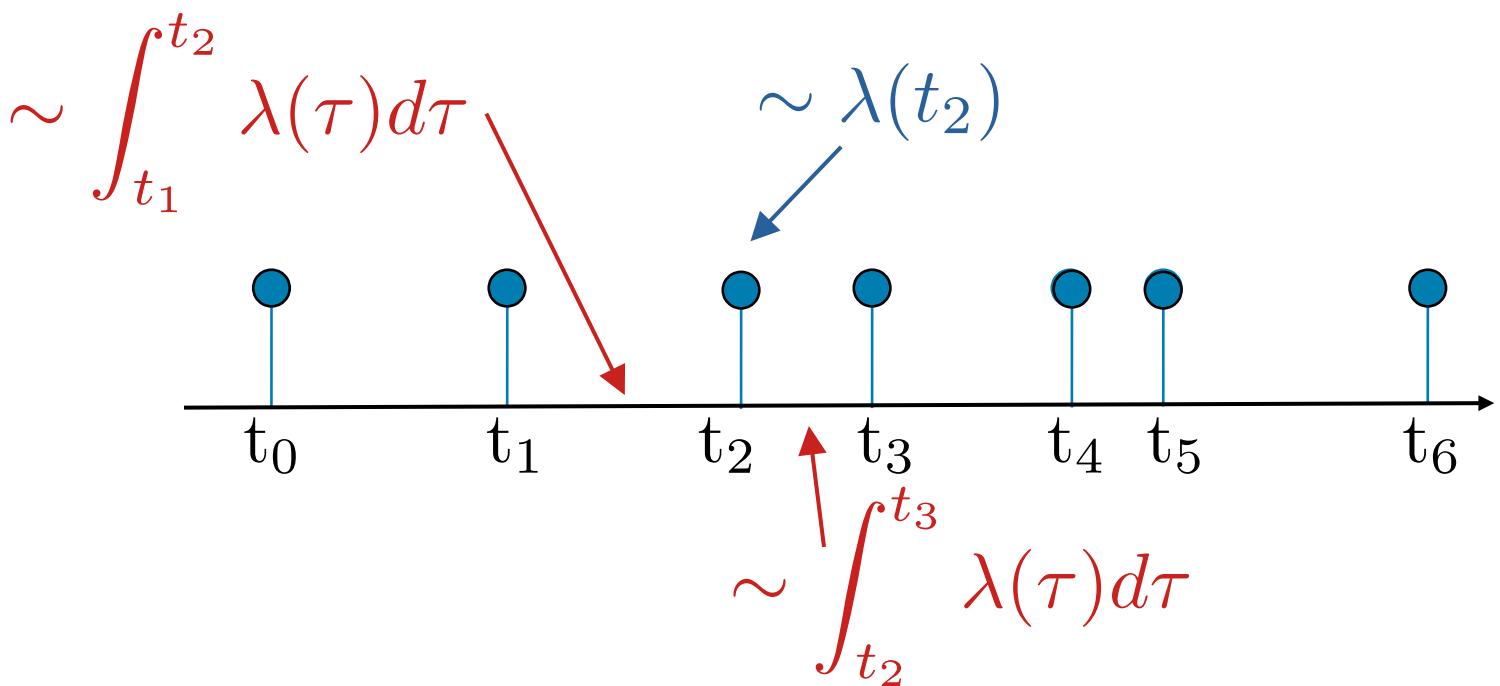
$$\begin{aligned}\mathcal{L}_{\text{IC-TH-LL}}(\theta) &= \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1})}_{\text{missing event counts}} \\ &= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i)\end{aligned}$$



# IC-TH: Novel log-likelihood function

**Hawkes process**

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



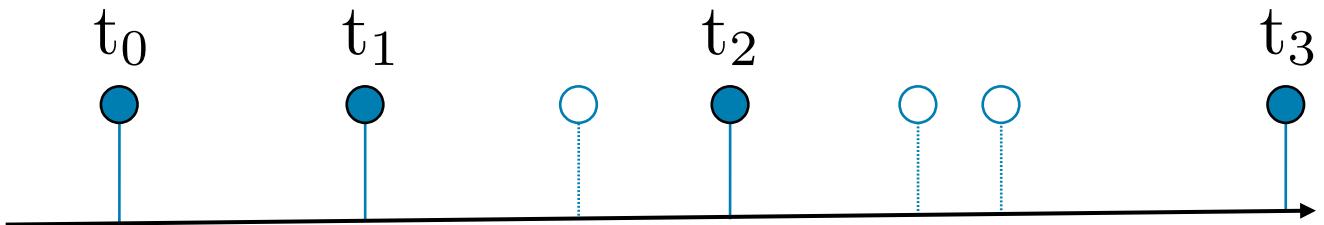
$$\begin{aligned}\mathcal{L}_{\text{IC-TH-LL}}(\theta) &= \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1})}_{\text{missing event counts}} \\ &= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i)\end{aligned}$$

# IC-TH: Novel log-likelihood function

**Interval-Censored Transformer Hawkes**

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

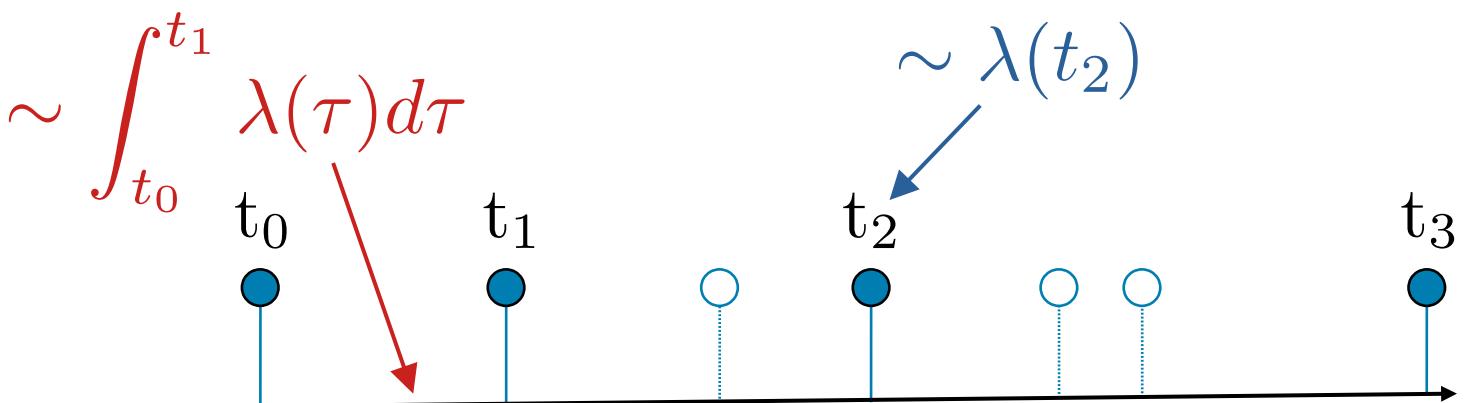
$$\begin{aligned}\mathcal{L}_{\text{IC-TH-LL}}(\theta) &= \underbrace{\sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_{i+1})}_{\text{missing event counts}} \\ &= \sum_{i \in \mathcal{H}_u^*} c_i \log \Xi(t_i, t_i)\end{aligned}$$



# IC-TH: Novel log-likelihood function

**Interval-Censored Transformer Hawkes**

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

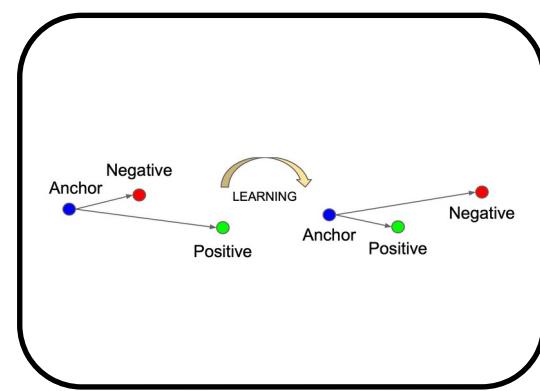


$$\sim Poi \left( c_3 = 1; \int_{t_1}^{t_2} \lambda(\tau) d\tau \right)$$

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

# IC-TH: contrastive learning

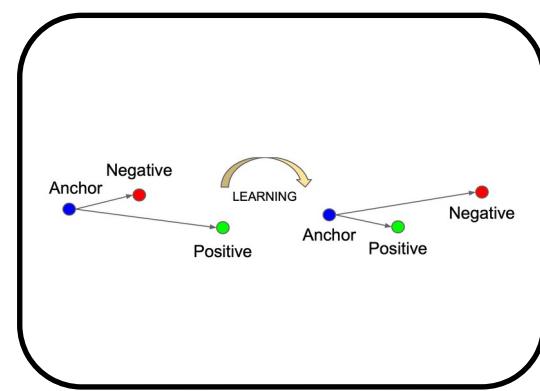
On a large, unlabeled dataset “contrast samples against each other to learn attributes that are common between data classes and attributes that set apart a data class from another.”



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

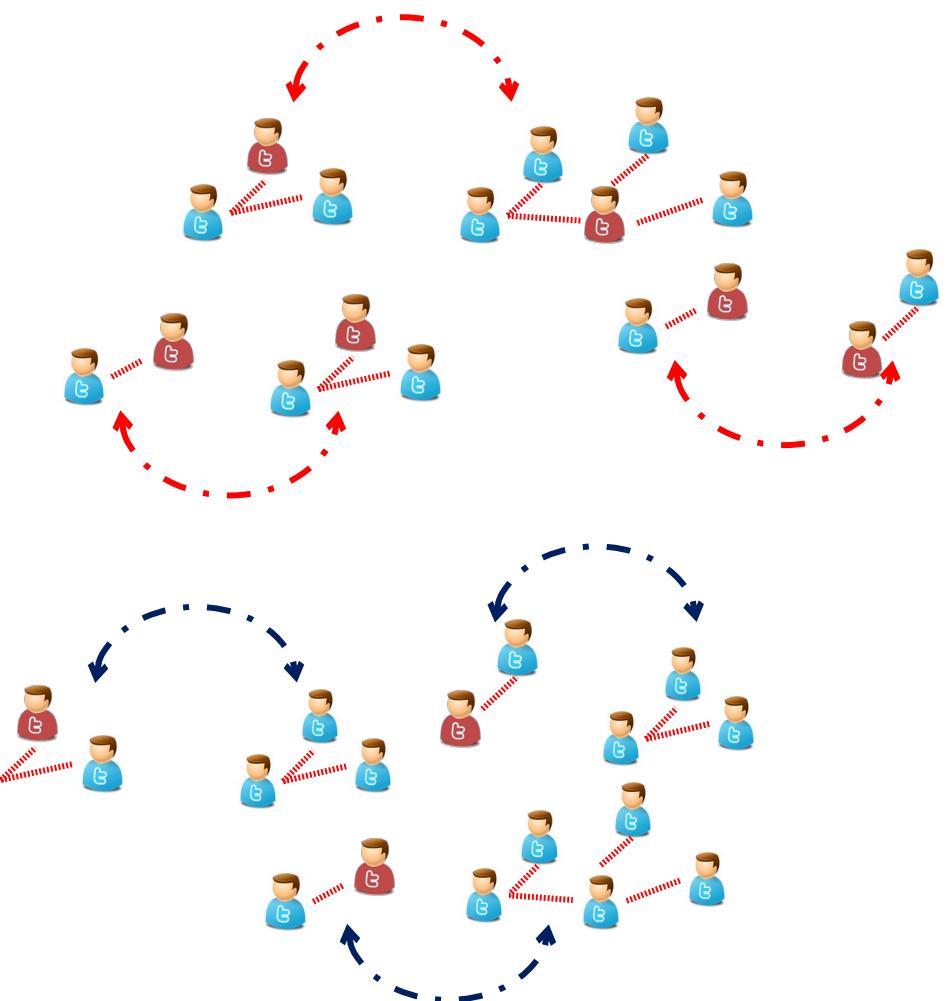
# IC-TH: contrastive learning

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

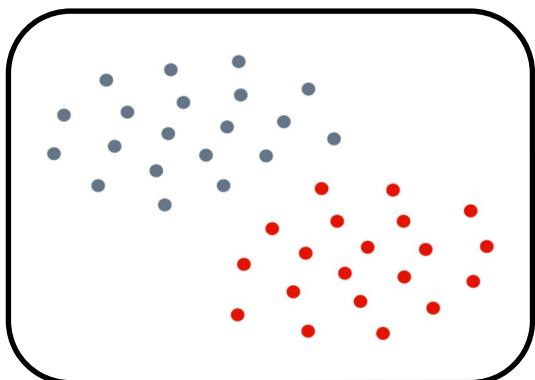


## Negative pairs

given user only appears in one cascade

# Presentation plan

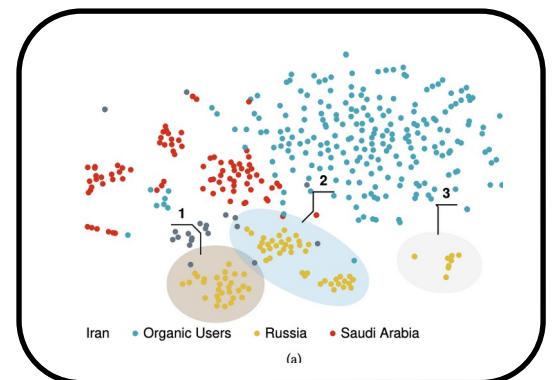
- Motivation and Challenges
- Interval-Censored Transformer Hawkes (IC-TH)
- Experiments and findings



Effects of data loss



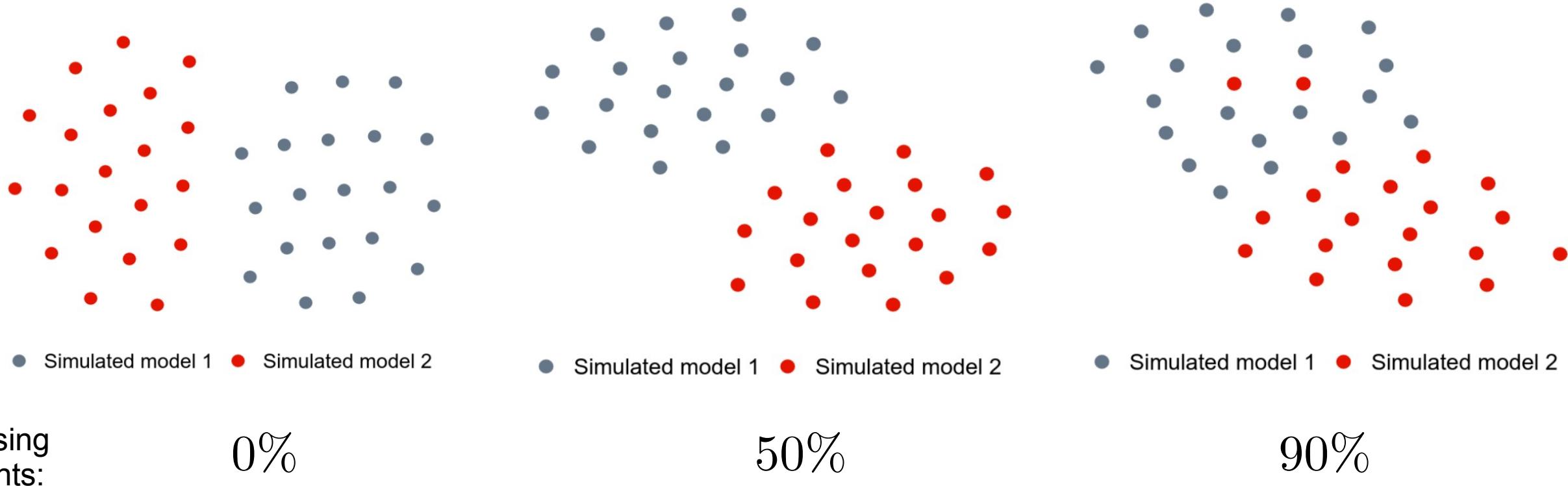
Category prediction



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



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

The Sydney Morning Herald  
INDEPENDENT. ALWAYS.

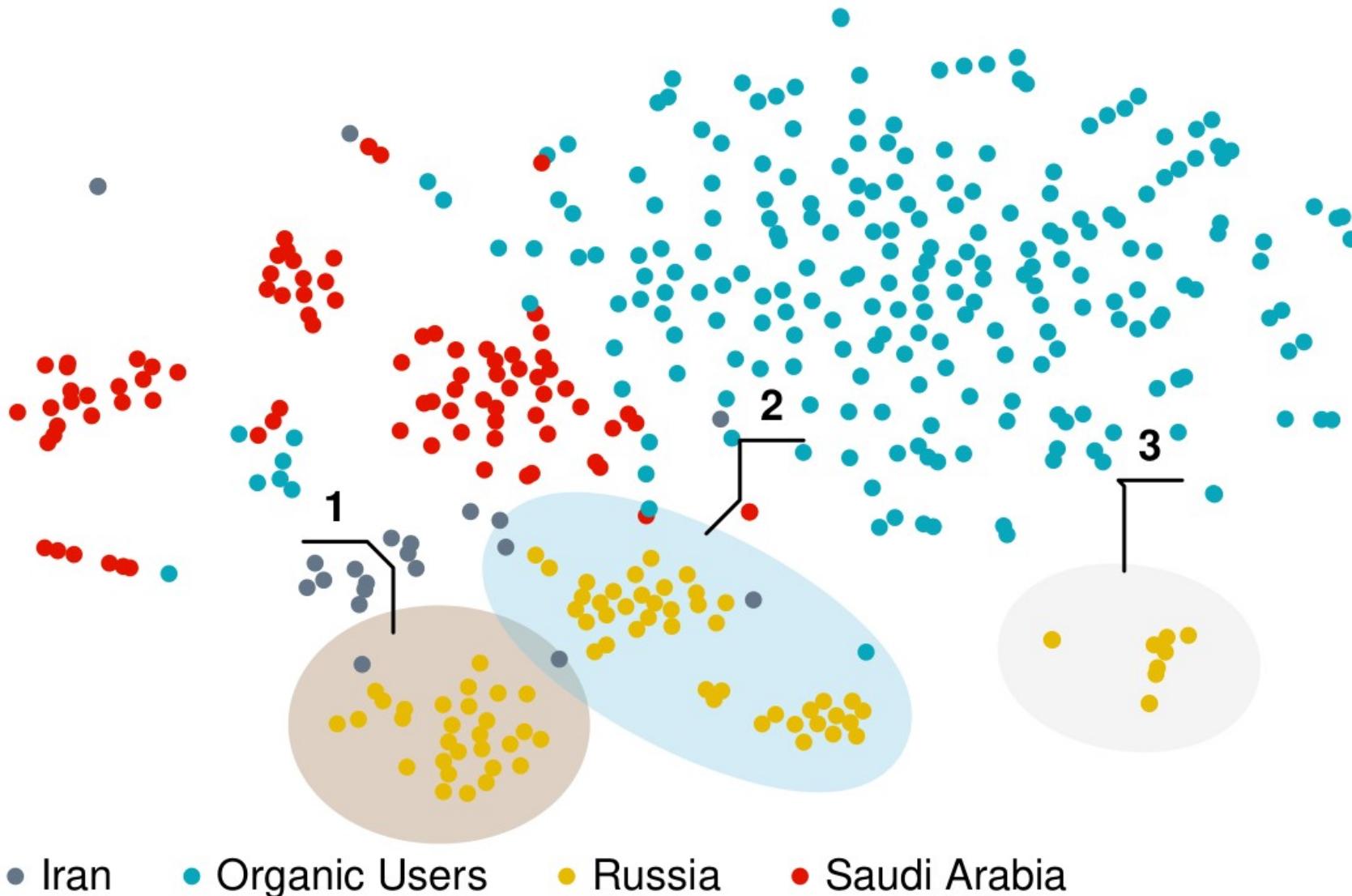


# Predicting the category of content

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	-	<b>0.987</b>
Pre-trained IC-TH	<b>0.503</b>	<b>0.853</b>	<b>0.987</b>

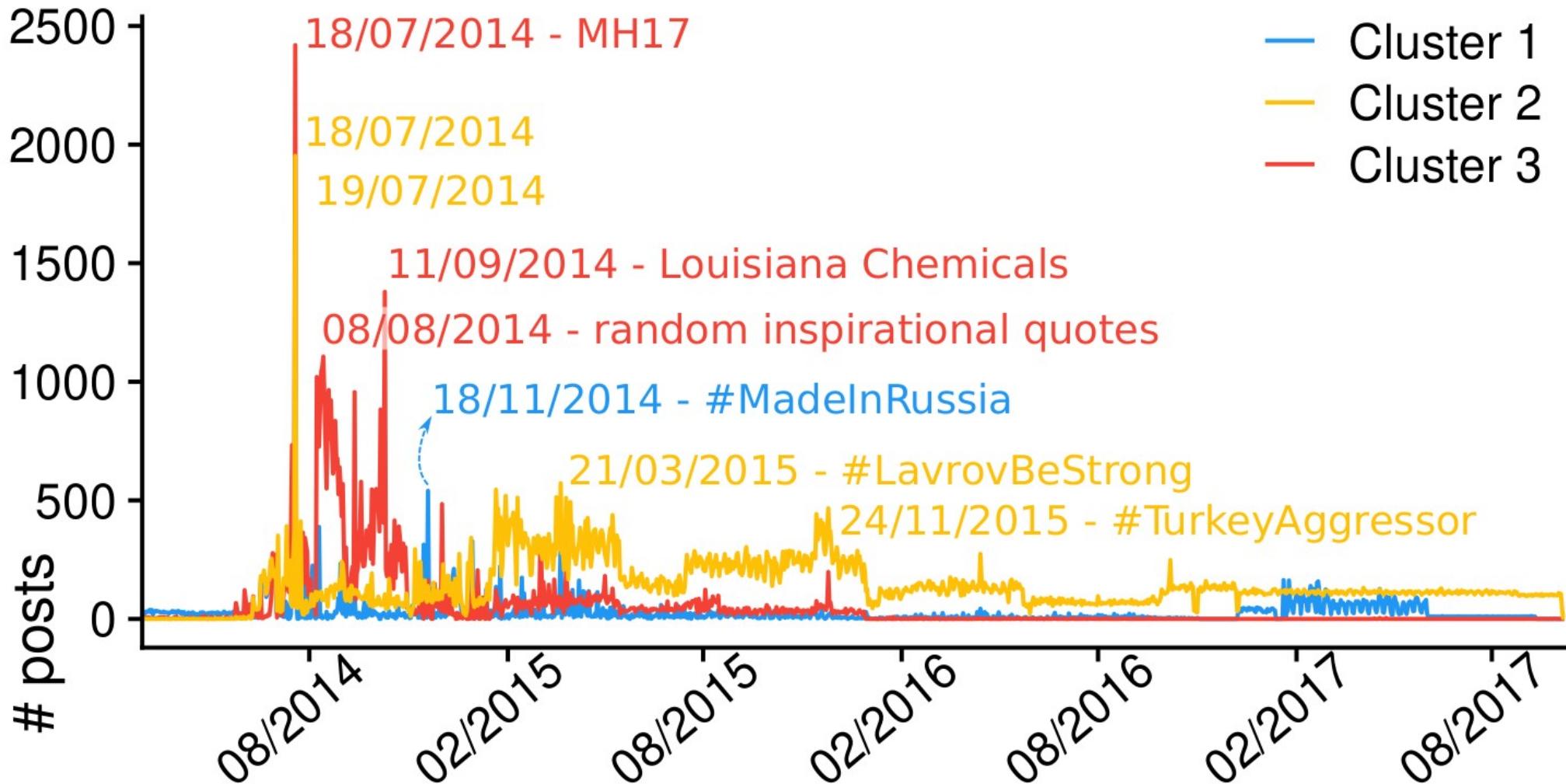
IC-TH outperforms all baselines;  
Mixed data format + loglikelihood contribute most to performance  
Missing counts and pre-training lead to moderate performance increases.

# 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 C<sub>2</sub> and C<sub>3</sub>

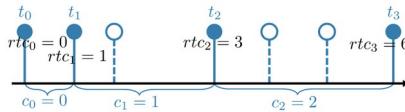


C<sub>2</sub>



C<sub>3</sub>

# Conclusion



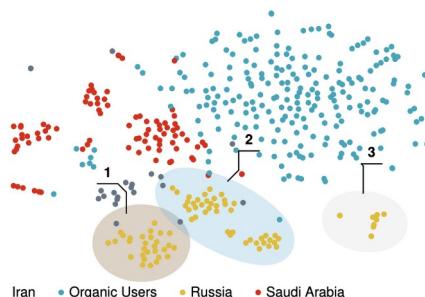
A unified representation and a novel log-likelihood for event times and missing events for the Transformer Hawkes architecture.



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



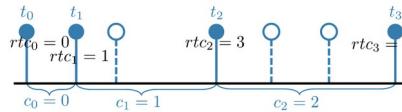
Predict the type of users and online content based solely on how the social systems react to them.



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

# Conclusion

Thank you!



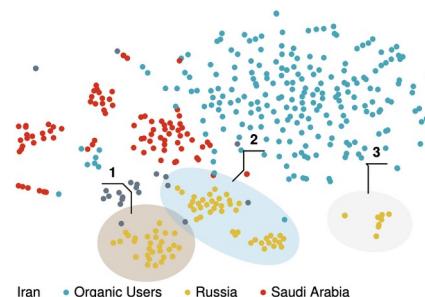
A unified representation and a novel log-likelihood for event times and missing events for the Transformer Hawkes architecture.



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.