



**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>



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

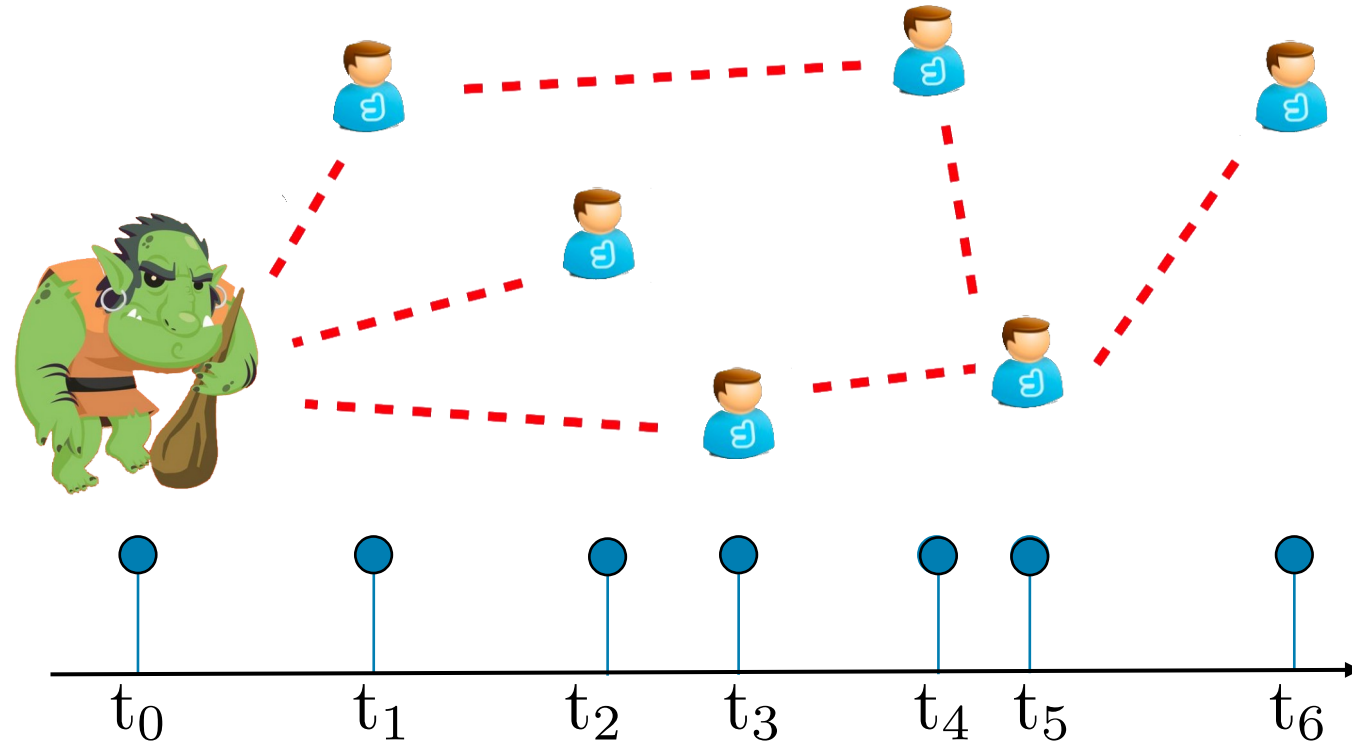
Content- and user-based detection tools:

language nuances, language drift and adversarial attacks

IO are designed to elicit particular reactions from the target audience

RQ1:

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



Challenge: partial missing data

Missing tweets

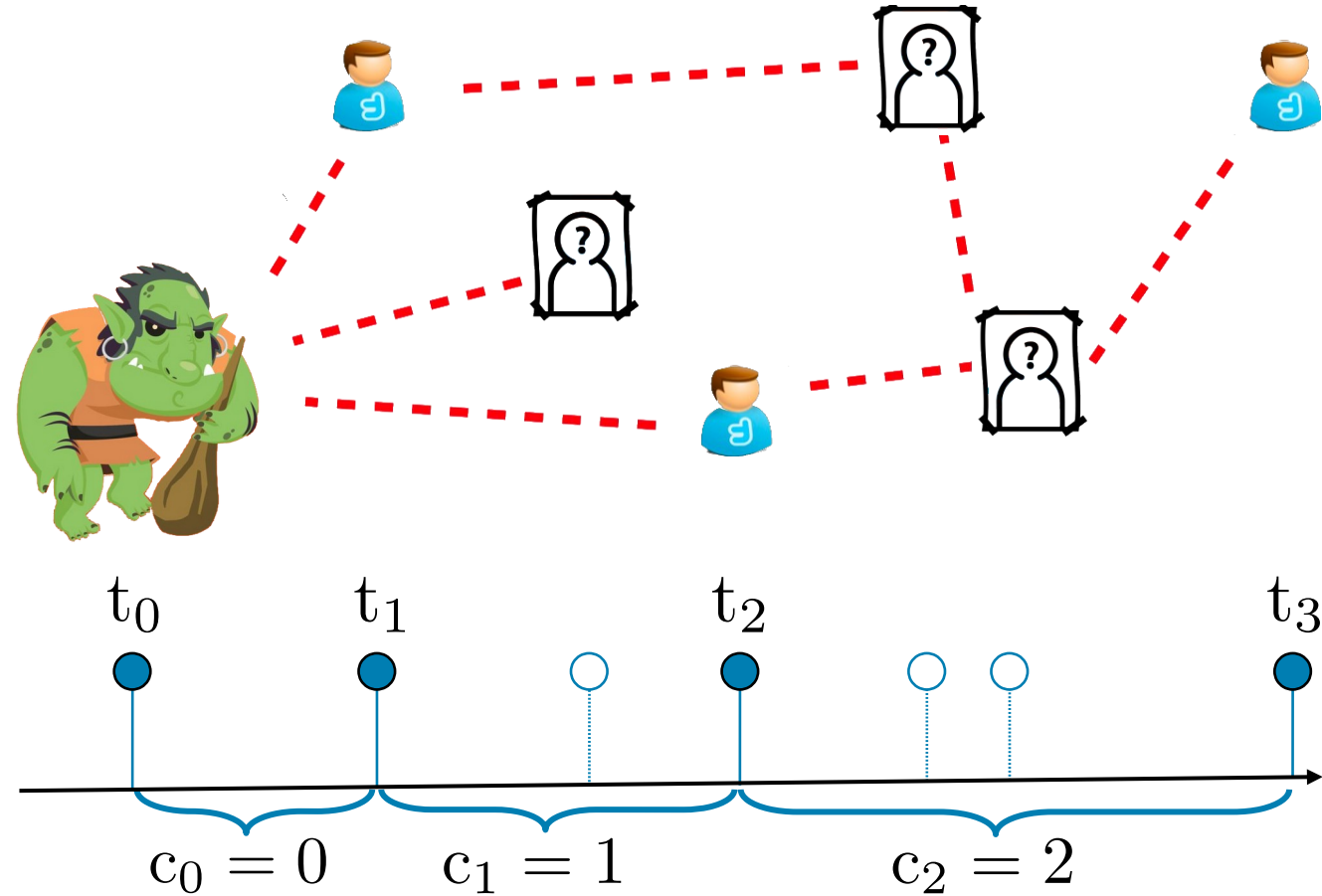
Restrictions from Twitter API [Wu et al, 2020]

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

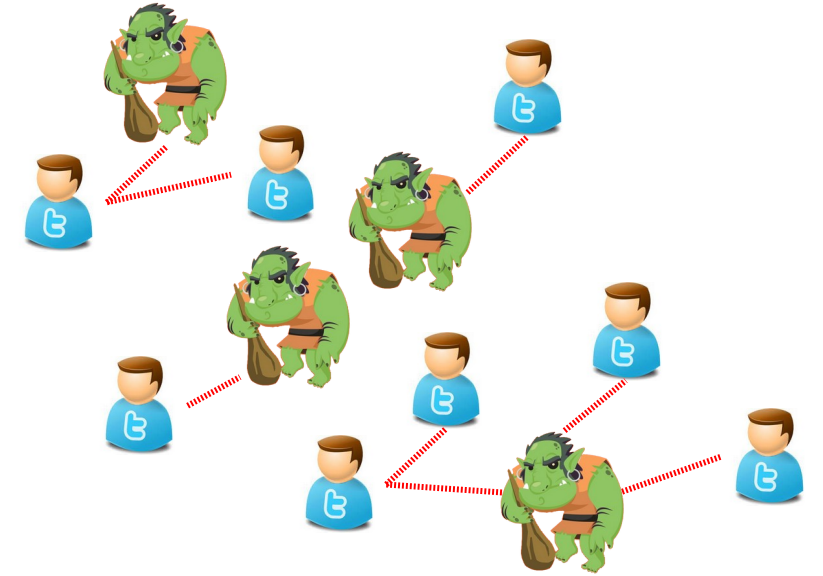
Abundant amount of unlabelled data

Public datasets

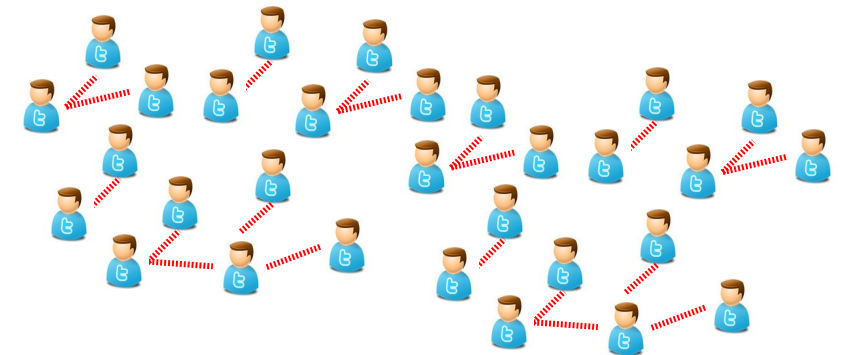
APIs

RQ3:

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



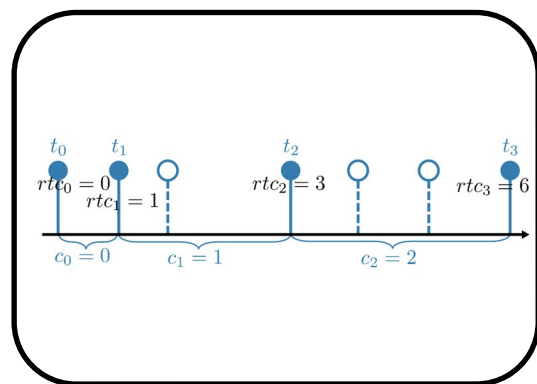
Labelled Cascades



Unlabelled Cascades

Presentation plan

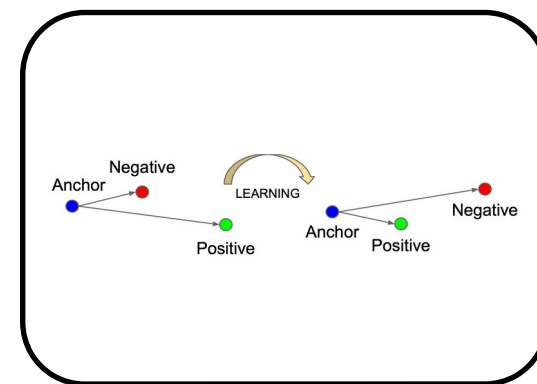
- Motivation and Challenges
- Interval-Censored Transformer Hawkes (IC-TH)



Unified representation times & event counts

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

Novel log-likelihood function

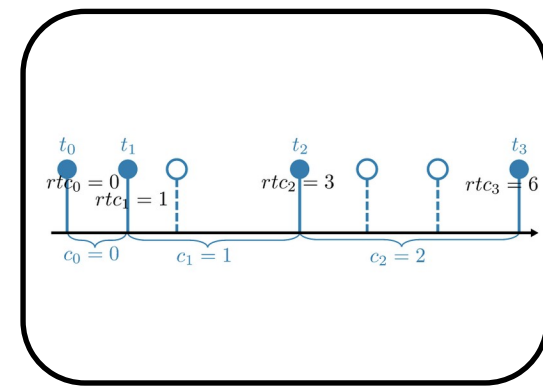
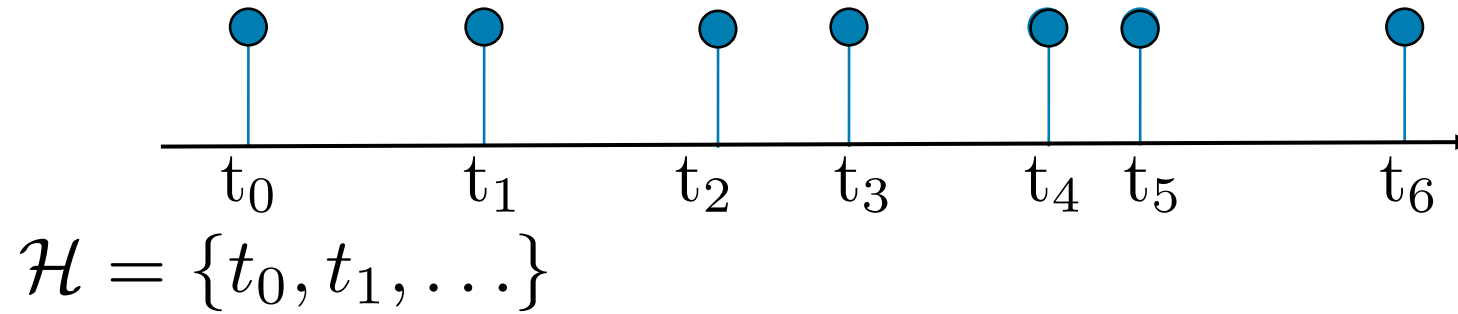


Model Pre-training via Contrastive learning

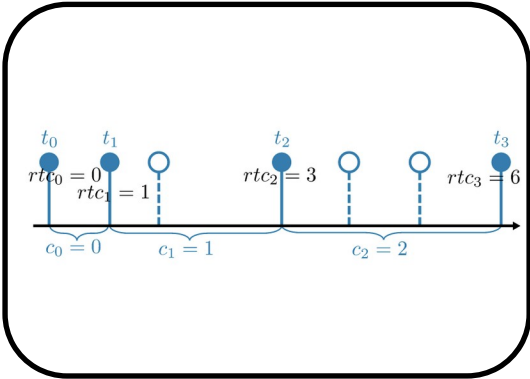
- Experiments and findings

IC-TH: a mixed data format

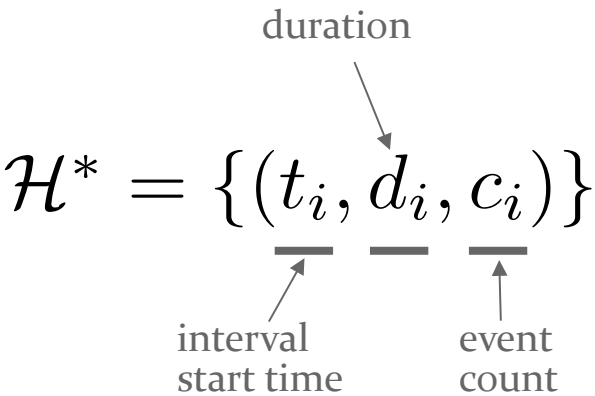
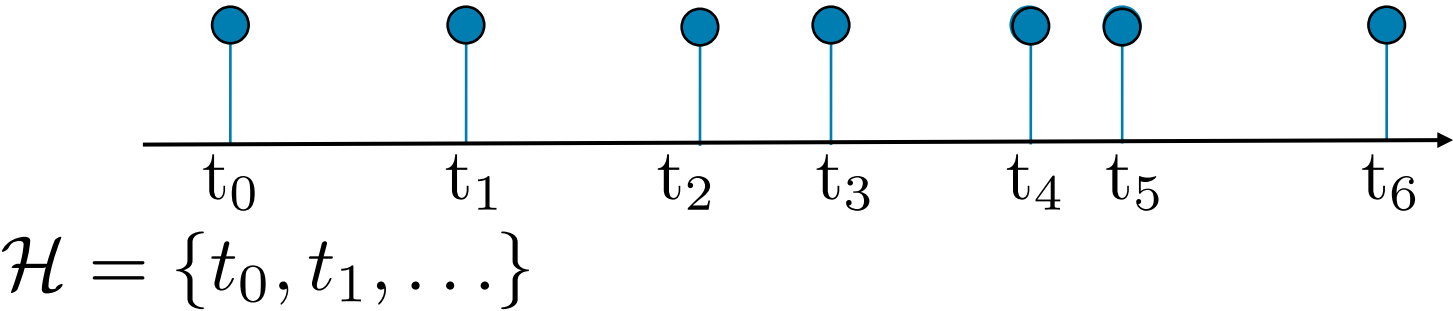
Hawkes & Transformer Hawkes



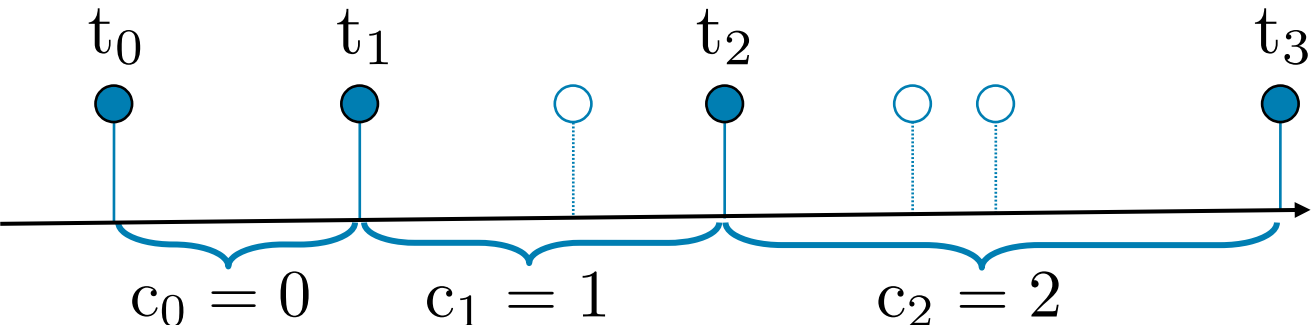
IC-TH: a mixed data format



Hawkes & Transformer Hawkes



Interval-Censored Transformer Hawkes



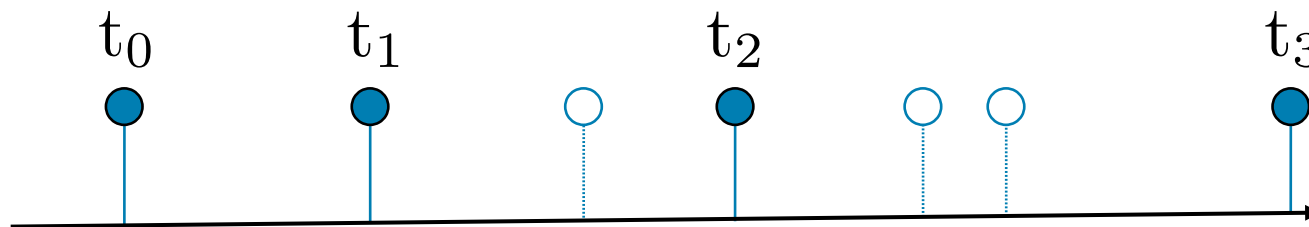
$$\mathcal{H}^* = \{(t_0 - dt, 2dt, 1), (t_0, t_1 - t_0, 0), (t_1 - dt, 2dt, 1), (t_1, t_2 - t_1, 1), \dots\}$$

IC-TH: Novel log-likelihood function

Interval-Censored Transformer Hawkes

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

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

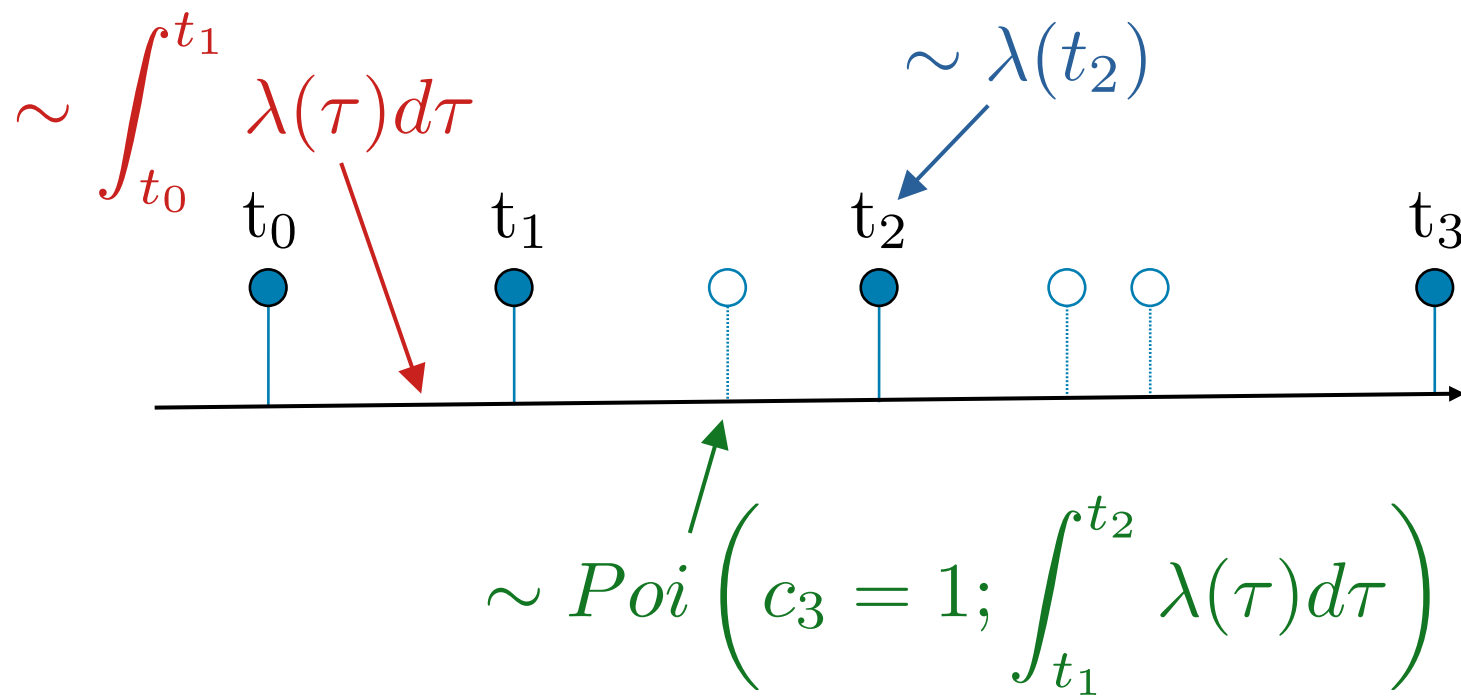


IC-TH: Novel log-likelihood function

Interval-Censored Transformer Hawkes

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

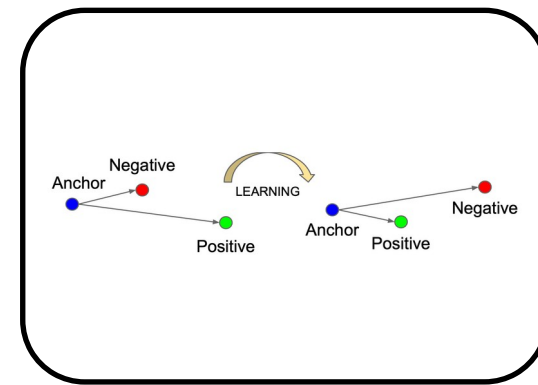
$$\log L(\theta) = \sum_{i \in \mathcal{H}_u^*} c_i \log \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau + \sum_{i \in \mathcal{H}_c^*} \log \lambda(t_i) - \sum_{i \in \mathcal{H}^*} \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau$$



IC-TH: contrastive learning

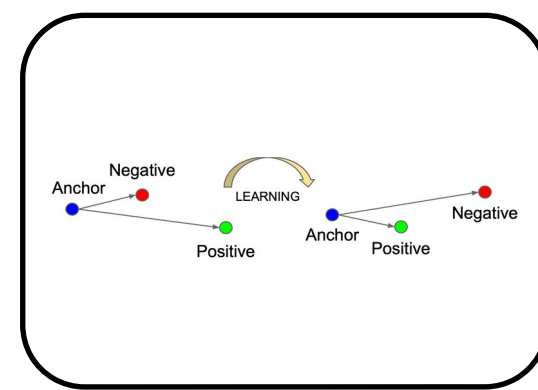
On a large, unlabeled dataset “contrast samples against each other to learn attributes that are common between data classes and attributes that set apart a data class from another.”

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



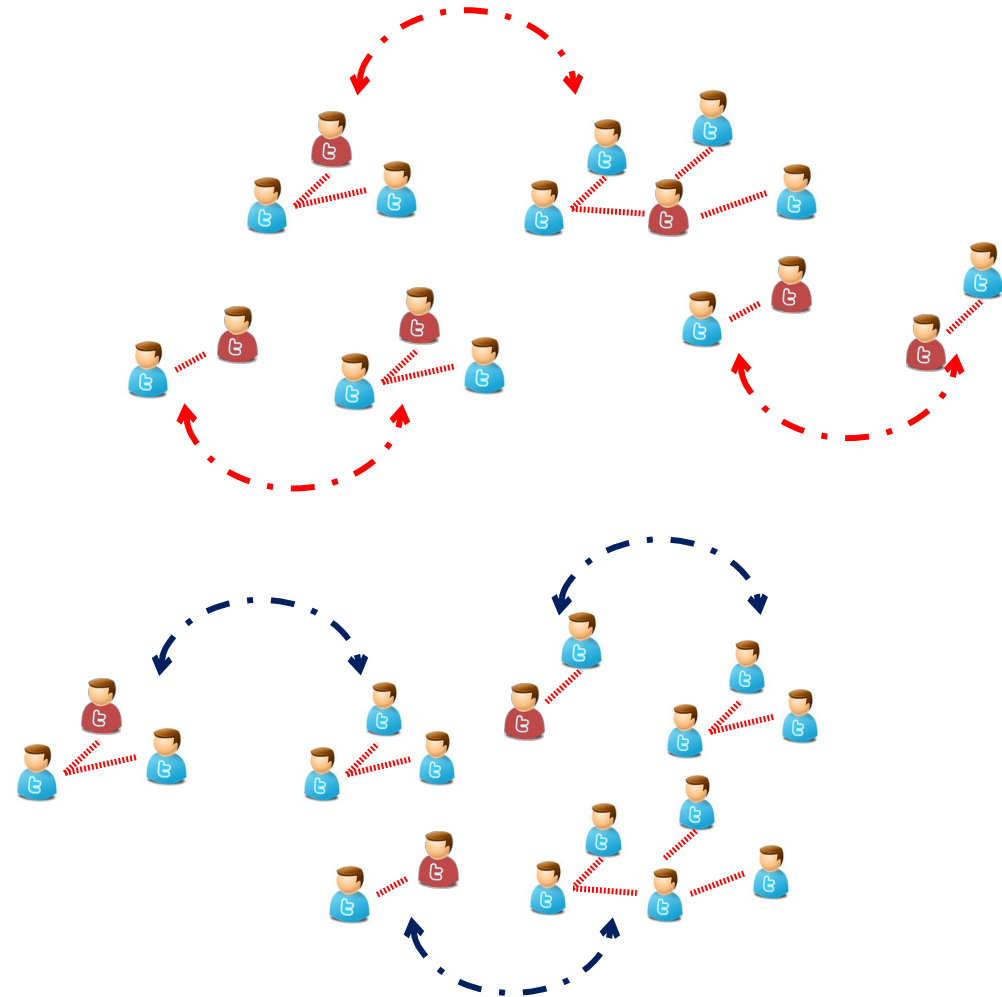
IC-TH: contrastive learning

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

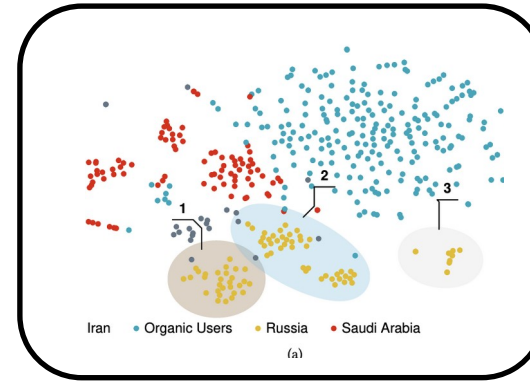
given user only appears in one cascade

Presentation plan

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



Category prediction



Analysis of Information
Operations Dataset

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: 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:

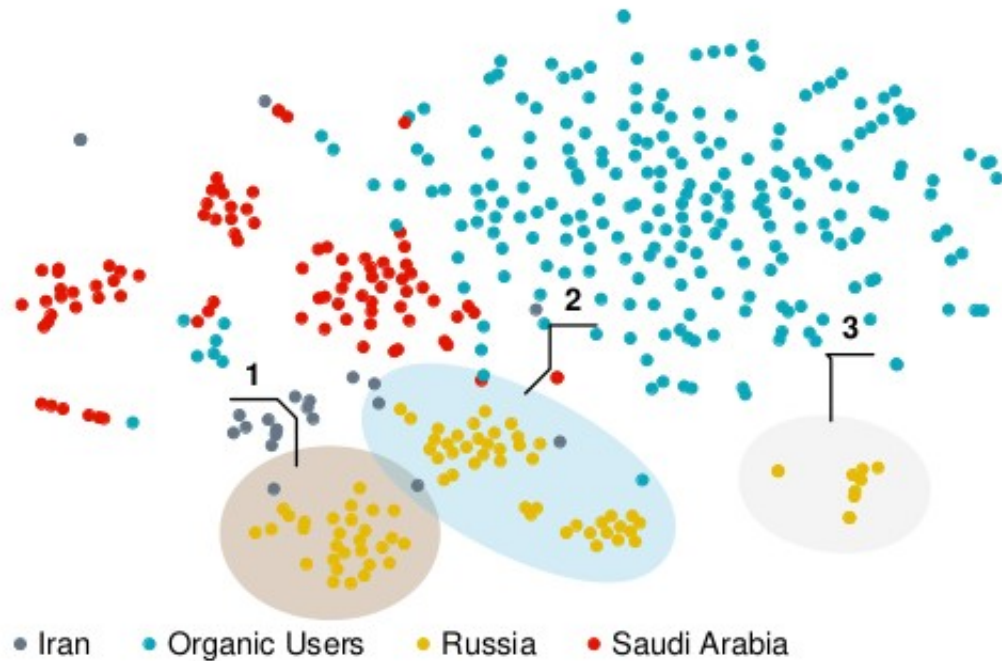
Russia, Iran, Saudi Arabia, Organic users

	Models	
[Kong et al, 2020]	Discrete Mixture Models	0.968
[Zuo et al, 2021]	Transformer Hawkes	0.983
	IC-TH w/o missing counts	0.985
	IC-TH	0.987

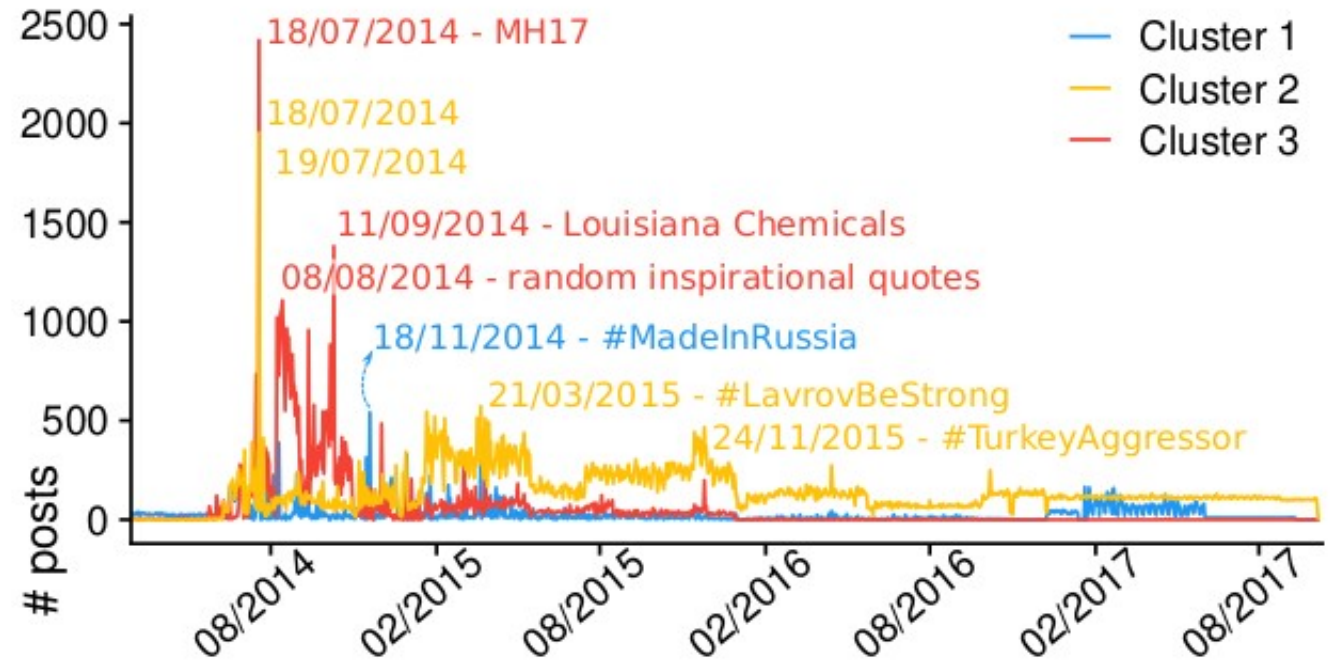


- IC-TH outperforms all baselines;
- Mixed data format + loglikelihood contribute most to performance
- Missing counts – moderate performance increases.

Strategies of Russia-backed Information Operations



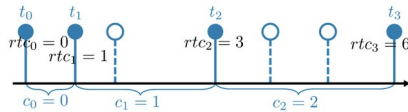
IC-TH clusters IO agents from specific countries based solely on the timing of the cascades in which they participate; it identifies even individual “troll farms”.



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

Conclusion



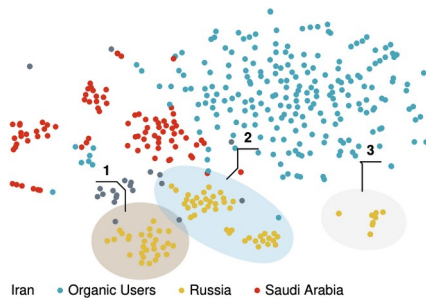
A unified representation and a novel log-likelihood for event times and missing events for the Transformer Hawkes architecture.



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



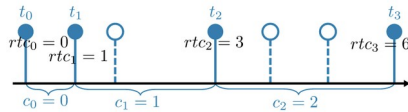
Predict the type of users and online content based solely on how the social systems react to them.



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

Conclusion

Thank you!



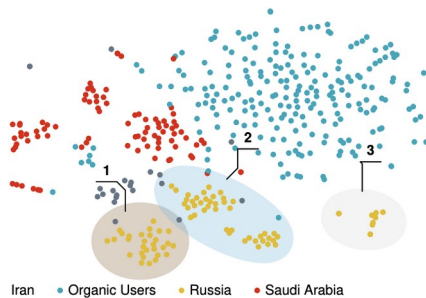
A unified representation and a novel log-likelihood for event times and missing events for the Transformer Hawkes architecture.



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.

Who are our online opinion leaders?



Common traits:

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



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
📅 Joined October 2014
📅 Born on October 02



Tennessee GOP
@TEN_GOP

I love God, I Love my Country

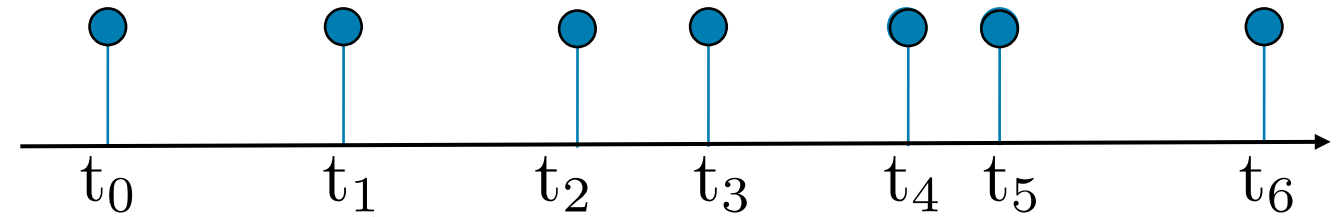
📍 Tennessee, USA
📅 Joined November 2015

Common traits:

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

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

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

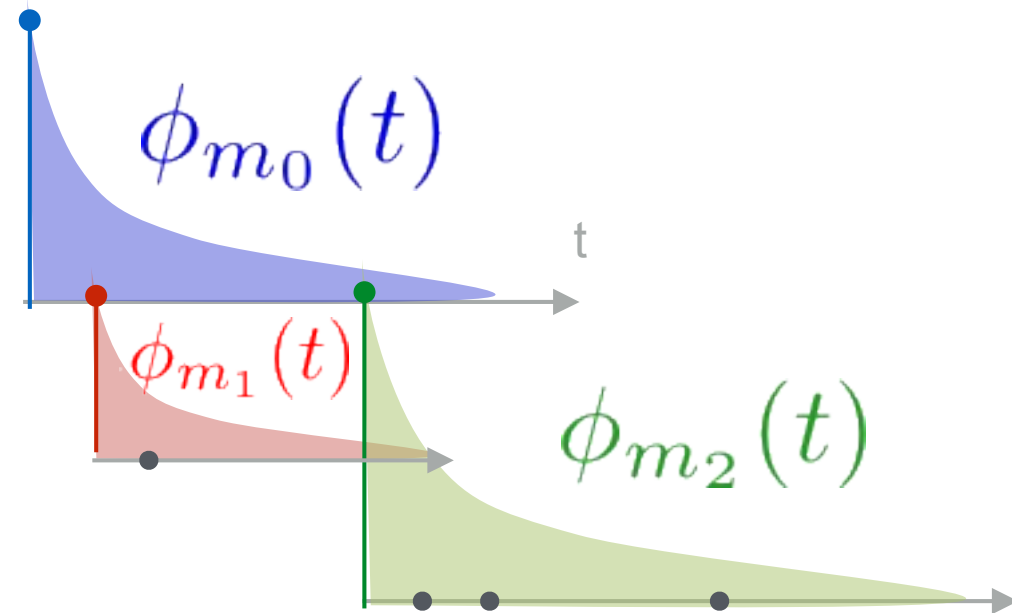


Event intensity

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

base intensity
(exogenous)

self-excitation
(endogenous)



Transformer Hawkes [Zuo et al, 2021]

Event intensity

Softplus function

$$\lambda(t) = \underbrace{f(w^\top h(t))}_{\text{Hidden-state}}$$

Multi-head self-attention module

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

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

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

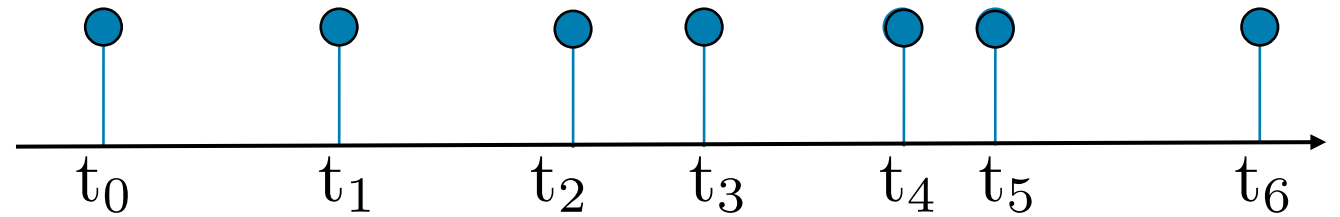
$$\text{head}_i = \text{Softmax} \left(\frac{XW_i^Q (XW_i^K)^\top}{\sqrt{d_k}} \right) XW_i^V$$

IC-TH: Novel log-likelihood function

Hawkes process

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

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

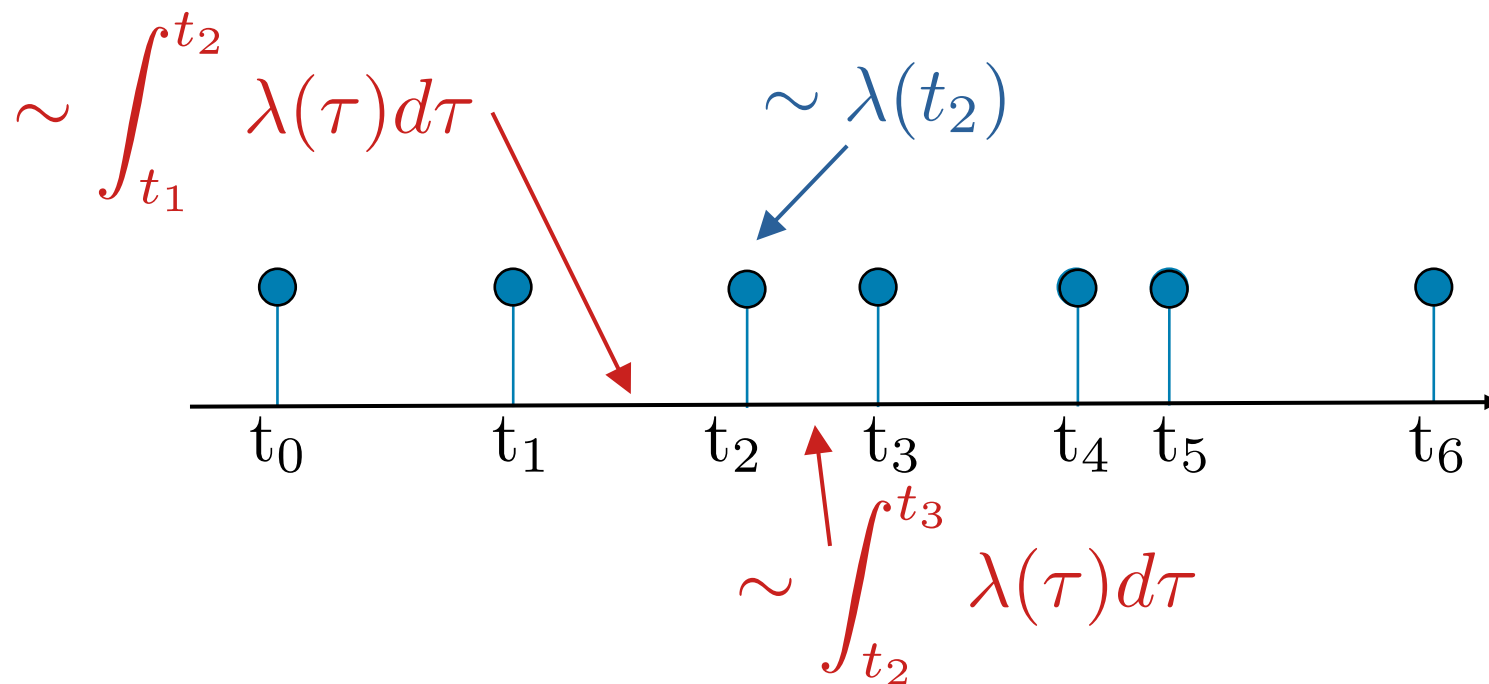


IC-TH: Novel log-likelihood function

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

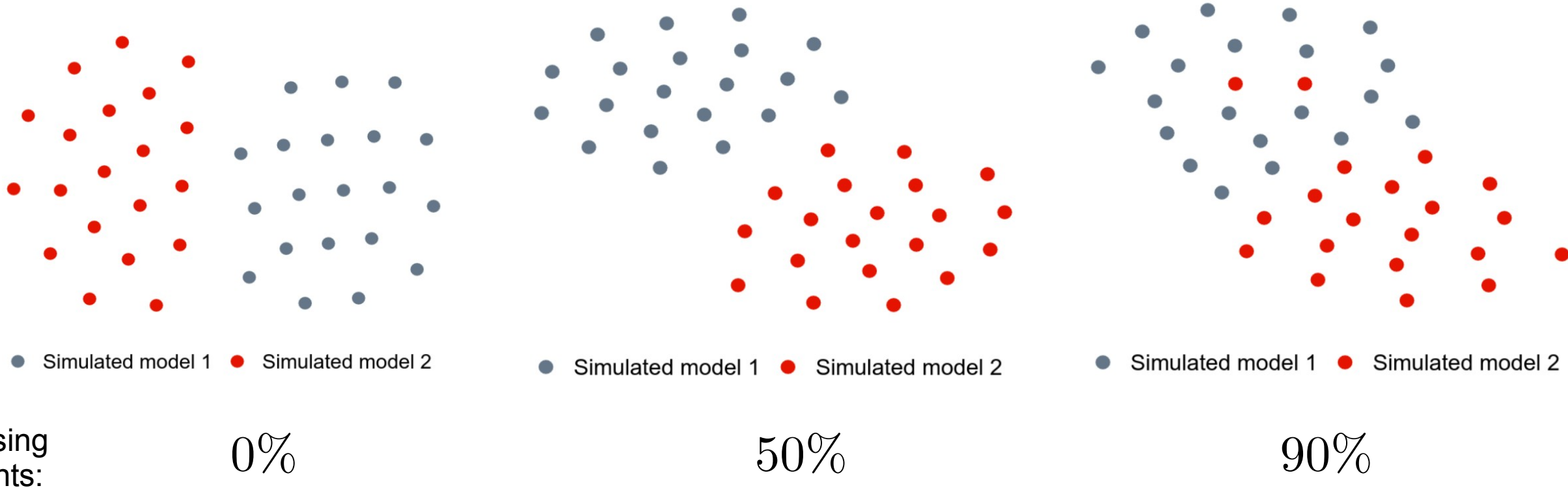
Hawkes process

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



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: Reputable&controversial news sources, tweeted YouTube



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




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

Classes – states sponsoring IO:
Reputable, controversial

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.