AFRL-AFOSR-JP-TR-2022-0018

Linking Online Attention to Measurable Actions

Xie, Lexing
AUSTRALIAN NATIONAL UNIVERSITY RESEARCH OFFICE ACTON (AUSTRALIA)
10C EAST RD
ACTON, ,
AU

03/31/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20220331	Final	START DATE 20190830	END DATE 20210829

4. TITLE AND SUBTITLE
Linking Online Attention to Measurable Actions

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA2386-19-1-4078		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Lexing Xie

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AUSTRALIAN NATIONAL UNIVERSITY RESEARCH OFFICE ACTON (AUSTRALIA)
10C EAST RD
ACTON
AU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD UNIT 45002 APO AP 96338-5002	AFRL/AFOSR IOA	AFRL-AFOSR-JP-TR-2022-0018

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project develops methods and models for understanding and predicting attention dynamics across online platforms. The project consists of five successive topics focusing on measurements, data sampling, and predictive models for social processes. First, we show
ideological asymmetries in digital space and provide a set of methods to quantify attention dynamics across different social platforms, especially YouTube and Twitter on long-running controversial topics. Second, we measure the correlation between online behavior and offline attitudes and actions, grounded on the theory of discursive opportunities. Third, we present a first study on cross-partisan communications on YouTube comments and find that the crosstalk is not symmetric. Fourth, we present a first study on measurement
errors under subsampled Twitter data streams, and discuss noises and potential biases in social data. Lastly, we develop three different models that explain how social process unfolds: a mathematical relationship between self-exciting processes and stochastic epidemic models; a succinct neural model that universally approximates any point process; and a dual mixture model that is particularly suited to long-tailed data with both popular and unpopular content.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	46

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
ALAN LIN	227-7009

AOARD/AFOSR 19IOA078 – Final Report

Project title: Linking Online Attention to Measurable Actions

Principle investigator:
Lexing Xie, lexing.xie@anu.edu.au
Key personnel:
Jooyoung Lee, jooyoung.lee@anu.edu.au
Siqi Wu (ANU and University of Michigan) siqi.wu@anu.edu.au
Organizations:
School of Computing
The Australian National University, Acton, ACT 2601, Australia.
Computational Media Lab: http://cm.cecs.anu.edu.au

Collaborator (via AFSOR 19RT0797)
Yu-Ru Lin yurulin@pitt.edu
School of Computing and Information,
University of Pittsburgh, Pittsburgh, PA 15260, USA.
Computational Social Dynamics Lab: https://picsolab.github.io
Period of Performance:
Sept 1 2019 – Aug 31 2021

Project abstract

This project develops methods and models for understanding and predicting attention dynamics across online platforms. The project consists of five successive topics focusing on measurements, data sampling, and predictive models for social processes. First, we show ideological asymmetries in digital space and provide a set of methods to quantify attention dynamics across different social platforms, especially YouTube and Twitter on long-running controversial topics. Second, we measure the correlation between online behavior and offline attitudes and actions, grounded on the theory of discursive opportunities. Third, we present a first study on cross-partisan communications on YouTube comments and find that the cross-talk is not symmetric. Fourth, we present a first study on measurement errors under sub-sampled Twitter data streams, and discuss noises and potential biases in social data. Lastly, we develop three different models that explain how social process unfolds: a mathematical relationship between self-exciting processes and stochastic epidemic models; a succinct neural model that universally approximates any point process; and a dual mixture model that is particularly suited to long-tailed data with both popular and unpopular content.

Project Highlights

We begin by summarizing six project highlights during this performance period. These contributions are further detailed in the **Technical Sections** of this report.

- **Measurements**: we developed a set of metrics which quantify attention levels on social media and compare attention dynamics of opposing political users in different social movements.
- **Prediction**: we construct online attention metrics to explain offline behaviors. We systematically formulate prediction tasks in order to show which online features predict offline behaviors.
- **Data sampling**: we show that Twitter rate limit message is an accurate indicator for the volume of missing tweets and present how to estimate the entity frequency and ranking of the complete data using only the sample data.
- Three new models:
  - **Linking SIR and HawkesN**: this model is a general connection between the two model classes via three new mathematical components. The first is a generalized version of stochastic Susceptible-Infected-Recovered (SIR) model with arbitrary recovery time distributions; the second is the relationship between the (latent and arbitrary) recovery time distribution, recovery hazard function, and the infection kernel of self-exciting processes; the third includes methods for simulating, fitting, evaluating and predicting the generalized process.
  - **UNIPoint**: point processes are a useful mathematical tool for describing events over time. We provide a proof that a class of learnable functions can universally approximate any valid intensity function. This bridges the gap on the open questions of how to precisely describe the flexibility of point process models and whether there exists a general model that can represent all point processes.
  - **Dual mixture model**: it is well-known that online behavior is long-tailed, with most cascaded actions being short and a few being very long. A prominent drawback in generative models for online events is the inability to describe unpopular items well. We address these shortcomings by proposing dual mixture self-exciting processes to jointly learn from groups of cascades.

Publication list

The following publications result from this project.

- Jooyoung Lee, Siqi Wu, Ali Mert Ertugrul, Yu-Ru Lin, and Lexing Xie. *Whose Advantage? Measuring Attention Dynamics across YouTube and Twitter on Controversial Topics*, Accepted at *International AAAI Conference on Web and Social Media*, 2022.
  Summary: This work proposes a set of metrics for comparing attention consumption patterns between left-leaning and right-leaning videos across two platforms. It is shown that left-leaning videos are more viewed and more engaging, while right-leaning videos are more tweeted and have longer attention spans. It is also found that the follower networks of early adopters on left-leaning videos are of higher centrality, whereas tweet cascades for right-leaning videos start earlier in the attention lifecycle.
  Technical sections: Section 1

- [63] Siqi Wu and Paul Resnick. *Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives*. In *Fifteenth International AAAI*

Conference on Web and Social Media, Pages 808-819, 2021. https://ojs.aaai.org/index.php/ICWSM/article/view/18105
Summary: This work presents the study of cross-partisan discussions between liberals and conservatives on YouTube. We find a large amount of cross-partisan commenting, but much more frequently by conservatives on left-leaning videos than by liberals on right-leaning videos. We find that people tend to be slightly more toxic when they venture into channels with opposing ideologies, however they also receive much more toxic replies.
Technical sections: Section 3

- [66] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. *Variation across Scales: Measurement Fidelity under Twitter Data Sampling*. In *Fourteenth International AAAI Conference on Web and Social Media*, Pages 715-725, 2020. https://ojs.aaai.org/index.php/ICWSM/article/view/7337
  Summary: This work presents a set of in-depth measurements on the effects of Twitter data sampling. We validate that Twitter rate limit messages closely approximate the volume of missing tweets. We show the effects of sampling across different subjects (entities, networks, cascades), which may in turn distort the results and interpretations of measurement and modeling studies. For counting statistics such as number of tweets per user and per hashtag, we find that the Bernoulli process with a uniform rate is a reasonable approximation for Twitter data sampling.
  Technical sections: Section 4

- [33] Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. *Modeling Information Cascades with Self-exciting Processes via Generalized Epidemic Models*. In *13th International Conference on Web Search and Data Mining*, Pages 286–294, 2020. https://doi.org/10.1145/3336191.3371821
  Summary: This work establishes a general connection between the two model classes via three new mathematical components. The first is a generalized version of stochastic Susceptible-Infected-Recovered (SIR) model with arbitrary recovery time distributions; the second is the relationship between the (latent and arbitrary) recovery time distribution, recovery hazard function, and the infection kernel of self-exciting processes; the third includes methods for simulating, fitting, evaluating and predicting the generalized process.
  Technical sections: Section 5.1

- [57] Alexander Soen, Alexander Mathews, Daniel Grixti-Cheng, and Lexing Xie. *UNIPoint: Universally Approximating Point Processes Intensities*. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, Pages 9685-9694, 2021. https://ojs.aaai.org/index.php/AAAI/article/view/17165
  Summary: Focusing on the widely used event intensity function representation of point processes, we provide a proof that a class of learnable functions can universally approximate any valid intensity function. We also design and implement UNIPoint, a novel neural point process model, using recurrent neural networks to parameterise sums of basis function upon each event.
  Technical sections: Section 5.2

- [32] Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. *Describing and Predicting Online*

Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *29th ACM International Conference on Information and Knowledge Management (CIKM)*, Pages 645-654, 2020. https://doi.org/10.1145/3340531.3411861
Summary: This work addresses the shortcomings of generative models by proposing dual mixture self-exciting processes to jointly learn from groups of cascades.
Technical sections: Section 5.3

Software and Demo

This project also led to the following demo and software resources:

- Software package "Twitter-intact-stream" for constructing the complete data streams on Twitter [66] is available in a GitHub repository https://github.com/avalanchesiqi/twitter-sampling.
- The pre-trained Hierarchical Attention Network (HAN) model [63] to classify a YouTube user as conservative or liberal is available in a GitHub repository https://github.com/avalanchesiqi/youtube-crosstalk/tree/main/hnatt.
- Software and tutorial for fitting and simulating the HawkesN process [33] for social media is available in a GitHub repository https://github.com/qykong/generalized-sir-and-hawkes.

Datasets

- 134M YouTube comments from 9.3M users on 274K political videos from 1,267 US partisan media [63]. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KF5JC5.
- High sampling rate covid-tweets that are available between March and December 2020 from our work on twitter sampling [66]. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GW9GDM.
- Cross-platform dataset of three controversial topics – `BLM`, `Abortion`, `Gun Control` will soon be released. Consisting of more than one thousand YouTube videos, their popularity history, and tweet history.

Technical Section

In the following technical sections, we report in detail the main technical contributions achieved for this AFOSR project. We detail the contributions outlined in the **Project Highlights** section – quantitative analysis on cross platform attention measures is described in Sec. 1, connecting online features to predict offline behaviors is detailed in Sec. 2, findings of cross-partisan discussions between liberals and conservatives on YouTube are presented in Sec. 3, impacts of Twitter data sampling are shown in Sec. 4, and three new models (Linking SIR and HawkesN, UNIPoint, and Dual mixture model) are described in Sec. 5. Selected results for each section follow in Sec. 6. Full technical details are in our papers [63, 66, 33, 57, 32].

1 Measurements: cross platform attention

We designs several sets of metrics for the cross-platform data, in order to compare content across different political ideologies, and examine whether the differences are consistent across topics, across platforms, and over time.

We choose three controversial topics: `Abortion`, `Gun Control`, and `Black Lives Matter` (`BLM`). We use video hyperlinks to connect the content from YouTube to Twitter. A motivating example is given in Figure 1. We plot the time series of daily view count for the collected `BLM` videos from YouTube (top panel) and daily volume of tweets mentioning these `BLM` videos from Twitter (bottom panel). Both time series are further disaggregated by video uploaders' political leanings. Visually, the view count series of both left- and right-leaning videos are relatively stable in year 2017, except a sharp spike caused by the "Unite the Right rally[1]" event in Charlottesville, VA, US. On the bottom panel, the tweet count series of right-leaning videos has many spikes, which can be attributed to the upload of new videos from far-right YouTube political commentators. The measurements on YouTube and Twitter present a contrasting story: if we focus on the two weeks period after the rally, left-leaning videos attract more attention on YouTube (measured by views, left: 27.2M, right: 13.9M) while right-leaning videos have higher exposure on Twitter (measured by tweets, left: 37.5K, right: 52.3K). This example demonstrates the need for cross-platform analysis – findings on one platform may not generalize to another platform.

We design a set of metrics from publicly available data on YouTube and Twitter, which include total views, video watch engagement, tweet reactions, the evolution of attention over time, and networks among tweets and Twitter users. On YouTube, we find that left-leaning videos accumulate more views, are more engaging, and have higher viral potential than right-leaning videos. In contrast, right-leaning videos have higher numbers of total tweets and retweets on Twitter. Statistics on the unfolding speed for views and tweets show that the attention on left-leaning videos attenuates faster, while those on right-leaning videos persist for longer. These findings are not generalized across topics, i.e., we observe significant differences for `Abortion` and `Gun Control`, but not for `BLM`. These findings expand the current wisdom that liberals, who often interact with left-leaning content, are more diverse, while conservatives are more coordinated in two ways. The first is exposing the novel facet that left-leaning content attract more attention in a shorter period of time, the second is the need to contrast temporal attention statistics between platforms, such as right-leaning cascades tend to start earlier and views on right-leaning content sustain longer. Our observations paint a richer picture of attention patterns across the political spectrum, provide a basis for further studying political framing and group behavior, and supply fundamental metrics for

[1] https://en.wikipedia.org/wiki/Unite_the_Right_rally

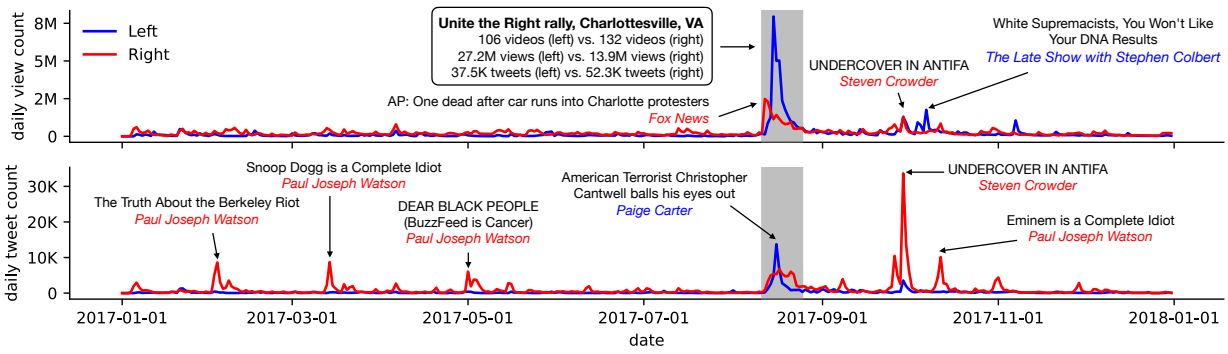


Figure 1: Attention time series (top: daily view count; bottom: daily tweet count) related to BLM throughout 2017. The view count series of both left- and right-leaning videos have a handful of sharp peaks, while the tweet count series of right-leaning videos peaks more frequently. Notable video releases (with the most views and tweets) are labeled. YouTube channel titles are *italicized* and colored by video leanings.

understanding influences that transcend platforms.

1.1 Aggregate Attention on YouTube and Twitter

We present four metrics for the total video attention on YouTube.

Total view count sums up a video's view count time series until day 120.

Relative engagement is a metric proposed in [64] for quantifying the average video watching behavior. Specifically, for each video, we first compute *average watch percentage*, defined as the total watch time divided by total number of views (both at 120 days) and then normalized by video length. The relative engagement score is the percentile ranking of average watch percentage among videos of similar lengths. It is a normalized score between 0 and 1. A higher score means more engaging, e.g., a score of 0.8 suggests that this video is on average watched for longer time than 80% videos of similar length. Note that relative engagement is shown to be stable over time, hence there is no need to examine the temporal variations of watch time, as it would strongly correlate with view counts. In this work, relative engagement is computed based on a publicly available collection of 5.3M YouTube videos [64].

Fraction of likes measures the video reaction – provided by YouTube as the total counts of likes and dislikes and collected via the thumb-up and thumb-down icon on the video page. A lower fraction of likes indicates a more diverse audience reaction to the video content. Note that the majority of videos receive a lot more likes than dislikes.

Viral potential is a positive number, representing the *expected* number of views a video will obtain if mentioned by an *average* tweet [51]. More specifically, it is the area under the impulse response function of an integral equation known as Hawkes Intensity Process (HIP) [52], which is learned for each video by using the first 120 days of tweeting and viewing history. We use this quantity rather than simply dividing the number of views by the number of tweets, because the model takes into account views that are yet to unfold due to its sustained circulation via sharing and tweeting.

On Twitter, tweets can be categorized into four types: original tweets, retweets, quotes, and replies. This leads to five counting metrics: the total number of ***tweets, original tweets, retweets, quoted tweets***, and ***replies***.

1.2 Views and Tweets over Time

Viewing half-life is measured as the number of days to achieve half of its total views at day 120.
Tweeting half-life is measured as the number of days to achieve half of its total tweets at day 120.
Tweeting lifetime is time gap between the first and the last tweets. We do not measure lifetime on viewing because the view count of a video rarely becomes zero even towards the end of the measurement period, but tweets tend to unfold much sooner.
Tweeting inter-arrival time is the average time difference between every two consecutive tweets about each video.
Accumulation of views and tweets. In addition to the summary metrics above, we also compare the attention accumulation on the left- and right-leaning videos on a daily basis. On each day, we compute the fraction of the total views that each video has achieved. This leads to two sets of samples $\{v_i^{(L)}\}_{i=1}^n$ and $\{v_j^{(R)}\}_{j=1}^m$, where n is the number of left-leaning videos and m is the number of right-leaning videos. We then compute the normalized Mann-Whitney U (MWU) statistic [40],

$$\bar{U}_v = \frac{1}{nm} \sum_i \sum_j \{I[v_i^{(L)} > v_j^{(R)}] + 0.5\, I[v_i^{(L)} = v_j^{(R)}]\}$$

Here $I[\cdot]$ is the indicator function that takes value 1 when the argument is true, 0 otherwise. The U statistic intuitively corresponds to the fraction of sample pairs $(v_i^{(L)}, v^{(R)_j})$ where the sample from left-leaning distribution is larger, accounting for ties. If the distributions of $v^{(L)}$ and $v^{(R)}$ are indistinguishable, then $\bar{U}$ would be around 0.5. We compute the statistic $\bar{U}_t$ on tweets in the same fashion, and both statistics are computed for each day. These two series of statistics allows us to quantify the differences between left- and right-leaning content, and compare the trends on the accumulation of views and tweets over time.

1.3 Videos' Tweet Cascades

We define that a *cascade* consists of a root tweet and all of its retweets, replies, and quotes. It is well-known that the vast majority of cascades in online diffusion networks are very small and only a very small fraction of cascades would become very big [21].

Based on the number of tweets in a cascade, we divide the cascades into isolated (only root tweet), small (2-4 tweets), and large (≥ 5 tweets) groups. For videos of each leaning on each topic, we compute ***the fractions of isolated/small/large cascades*** and ***the fraction tweets in each type of cascade***. These metrics quantify the structure of online diffusion and allow us to compare behavior on political controversial topics with what was known about tweeted videos in general.
Cascade start time is the percentage of accumulated views of the video when the root tweet of the cascade is posted. It measures how much view attention is accumulated on YouTube before the infusion on Twitter starts. We choose to describe cascade timing relative to the accumulation of view, rather than in absolute number of days since upload, because (1) such relative time more directly correlates the amount of cascades with respect to the views they can potentially drive (rather than through another variable, days), and (2) the percentage of views provides more granularity, since many video have all views and tweets unfold within a few days after upload.

1.4 Networks among Early Adopters on Twitter

For each video, we obtain its follower network among the early adopters. If there exists a following relationship between a pair of users, a directed edge is established. This results in one network for each shared videos. We compute a set of statistics per video, and then compare their distributions on each topic for left- and right-leaning videos. We describe two key metrics here.

Gini coefficient of indegree centrality. We calculate the indegree centrality for each node in the network. To have a video-level metric, we use the Gini coefficient, which ranges from 0 to 1 and quantifies the inequality of distribution. Overall, the Gini coefficient of indegree centrality quantifies the degree of inequality of the indegree distribution. A higher value indicates that a few early adopters are followed more by other early adopters, and a lower value indicates that the indegree distribution is more equal.

Gini coef. of closeness centrality captures the dispersion in inverse of average shortest path length from one early adopter to all other early adopters of the given video. Higher coefficient implies that a few early adopters can reach the rest of the early adopters within a few hops.

2 Prediction: predicting offline attitudes from online signals

We correlate online attention metrics from Twitter and YouTube with a range of offline opinions and surveys. We curate a comprehensive set of state-level offline data for the three topics based on surveys of abortion rate, gun ownership, fatality, and protests. We design volume- and group-based features from Twitter and YouTube measures. We conduct extensive prediction experiments to examine how indicative online signals are of related public attitudes or events observed for each topic. Our results show robust improvement over baseline models with historical offline data across nine prediction tasks.

We first describe nine different offline variables, define a set of features that profile video leaning and user ideology, followed by the prediction setting.

2.1 Signals of offline attitudes and behaviors

Broadly speaking, there are two desirable properties of offline data. The first is covering our study period, ideally spanning the time before, during, and after the data collection span. The second is being disaggregated, since the nature of controversial topics dictates that a global variable reflecting some average opinion is not meaningful. Common dimensions for disaggregation include age group, geography, socio-economic status, and other demographic attributes. Here we choose to focus on large geographic areas, because this can be inferred from Twitter profile, and arguably less sensitive than age or other demographic attributes. Given the major presence of the topics in the US, we look for opinions and behaviors statistics by state. While the three topics are of national interest and important to policy making, fine-grained datasets that are disaggregated by geographies or split into shorter time periods are largely unavailable or not up to date. We identified two different sources for each topic [16, 49, 31, 19, 9, 18], ranging from government organization, think tank, academic research, and datasets curated by the community. For the `Abortion` and `BLM` topics, one task builds on the source of public opinion, a second task on behaviors (abortion rate and protest, respectively), and a third on the change in behaviors. For `Gun Control`, both sources cover behavioral statistics (ownership and violence), we also predict change in gun violence.

Note that this state-level correlation study restricts the range of measurements that can be considered as meaningful features, i.e., it has to be directly associated with geo-located Twitter users and tweets.

The three target variables for `Abortion` are:

A1 **Public support** (for legalizing abortion) [16]: We use the *public opinion in 2014* from Pew Research Center [16] as one outcome variable. To our best knowledge, this is the most up-to-date, state-level public opinion data on this topic. Note that the data was collected for public attitude in 2014, asking participants to answer retrospectively for 2014, and the results were made available in 2017. This exemplified collecting such data is a time-consuming undertaking. As our online data span across 2017, this analysis is a correlation that goes *backward in time* to help understand how online signals correlate with historic trends. The most relevant *other* variable, *abortions by state of residence in 2014* was chosen as the baseline variable.

A2 **Abortion rate** [49]: We use the *abortions per 1000 women in 2017* from the Guttmacher Institute [49] as one outcome variable. While the abortion rates do not necessarily reflect the public opinions on the topic and there are other reasons can impact the abortion rates, the disparities in abortion rates over states may still reflect the socioeconomic states that are correlated with the public support – e.g., the costs of obtaining an abortion decrease while public support for legal abortion increases [35]. This prediction task aims to establish whether there are correlations between the online signals and the abortion trend. The past abortion rates (in 2014) were chosen as the baseline variable.

A3 **Change in abortion rate**: With the lack of public opinion data after 2017, and the general observations that public support for legal abortion has been stable over time [16], we chose to predict the change in abortion rate. We calculate the difference of the abortion rates in 2014 and 2017 as the outcome variable, and the task is to predict the change using the offline signals with the baseline variable being the past abortion rates (in 2014).

The three outcome variables for `Gun Control` are:

G1 **Gun ownership** [31]: We use the *gun ownership* survey for 2015 [31]. This is the closest state-level public opinion data to our study period. We consider the statistics of gun ownership as a proxy for the opposing stance for gun control, as argued by [48], those "who own guns largely disagree with non-owners on gun policy, but some proposals have supports from both groups". Similar to task **A1**, this is a backward correlation task. The historic prevalence of gun violence across states (*fatal injuries by firearms in 2007-2016*) was chosen to be the baseline variable.

G2 **Gun violence** [19]: We use the (*fatal injuries by firearms in 2018*) as the outcome variable for a *forward* correlation task, with the previous gun violence data (*fatal injuries by firearms in 2007-2016*) as the baseline variable.

G3 **Change in gun violence**: We calculate the difference in gun violence incidence between 2017 and 2018 as the outcome variable, with the historic prevalence of gun violence as the baseline variable.

For BLM, the relevant statistics nearly cover our study period, allowing us to design three forward correlation tasks:

B1 **Support** for BLM [9], We use the state-level *BLM support in 2018* (Jan. 2018) from Civiqs [9] as the outcome variable. As for baseline, although the information *BLM support in 2017* is available and can be most relevant, we found the support nearly unchanged across the two years (with more than 98% correlations between consecutive surveys) and thus the past information was not meaningful to serve as a baseline. Instead, we use the past observation of the protest events *BLM protests in 2017*, available from *Elephrame* [18], as a baseline variable.

B2 **Protests** from *Elephrame* [18]: We use the number of *BLM protests in 2018* in each state as the outcome variable with the past protest prevalence (in 2017) as the baseline variable.

B3 **Change in protests**: We calculate the difference in the number of protests per state from 2017 to 2018 as the outcome variable, with the past protest prevalence as the baseline variable.

A note on terminologies: the goal of our study is to examine whether or not there are sufficient *correlations* between online measurements and the offline variables presented above. This setting does not support *causal* reasoning. Thus the terms of *forward* and *backward* correlation (in time) in this Section. On the other hand, performing statistical estimation on hold-out data is called *prediction*, as commonly used in machine learning literature. The next two subsection uses *prediction* in this sense *without* implying making predictions for the future.

2.2 Features of online attention on Twitter and YouTube

For the prediction input, we need to construct a set of state-level features of online attention. Since YouTube does not provide geographical sources of view count to the general public, we use the geo-located tweets from Twitter to estimate the number of views from different states on YouTube.

Parsing tweet locations and estimating geographical YouTube views. We infer a Twitter user's location by detecting US state names, state abbreviations, and US city names[2] in the user profile text. Cities with multiple possible states are discarded to avoid introducing data noises. In total, we successfully map the state-level location for 27,152 (35.6%) users / 38,161 (35.7%) tweets for Abortion, 55,349 (35.5%) users / 97,727 (36.1%) tweets for Gun Control, and 85,519 (35%) users / 197,173 (35.2%) tweets for BLM.

For YouTube, we use the video *virality* scores derived from the HIP model (see [52]) together with the geo-located tweets to estimate the amount of views originated from each of the 50 states. Specifically, the estimated view count for a given YouTube video from a state is computed by aggregating the observed geo-located tweets from this state following the HIP model. The total interest on YouTube for a topic is computed by summing the estimated views over all videos belonging to that topic.

Constructing features. We construct a list of predictive features for each state by using 9 metrics that can be categorized into 3 groups: (a) the number of *unique users*; (b) numbers of *original tweets*, *retweets*, *replies*, and *quotes*, (c) numbers of expected YouTube views generated by *original tweets*, *retweets*, *replies*, and *quotes* – by weighting each tweet with the *virality score* estimated

[2]The list of US cities: https://www.britannica.com/topic/list-of-cities-and-towns-in-the-United-States-2023068

Table 1: Base matrix that generates the predictive features. In our study, liberals mostly have aligned stance with left-leaning videos.

	Liberal	Conservative
Left-leaning	L-lib(x)	L-con(x)
Right-leaning	R-lib(x)	R-con(x)

for the video it mentions. For a given metric *x*, we use a 2x2 matrix (see Table 1) to reflect the conceptual breakdown of activities performed by users of different ideologies on videos of different leanings. We compute three types of features that contrast the activities on left- and right- leaning videos. There are two important considerations that led to the current feature design. The first is that it is the relative (rather than absolute) prevalence between the different political forces (e.g. whether left/liberal is more prevalent than right/conservative that determines the overall political polarity (e.g. voting outcomes, or support for gun control). The relative prevalence is also invariant to the population of the state, whereas the total prevalence is not. The second is the practical constraint that each feature needs to be geolocated to US states. This makes aggregate YouTube views and engagement metrics not applicable, and only those that are directly connected to Tweets are applicable (e.g. virality score). Note that the normalization factor is the total volume, calculated as, *total* = *L-lib*(x) + *L-con*(x) + *R-lib*(x) + *R-con*(x).

- *activity-L*: normalized volume of metric *x* on the left-leaning videos, computed as (*L-Lib*(x) + *L-con*(x)) / *total*.
- *activity-R*: normalized volume of metric *x* on the right-leaning videos, computed as (*R-lib*(x) + *R-con*(x)) / *total*.
- *tension*: the ratio of the activities on the left-leaning videos to those of the right-leaning videos, computed as (*L-Lib*(x) + *L-con*(x)) / (*R-lib*(x) + *R-con*(x)).
- *disparity-L*: the difference between the percentages of metric *x* of left-leaning and right-leaning groups related to the left-leaning videos, computed as (*L-Lib*(x) - *L-con*(x)) / *total*.
- *disparity-R*: the difference between the percentages of metric *x* of left-leaning and right-leaning groups related to the right-leaning videos, computed as (*R-lib*(x) - *R-con*(x)) / *total*.

For example, for the *retweet* in topic `BLM`, its *disparity-L(retweet)* feature is computed as the the difference between the fractions of retweets by *liberal* and *conservative* users on the *left-leaning* videos, i.e. *L-Lib(retweet) - L-con(retweet) / total*.

The disaggregation by leanings and groups increases the size of the feature set and feature sparseness. We choose to remove features derived from the number of replies and quoted tweets due to their sparseness (e.g., states may have zero quoted tweets). And we keep one metric, the total number of *unique users* per state without disaggregating it into any of the four types. In total, we have at most 1 + 4 metrics x 5 types = 21 features for each state for each topic.

2.3 Prediction setup and evaluation

Each of the prediction tasks was formulated as a regression task with a baseline variable and the variables for online signals as features. The *baseline* variable is the most relevant offline data, typically the most similar *other* offline variable, or the previous value in time. This reflects our "best knowledge" before obtaining any online signals. The regression task estimates $f([X_d^b; X_d^{tw}; X_d^{yt}]) \rightarrow y_d$ where X_d^b is the baseline feature, X_d^{tw} is the set of Twitter engagement features and X_d^{yt} is the set of YouTube virality features derived from location (state) d. y_d is the outcome variable, describing the state about a given *topic* at location d.

We standardize all features before To avoid over-fitting, we incorporate different regularization techniques including Ridge, Lasso, and Elastic Net in linear regression to eliminate insignificant or correlated features. We use cross-validation to determine the hyper-parameters. Our experiment shows that the results of using Lasso and Elastic Net regularization are similar (with only 1-2% difference in all prediction tasks) and both are significantly better than the Ridge regression. Thus the final results were reported based on Lasso regression. The performance was evaluated in terms of *RMSE* (root-mean-square-error) and R^2 based on 10-fold cross-validation.

In each prediction task, we test the impact of features through four different models: (1) *Baseline* only includes the specific baseline (offline) variable, (2) *Model-All* includes *activity*, *tension* and *disparity* features, as well as the baseline variable, (3) *Model-Tension* includes the *tension* and *activity* features, and (4) *Model-Disparity* includes *disparity-L*, *disparity-R* and *activity* features.

3 Prevalence analysis: cross-partisan discussions on YouTube

We studied three questions related to the prevalence of cross-partisan discussions between liberals and conservatives. We first quantified the portions of cross-partisan comments from a user-centric view, then and from a video-centric view. Finally, we investigated whether the extent of cross-talk varied on different media types.

3.1 How Often Do Users Post on Opposing Channels?

We empirically measured how often a YouTube user commented on videos with opposite ideologies. For active users with at least 10 comments, we counted the frequencies that they posted on left-leaning and right-leaning videos. The HAN model classified 90.1% of the active users as either liberal or conservative. Among them, 62.2% of liberals posted at least once on right-leaning videos, while 82.3% conservatives posted at least once on left-leaning videos.

Figure 2 plots the fraction of users' comments that were cross-partisan, as a function of how prolific the users were at commenting. Overall, conservatives posted much more frequently on left-leaning videos (median: 22.2%, mean: 33.9%) than liberals on right-leaning videos (median: 4.8%, mean: 15.6%). The fractions of cross-partisan comments from liberals were largely invariant to user activity level (Figure 2a). By contrast, prolific conservatives disproportionately commented on left-leaning videos (Figure 2b). The few most prolific conservative commenters with more than 10,000 comments made more than half their comments on left-leaning videos, suggesting a potential trolling behavior. Nevertheless, even for less prolific conservatives, they still commented on left-leaning videos far more frequently than liberals did on right-leaning videos.

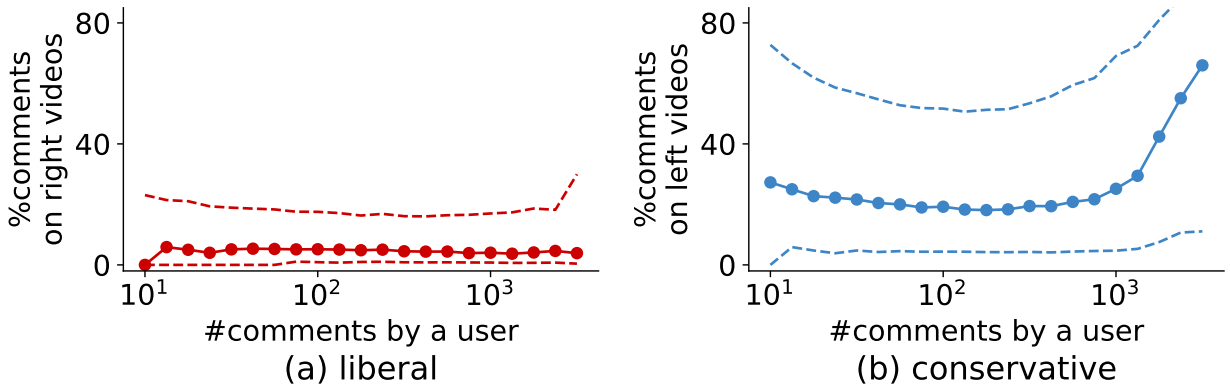


Figure 2: Analysis of users' cross-partisan interaction: conservatives were more likely to comment on left-leaning videos than liberals on right-leaning videos. The x-axis shows the total number of comments from the user, split into 21 equally wide bins in log scale. The y-axis shows the percentage of comments on videos with opposing ideologies. Within each bin, we computed the 1st quartile, median, and 3rd quartile. The line with circles connects the medians, while the two dashed lines indicate the inter-quartile range.

3.2 How Many Cross-Partisan Comments Do Videos Attract?

We also quantified cross-partisan discussions on the video level. For videos with at least 10 comments, we counted the number of comments posted by liberals and posted by conservatives. Figure 3 plots the fraction of cross-partisan comments on (a) left-leaning and (b) right-leaning videos. We make two observations here:

First, higher fraction of cross-partisan comments occurred on left-leaning videos (median: 28%, mean: 29.5%) than on right-leaning videos (median: 8.6%, mean: 13.4%). This result provides a new angle for explaining the creation of conservative echo chambers online [20, 36]. For random viewers who watched right-leaning videos and browsed the discussions there, they would be exposed to very few comments from liberals. On the other hand, users might experience relatively more balanced discussions on the left-leaning videos since about one in three comments there was made by conservatives.

Second, the correlations between video popularity and cross-partisan comments were opposite for left-leaning and right-leaning videos. When left-leaning videos attracted more views, the fraction of comments from conservatives became lower. On the contrary, the fraction of comments from liberals was higher on right-leaning videos when the videos attracted more views. This finding reveals potentially different strategies when politically polarized users carry out cross-partisan communication: while conservatives occupy the discussion spaces in less popular left-leaning videos, liberals largely comment on high profile right-leaning videos.

3.3 Which Media Types Attract More Cross-Partisan Comments?

We examined the extent of cross-talk in the four media types. We excluded videos with less than 10 comments and then removed channels with less than five videos. For each channel, we computed the mean fraction of cross-partisan comments over all of its videos, dubbed $\bar{\eta}(c)$. The metric $\bar{\eta}(c)$ can be interpreted as the expected rate of cross-talk appearing on an average video of a given

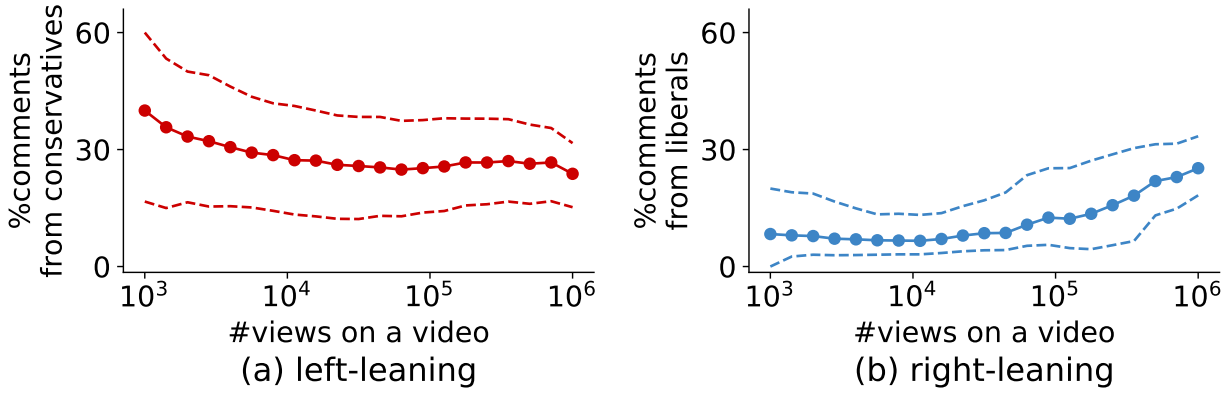


Figure 3: Analysis of videos: there were more cross-partisan comments on left-leaning videos than these on right-leaning videos. The x-axis shows the video view counts and it is divided into 21 equally wide bins in log scale. Each bin contains 553 to 7,527 videos. The lines indicate median and inter-quartile range.

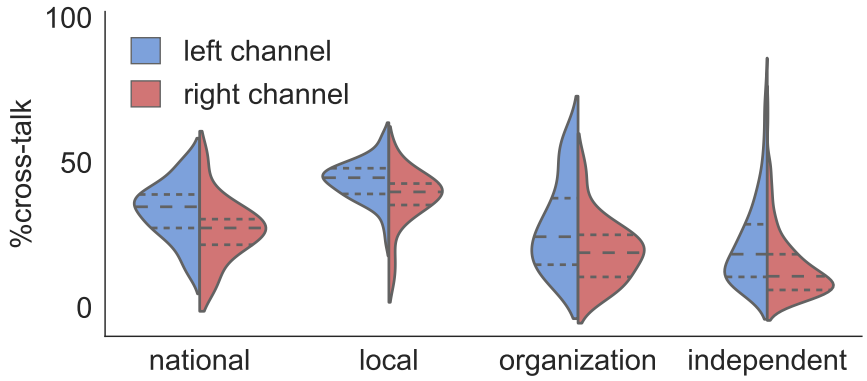


Figure 4: Analysis of media types. Local news channels attracted balanced audiences. By contrast, on right-leaning independent media, very few comments were posted by liberals. The outlines are kernel density estimates for the left-leaning and right-leaning channels. The center dashed line is the median, whereas the two outer lines denote the inter-quartile range.

channel c.

Figure 4 shows the distributions of $\bar{\eta}(c)$ in a violin plot, disaggregated by media types and political leanings. Right-leaning channels received relatively fewer cross-partisan comments than the corresponding left-leaning channels across all four media types (statistically significant in one-sided Mann-Whitney U test at significance level of 0.05). In particular, half of right-leaning independent media had fewer than 10.7% comments from liberals, exposing their audience to a more homogeneous environment. For example, "Timcast", who was the most commented right-leaning independent media in our dataset, had only 3.3 comments from liberals in every 100 comments. This phenomenon stresses the potential harm for those who engage with ring-wing political commentary, because the discussions there often happen within the conservative echo chambers, which may in turn foster the circulation of rumors and misinformation [24]. On the other hand, the two prominent US news outlets – CNN and Fox News – were both crucial ground for cross-partisan

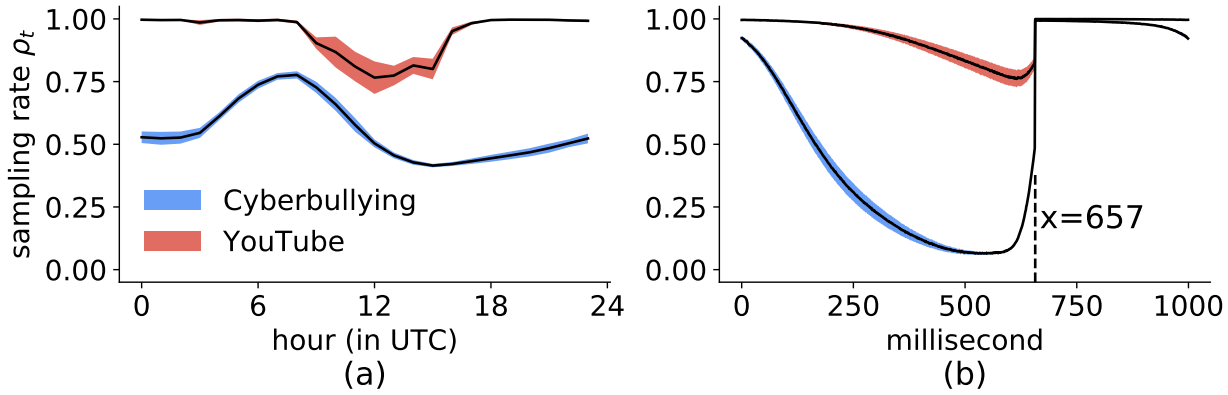


Figure 5: Sampling rates are uneven **(a)** in different hours or **(b)** in different milliseconds. black line: temporal mean sampling rates; color shades: 95% confidence interval.

discussions, having $\bar{\eta}(c)$ of 35.8% and 31%, respectively.

4 Data sampling: measurement fidelity under Twitter data sampling

Researchers have used Twitter data as a lens to understand political elections [6], social movements [13], information diffusion [69], and many other social phenomena. Twitter offers two streaming APIs for free, namely *sampled* stream and *filtered* stream. The filtered stream tracks a set of keywords, users, languages, and locations. When the matched tweet volume is above a threshold, Twitter subsamples the stream, which compromises the completeness of the collected data. In this work, we focus on empirically quantifying the data noises resulted from the sampling in the filtered stream and its impacts on common measurements.

In this work, we study the randomness of Twitter sampling – do all tweets share the same probability of missing? This is relevant because uniform random sampling creates representative samples. When the sampling is not uniform, the sampled set may suffer from systematic biases, e.g., some tweets have a higher chance of being observed. Consequently, some users or hashtags may appear more often than their cohorts. We tackle the uniformity of the sampling when accounting for the tweet timestamp, language, and type.

Tweet timestamps. Figure 5(a) plots the hourly sampling rates. CYBERBULLYING dataset has the highest sampling rate (ρ_t=78%) at UTC-8. The lowest sampling rate (ρ_t=41%) occurs at UTC-15, about half of the highest value. YOUTUBE dataset is almost complete (ρ_t=100%) apart from UTC-8 to UTC-17. The lowest sampling rate is 76% at UTC-12. We posit that the hourly variation is related to the overall tweeting dynamics and the rate limit threshold (i.e., 50 tweets per second): higher tweet volumes yield lower sampling rates. Figure 5(b) shows the sampling rate at the millisecond level, which curiously exhibits a periodicity of one second. In CYBERBULLYING dataset, the sampling rate peaks at millisecond 657 (ρ_t=100%) and drops monotonically till millisecond 550 (ρ_t=6%) before bouncing back. YOUTUBE dataset follows a similar trend with the lowest value (ρ_t=76%) at millisecond 615. This artifact leaves the sample set vulnerable to automation tools. Users can deliberately schedule tweet posting time within the high sampling rate period for

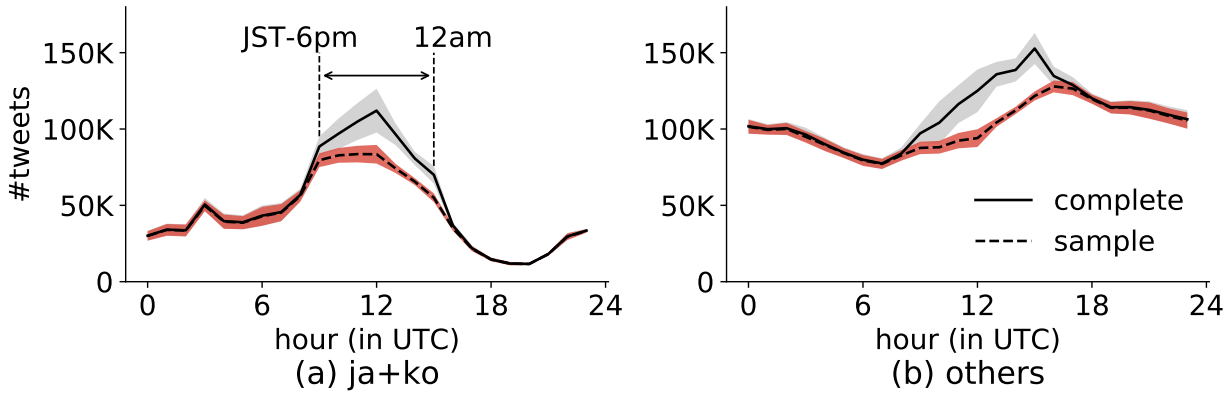


Figure 6: Hourly tweet volumes in YOUTUBE dataset. **(a)** Japanese+Korean; **(b)** other languages. black line: temporal mean tweet volumes; color shades: 95% confidence interval.

	CYBERBULLYING		YOUTUBE	
	complete	sample	complete	sample
%root tweets	14.28%	14.26%	25.90%	26.19%
%retweets	64.40%	64.80%	62.92%	62.51%
%quotes	7.37%	7.18%	3.44%	3.40%
%replies	13.94%	13.76%	7.74%	7.90%

Table 2: The ratios of the 4 tweet types (root tweet, retweet, quote, and reply) in the complete and the sample sets.

inflating their representativeness, or within the low sampling rate period for masking their content in the public API.

Tweet languages. Some languages are mostly used within one particular timezone, e.g., Japanese and Korean[3]. The temporal tweet volumes for these languages are related to the daily activity patterns in the corresponding countries. We break down the hourly tweet volumes of YOUTUBE dataset into Japanese+Korean and other languages. The results are shown in Figure 6. Altogether, Japanese and Korean account for 31.4% tweets mentioning YouTube URLs. The temporal variations are visually different – 48.3% of Japanese and Korean tweets are posted in the evening of local time (JST-6pm to 12am), while tweets in other languages disperse more evenly. Because of the high volume of tweets in this period, sampling rates within UTC-9 to UTC-15 are lower (see Figure 5(a)). Consequently, "ja+ko" tweets are less likely to be observed (89.0% in average, 80.9% between JST-6pm and 12am) than others (92.9% in average).

Tweet types. Twitter allows the creation of 4 types of tweets. The users create a *root tweet* when they post new content from their home timelines. The other 3 types are interactions with existing tweets: *retweets* (when users click on the "Retweet" button); *quotes* (when users click on the "Retweet with comment" button); *replies* (when users click on the "Reply" button). The relative ratios of different types of tweets are distinct for the two datasets (see Table 2). CYBERBULLYING has higher ratios of retweets, quotes, and replies than YOUTUBE, implying more interactions among users. However, the ratios of different types are very similar in the sampled versions of both

[3]Japanese Standard Time (JST) and Korean Standard Time (KST) are the same.

datasets (max deviation=0.41%, retweets in YOUTUBE dataset). We conclude that Twitter data sampling is not biased towards any tweet type.

5 Modeling social process

Here, we detail our proposed models for the diffusion of information in online media and temporal event dynamics under external influence.

5.1 Information cascades: self-exciting processes via generalized epidemic models

HawkesN process is a finite-population variant of the Hawkes process [50]. Assuming the diffusion occurs in a fixed population of size N, the event intensity is modulated by the proportion of remaining population:

$$\lambda^H(t) = \frac{N - N_t}{N} \sum_{t_i < t} \phi(t - t_i) \tag{1}$$

N_t is the number of events up to time t, the background intensity μ is set to zero and the first event happens at time 0, i.e., $N_0 = 1$.

The Hawkes process with exponential kernels and stochastic SIR process have been recently shown [50] to share a connection via the infection intensity function when the recovery time in the SIR model is *latent*. However, this result is restricted to one particular parametric family of self-exciting processes, whereas Hawkes processes allow a richer set of kernel functions, and an inequality of the connection has been overlooked. These observations lead to the question: **How to both broaden and deepen the connection between epidemic models and Hawkes processes?** The broadening is with respect to arbitrary recovery time distributions and kernel functions, while the deepening is with respect to the mathematical relationships between two model classes. To address these, we propose a generalized stochastic SIR process in which infected individuals recover independently following an arbitrary distribution of recovery times. Next, we link this process to a finite-population Hawkes process (dubbed *HawkesN* [50]) by showing that the Complementary Cumulative Distribution Function (CCDF) of the recovery time (in SIR), given the infection event history, is an upper bound of the HawkesN kernel. We derive relationships among three key functions: the kernel function in HawkesN, the SIR recovery time distribution, and the recovery hazard function.

5.1.1 SIR with general recovery distributions

First, we present a generalized stochastic SIR model with an arbitrary recovery time distribution, and next we reveal the connection between the general stochastic SIR and HawkesN. Then we extend the generalized SIR model with concepts from the Hawkes models.

The stochastic SIR process is defined by an infection event intensity function $\lambda^I(t)$ and a recovery event intensity function $\lambda^R(t)$ [67]

$$\lambda^I(t) = \beta \frac{S_t}{N} I_t; \quad \lambda^R(t) = \gamma I_t \tag{2}$$

where β and γ are known as the infection rate and the recovery rate in SIR terminology. The total infection rate is proportional to the susceptible population S_t and the infected population I_t.

The stochastic SIR process implicitly assumes that recovery times of infected individuals are exponentially distributed. We relax this assumption by letting recovery times follow an arbitrary distribution $f(t)$. The recovery intensity for each individual is given by the hazard function $h(t)$ [10], i.e., the recovery time distribution conditioned on recovering after time t:

$$h(t) = \frac{f(t)}{\int_t^\infty f(\tau)d\tau} \tag{3}$$

Considering that individuals recover independently, the overall recovery event intensity is the superposition of recovery intensities of the individuals still infected at time t:

$$\lambda^R(t) = \sum_{t_i^I \in \mathcal{H}_t^I} h(t - t_i^I) = \sum_{t_i^I \in \mathcal{H}_t^I} \frac{f(t - t_i^I)}{\int_{t-t_i^I}^\infty f(\tau)d\tau} \tag{4}$$

The overall infection event intensity remains unchanged.

To the best of our knowledge, this is the first work presenting this generalized SIR with arbitrary recovery distributions.

5.1.2 Marginalizing over recovery events

One of the challenges for using the SIR model for social media diffusion is that the definitions of infection and recovery are not straightforward. Infection events can be interpreted as posting, sharing or retweeting, and they are usually recorded in data traces; recovery events can be the times when these posts or discussion topics lose traction, which are rarely directly observable. This observation implies that one may treat recovery events as *latent*, and examine the expected process after marginalizing over them.

We use $\mathbb{E}_{\{t_i^R | \mathcal{H}_t^C\}}\left[\lambda^I(t)\right]$ to denote the expected infection intensity over all recovery event times up to time t:

$$\begin{aligned}
\mathbb{E}_{\{t_i^R | \mathcal{H}_t^C\}}\left[\lambda^I(t)\right] &\overset{\text{(a)}}{=} \beta \frac{S_t}{N} \sum_{t_i^I \in \mathcal{H}_t^C} \int_{t-t_i^I}^\infty f(\tau_i \mid \mathcal{H}_t^C) d\tau_i \\
&\overset{\text{(b)}}{\geq} \beta \frac{S_t}{N} \sum_{t_i^I \in \mathcal{H}_t^C} \int_{t-t_i^I}^\infty f(\tau_i) d\tau_i
\end{aligned} \tag{5}$$

Eq. (5a) follows from [50]. Step (b) is because, given $\mathcal{H}_t^C$ an infection history observed up to time t, the recovery event time of the i^{th} individual t_i^R ($i \in U(\mathcal{H}_t^C)$) is dependent on the entire $\mathcal{H}_t^C$. Figure 7 illustrates this dependence with the red recovery event being an invalid candidate for t_1^R given $\mathcal{H}_t^C = \{t_1^I, t_2^I, t_3^I, t_4^I\}$. Intuitively, if the first individual recovers at the time of the red event, there will be zero infected individuals afterwards, rendering impossible the rest of the diffusion. We simplify the dependence using the inequality in Eq. (5b) to the recovery time distribution $f(t)$. We show that

$$\int_{t-t_i^I}^\infty f(\tau_i \mid \mathcal{H}_t^C) d\tau_i \geq \int_{t-t_i^I}^\infty f(\tau_i) d\tau_i \tag{6}$$

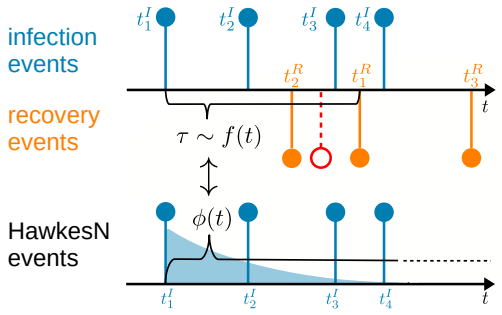


Figure 7: A sample stochastic SIR process including an infection event history until time t, i.e., $\mathcal{H}_t^C = \{t_1^I, ..., t_4^I\}$, and recovery events $\{t_2^R, t_1^R, t_3^R\}$. Infected individuals recover at time intervals τ following a distribution $f(t)$. The bottom plot presents a corresponding realization of HawkesN events. HawkesN events generate descendants with the intensity rate $\phi(t)$. A connection between $f(t)$ and $\phi(t)$ is explored when $f(t)$ is assumed of arbitrary parametric forms. The red color marks an invalid recovery event given upcoming infections.

with the left and right terms being equal when $t_i^I = \max\{\mathcal{H}_t^C\}$.

Comparing Eq. (5b) and eq. (1), both $N - N_t$ (for HawkesN) and $S_t = N - C_t$ (for SIR) stand for the size of remaining susceptible population — hence the scaling factors S_t/N and $(N - N_t)/N$ are equivalent. Also, both Eq. (5b) and eq. (1) sum over the infected population, and the integral in Eq. (5a) is a function of time since infection $t - t_i^I$. Therefore, marginalizing the recovery events reduces the infection intensity of the stochastic SIR to a lower bound — the HawkesN intensity — as long as the following relationship between the HawkeN kernel and recover time distribution holds:

$$\phi(t) = \beta \int_t^\infty f(\tau) d\tau \tag{7}$$

We can express $f(t)$ in terms of $\phi(t)$. $f(t)$ is a probability density function which implies $f(t) \geq 0$ and $\int_0^\infty f(\tau)d\tau = 1$, leading to $\phi(0) = \beta$:

$$f(t) = -\frac{\phi'(t)}{\phi(0)} \tag{8}$$

where we assume $\lim_{t\to\infty} f(t) = 0$. eq. (7) and eq. (8) spell out the closed-form relationship between the recovery time distribution $f(t)$ of the stochastic SIR and the kernel function $\phi(t)$ of the HawkesN process. From eq. (7), we note that this relationship only holds when $\phi(t)$ is a monotonically decreasing function. Incorporating eq. (8) into eq. (3), we can express the recovery hazard function in terms of the HawkesN kernel:

$$h(t) = -\frac{\phi'(t)}{\phi(t)} \tag{9}$$

Given that $\phi(t)$ is monotonically decreasing, $-\phi'(t)$ and $h(t)$ are non-negative.

Table 3 lists six examples of HawkesN kernels, with their corresponding recovery time distributions and recovery hazard functions. The first three rows show the linear, quadratic, and Gaussian kernels, followed by the Tsallis Q-Exponential kernel used in quantum optics and atomic physics [37]. The last two examples are the exponential kernel function and the power-law kernel function, widely used for financial data, geophysics, and information diffusion [27, 4, 42].

Table 3: Examples of HawkesN kernel functions $\phi(t)$, the corresponding SIR recovery time distributions $f(t)$ and hazard functions $h(t)$ following Eqs. (7)(8)(9). Parameter ranges: $\theta > 1$ for Tsallis Q-Exponential kernel, $\kappa > 0$, $\theta > 0$, $c > 0$ for all others.

HawkesN Kernel Name	HawkesN Kernel Function $\phi(t)$	SIR Recovery Time Distribution $f(t)$	SIR Recovery Hazard $h(t)$	Time Constraint t
Linear	$-\kappa\theta t + \kappa$	θ	$\frac{\theta}{-\theta t + 1}$	$(0, \frac{\kappa}{\theta})$
Quadratic	$\kappa\frac{\theta^2}{4}t^2 - \kappa\theta t + \kappa$	$-\frac{\theta^2}{2}t + \theta$	$\frac{\theta^2 t - 2\theta}{\theta^2 t^2 - 4\theta t + 4}$	$(0, \frac{2}{\theta})$
Gaussian	$\kappa e^{-\frac{t^2}{2\theta^2}}$	$\frac{t}{\theta^2}e^{-\frac{t^2}{2\theta}}$	$\frac{1}{\theta^2}t$	$(0, \infty)$
Tsallis Q-Exponential [37]	$\kappa\left[1 + (\theta - 1)t\right]^{\frac{1}{1-\theta}}$	$\left[1 + (\theta - 1)t\right]^{\frac{\theta}{1-\theta}}$	$1 + (\theta - 1)t$	$(0, \infty)$
Exponential [27]	$\kappa\theta e^{-\theta t}$	$\theta e^{-\theta t}$	θ	$(0, \infty)$
Power-law [42]	$\kappa(t + c)^{-(1+\theta)}$	$c^{1+\theta}(1 + \theta)(t + c)^{-(2+\theta)}$	$\frac{1 + \theta}{t + c}$	$(0, \infty)$

5.1.3 Marked stochastic SIR

In real data and apart from event times, additional information about individual events is available, such as the user profile of a retweet event or patient characteristics in epidemics. Mathematically, the event history $\mathcal{H}_m^C = \{(t_1^I, m_1), ..., (t_n^I, m_n)\}$ is a sequence of pairs of event times and extra event information also known as *event marks*. To leverage this information, marked variations of Hawkes process models are proposed to incorporate event marks as a scaling factor of kernel functions [27]. This idea leads to a marked variation of the HawkesN model, with the intensity function as:

$$\lambda_m^H(t) = \frac{N - N_t}{N} \sum_{(t_i^I, m_i) \in \mathcal{H}_m^I(t)} m_i^\rho \phi(t - t_i^I) \tag{10}$$

where ρ controls a warping effect for the mark. Using the generalized connection, we are able to obtain a marked stochastic SIR model, whose infection intensity function is

$$\lambda_m^I(t) = \beta \frac{S_t}{N} \sum_{(t_i^I, m_i) \in \mathcal{H}_m^I(t)} m_i^\rho \tag{11}$$

where, comparing to eq. (2), I_t was decomposed to $\sum_{(t_i^I, m_i) \in \mathcal{H}_m^I(t)} m_i^\rho$ to account for the individual mark information. The recovery intensity $\lambda_m^R(t)$ is identical to its unmarked counterpart in eq. (4).

5.1.4 Branching factor for SIR

The basic reproduction number R_0 is an important quantity in epidemic models for determining whether an epidemic is likely to occur [1]. This quantity conceptually connects to the branching factor n^* from Hawkes processes which is defined as the expected number of events generated by a single infection event [50], i.e., $n^* = \int_0^\infty \phi(\tau)d\tau$. Building upon this observation and eq. (7), we define R_0 for stochastic SIR with a general recovery time distribution as

$$R_0 = n^* = \beta \int_0^\infty \int_\eta^\infty f(\tau)d\tau d\eta \tag{12}$$

Algorithm 1 Simulating generalized stochastic SIR

Input: Recovery time distribution $f(t)$, parameters $\{N, \beta\}$
Output: Infection event times $\mathcal{H}^C$ and recovery event times $\mathcal{H}^R$

Set current time $T = 0$.
Initialize $\mathcal{H}^C = \{0\}$ with one initial infection at time 0.
Initialize $\mathcal{H}^R = \{\eta\}$ where $\eta \sim f(t)$ and $t_1^R = \eta$.
while $|\mathcal{H}^C| < N$ **do**
  $s = -\frac{log(u)}{\lambda *}$ where $u \sim U(0, 1)$
  Compute $\Lambda^I(t) = \int_0^t \lambda^I(\eta) d\eta$ from $\mathcal{H}^C, \mathcal{H}^R$
  $T = T + (\Lambda^I)^{-1}(s)$
  if $T = \infty$ **then**
    break *// No infection will occur*
  else
    $\eta \sim f(t)$ *// Draw recovery time, update histories*
    $\mathcal{H}^R = \mathcal{H}^R \cup \{T + \eta\}, \mathcal{H}^C = \mathcal{H}^C \cup \{T\}$
return $\mathcal{H}^C, \mathcal{H}^R$

For marked variations, this quantity is computed by taking expectation over the distribution of event marks. Particularly, for retweet cascades where the event marks are the count of user followers, a power law distribution $P(m) = (\alpha - 1)m^{-\alpha}$ of exponent $\alpha = 2.016$ is determined by [42]. We obtain

$$R_0 = n^* = \beta \frac{\alpha - 1}{\alpha - 1 - \rho} \int_0^\infty \int_\eta^\infty f(\tau) d\tau d\eta \tag{13}$$

5.2 Universally approximating Point processes intensities

To represent the influence of past events on future events, point process intensity functions $\lambda^\star(t)$ are often continuous between events $(t_{i-1}, t_i]$; with discontinuities only possible at events. For example, the intensity function of the Hawkes process has discontinuities at each event. Intuitively, this piece-wise continuous characterisation of the intensity function encodes the belief that the process only significantly changes its behaviour when new information (an event) is observed. As such, there are two behaviours of a point process we need to approximate: (1) the continuous intensity function segment between consecutive events, given a fixed event history; and (2) the change in the point process intensity function when an event occurs, so that we can approximate the jump dynamics between events.

We consider an intensity function $\lambda^\star(t)$ with fixed observation period $(0, T]$. The intensity function can be segmented by the event times of an event sequence $(t_0, t_1], (t_1, t_2], \ldots, (t_{N-1}, t_N], (t_N, t_{N+1}]$, where $t_{N+1} = T$. Given a piece-wise continuous intensity function, the segmented intensity function is continuous: $u_i(\tau) = \lambda^\star(t)$ for $t \in (t_{i-1}, t_i]$, where $\tau = t - t_{i-1} \in (0, t_i - t_{i-1}]$. Thus to approximate the intensity function between consecutive events, we learn a function $\hat{u}(\tau; p_i)$, parameterised by p_i, to approximate *any* of the segmented intensity functions $u_i(\tau)$, where each segment only differs in parameterisation. Then to approximate the jump dynamics of the intensity func-

tion we utilise the RNN approximation of a dynamic system, which dictates how the parameters p_i change over time.

To quantify the quality of an approximation, we use the uniform metric between two functions $f, g : X \to \mathbf{R}$,

$$d(f, g) = \sup_{x \in X} |f(x) - g(x)|. \tag{14}$$

This metric is the maximum difference of the two functions over a shared (compact) domain X. The uniform metric has been used to prove universal approximation properties for neural networks [28, 14] and RNNs [56]. Given classes of functions $\mathcal{F}$ and $\mathcal{G}$, $\mathcal{F}$ is a universal approximator of $\mathcal{G}$ if for any $\varepsilon > 0$ and $g \in \mathcal{G}$, there exists an $f \in \mathcal{F}$ such that $d(f, g) < \varepsilon$. An equivalently expression is: $\mathcal{F}$ is uniformly dense in $\mathcal{G}$.

5.2.1 Approximation Between Two Events

To approximate the time shifted non-negative functions $u_i(\tau)$, we first introduce transfer functions f_+ (Definition 1). We then prove that the class of composed function $f_+ \circ \mathcal{F}$ preserves uniform density (Theorem 1). Given this theorem, we provide a method for constructing uniformly dense classes with sums of basis functions $\Sigma(\phi)$ (Definition 2) which are in turn uniformly dense after composing with f_+ (Corollary 1). We further provide a set of suitable basis functions (Table 4).

Formally, we define the *M-transfer functions* which maps negative outputs of a function to positive values.

Definition 1. *A function $f_+ : \mathbf{R} \to \mathbf{R}_+$ is a M-transfer function if it satisfies the following:*

1. *f_+ is M-Lipschitz continuous;*
2. *$\mathbf{R}_{++} \subseteq f_+[\mathbf{R}]$;*
3. *And f_+ is strictly increasing on $f_+^{-1}[\mathbf{R}_{++}]$.*

Definition 1 provides a wide range of functions. In practice, it is convenient to use softplus function $f_{\text{SP}}(x) = \log(1 + \exp(x))$ which is a 1-transfer function — commonly used in other neural point processes [41, 46, 72]. Alternatively, $f_+(x) = \max(0, x)$ could be used; however, this is not differentiable at $x = 0$ which can cause issues in practice. Intuitively, M-transfer function are increasing functions which map to all positive values and have bounded steepness.

When a Gaussian process is used to define an inhomogenous Poisson process, the link functions serve a similar role to ensure valid intensity functions [38]. However, many of these link function violate the conditions of being a M-transfer function [17], i.e., the exponential link function $f_+(x) = \exp(x)$ and squared link function $f_+(x) = x^2$ are not M-Lipschitz continuous as they have unbounded derivatives; whereas the sigmoid link function $f_+(x) = \sigma(x)$ is a bounded function (violating condition 2).

Using M-transfer functions, we can show that a uniformly dense class of unbounded functions will be uniformly dense for strictly positive functions under composition. These functions are defined with domain $K \subset \mathbf{R}$, a compact subset, which can be set as $K = [0, T]$ for intensity functions.

Theorem 1. *Given a class of functions $\mathcal{F}$ which is uniformly dense in $C(K, \mathbf{R})$ and a M-transfer function f_+, the composed class of functions $f_+ \circ \mathcal{F}$ is uniformly dense in $C(K, \mathbf{R}_{++})$ for any compact subset $K \subset \mathbf{R}$.*

Proof. Let $f \in C(K, \mathbf{R}_{++})$ and $\varepsilon > 0$ be arbitrary. Since f_+ is strictly increasing and continuous on the preimage of $\mathbf{R}_{++}$ then f_+^{-1} exists, is continuous, and restricted to subdomain $\mathbf{R}_{++}$. Thus, there exists some $g \in C(K, \mathbf{R})$ such that $f = f_+ \circ g$.

As $\mathcal{F}$ is dense with respect to the uniform metric, for ε/M there exists some $h \in \mathcal{F}$ such that $d(h, g) < \varepsilon/M$. Thus for any $x \in K$,

$$\begin{aligned} |(f_+ \circ h)(x) - f(x)| &= |(f_+ \circ h)(x) - (f_+ \circ g)(x)| \\ &\leq M|h(x) - g(x)| < \varepsilon. \end{aligned}$$

We have $d(f_+ \circ h, f) < \varepsilon$. □

To approximate $u_i(\tau)$ using Theorem 1 we need a family of functions which are able to approximate functions in $C(K, \mathbf{R})$. We consider the family of functions consisting of the sum of basis functions $\phi(\cdot;\, p_j)$, where $p_j \in \mathcal{P}$ denotes the parameterisation of the basis function ϕ.

Definition 2. *Denote $\Sigma(\phi)$ as the class of functions corresponding to the sum of basis functions $\phi : \mathbf{R} \times \mathcal{P} \to \mathbf{R}$, with parameter space $\mathcal{P}$, as follows:*

$$\left\{ \hat{u} : \mathbf{R} \to \mathbf{R} \mid \hat{u}(x) = \sum_{j=1}^{J} \phi(x;\, p_j),\ p_j \in \mathcal{P},\ J \in \mathbf{N} \right\}.$$

The parameter space $\mathcal{P}$ of a basis function is determined by the parametric form of a chosen basis function $\phi(x;\, p_j)$. For example, the class composed of exponential basis functions could be defined with parameter space $\mathcal{P} = \mathbf{R}^2$ with functions $\{\phi : \mathbf{R} \to \mathbf{R} \mid \phi(x) = \alpha \exp(\beta x),\ \alpha, \beta \in \mathbf{R}\}$. Definition 2 encompasses a wide range of function classes, including neural networks with sigmoid [11, 28, 14] or rectified linear unit activations [58].

The Stone-Weierstrass Theorem provides sufficient conditions for finding basis function for universal approximation.

Theorem 2 (Stone-Weierstrass Theorem [54, 53])**.** *Suppose a subalgebra $\mathcal{A}$ of $C(K,\, \mathbf{R})$, where $K \subset \mathbf{R}$ is a compact subset, satisfies the following conditions:*

1. *For all $x, y \in K$, there exists some $f \in \mathcal{A}$ such that $f(x) \neq f(y)$;*
2. *For all $x_0 \in K$, there exists $f \in \mathcal{A}$ such that $f(x_0) \neq 0$.*

Then $\mathcal{A}$ is uniformly dense in $C(K,\, \mathbf{R})$.

Thus, by using Theorem 1 and the Stone-Weierstrass theorem, Theorem 2, we arrive at Corollary 1, which gives sufficient conditions for basis functions ϕ to ensure that $f_+ \circ \Sigma(\phi)$ is a universal approximator for $C(K, \mathbf{R}_{++})$.

Corollary 1. *For any compact subset $K \subset \mathbf{R}$ and for any M-transfer function f_+, if a basis function $\phi(\cdot\,;\, p)$ parametrised by $p \in \mathcal{P}$ satisfies the following conditions:*

Basis Function	Functional Form ϕ	Parameter Space $\mathcal{P}$
$\phi_{\text{EXP}}{}^{\dagger}$	$\alpha \exp(\beta x)$	$(\alpha, \beta) \in \mathbf{R}^2$
$\phi_{\text{PL}}{}^{\dagger}$	$\alpha(1+x)^{-\beta}$	$(\alpha, \beta) \in \mathbf{R} \times \mathbf{R}_+$
$\phi_{\text{COS}}{}^{\dagger}$	$\alpha \cos(\beta x + \delta)$	$(\alpha, \beta, \delta) \in \mathbf{R}^3$
$\phi_{\text{SIG}}{}^{\ddagger}$	$\alpha\sigma(\beta x + \delta)$	$(\alpha, \beta, \delta) \in \mathbf{R}^3$
$\phi_{\text{ReLU}}{}^{*}$	$\max(0, \alpha x + \beta)$	$(\alpha, \beta) \in \mathbf{R}^2$

Table 4: Basis function universal approximators for intensity functions between two consecutive events. † indicates functions that satisfy Corollary 1; ‡ one proven in [11]; and $*$ one proven in [58].

1. *$\sum(\phi)$ is closed under product;*
2. *For any distinct points x, $y \in K$, there exists some $p \in \mathcal{P}$ such that $\phi(x;\, p) \neq \phi(y;\, p)$;*
3. *For all $x_0 \in K$, there exists some $p \in \mathcal{P}$ such that $\phi(x_0;\, p) \neq 0$.*

Then $f_+ \circ \sum(\phi)$ is uniformly dense in $C(K, \mathbf{R}_{++})$.

The first condition of Corollary 1 is given such that the set of basis functions $\sum(\phi)$ is a subalgebra of $C(X, \mathbf{R})$. The later two conditions are the required preconditions for the Stone-Weierstrass Theorem to hold.

Given the conditions of Corollary 1, some interesting choices for valid basis functions $\phi(x;\, p)$ are the exponential basis function $\phi_{\text{EXP}}(x) = \alpha \exp(\beta x)$ and the power law basis function $\phi_{\text{PL}}(x) = \alpha(1+x)^{-\beta}$. These basis functions are similar to the exponential and power law Hawkes triggering kernels, which have seen widespread use in many domains [45, 4, 34, 52].

We note that the class of intensity functions in Theorem 1 and Corollary 1 are strictly positive continuous functions. However, these results generalise to non-negative continuous functions as our definition of intensity functions permits arbitrarily low intensity in $u_i(\tau)$ — where switching from arbitrarily low intensities to zero intensity results in arbitrarily low error with respect to the uniform metric on $(0, T]$.

In Table 4, we provide a selection of interesting basis functions to universally approximate $u_i(\tau) \in C(K, \mathbf{R}_{++})$. One should note that Corollary 1 only provides sufficient conditions, where some of the basis function in Table 4 do not satisfy the precondition. For example, the sigmoid basis function $\phi_{\text{SIG}}(x) = \alpha\sigma(\beta x + \delta),\ (\alpha, \beta, \delta) \in \mathbf{R}^3$ does not allow $\Sigma(\phi_{\text{SIG}})$ to be closed under product and thus does not satisfy the conditions of Corollary 1. However, the sum of sigmoid basis functions is equivalent to the class of single hidden layer neural networks [28, 14]. Thus, in additional to an appropriate transfer function it does have the universal approximation property for non-negative continuous functions through Theorem 1. Additionally, other basis functions used to define point process intensity functions can be used, such as radial basis functions [60] that are not generally closed under product but have universal approximation properties [47].

5.2.2 Approximation for Event Sequences

The approximations to $u_i(\tau)$ use a set of parameters, e.g. (α, β, δ) in Table 4. We denote these parameters vectors as $p_i \in \mathcal{P}$, and the approximated function segment as $\hat{u}_i(\tau; p_i)$. Since each

segment $\hat{u}_i(\tau; p_i)$ is uniquely determined by p_i, and the union of all segments approximates $\lambda^\star(t)$, we would only need to capture the dynamics in p_i.

We express p_i as the output of a dynamic system.

$$\begin{aligned} s_{i+1} &= g(s_i, t_i) \\ p_i &= \nu(s_i), \end{aligned} \tag{15}$$

where s_{i+1} is the internal state of the dynamic system, g updates the internal state at each step, and ν maps from the internal state to the output.

Theorem 3 (RNN Universal Approximation [56])**.** *Let $g : \mathbf{R}^J \times \mathbf{R}^I \to \mathbf{R}^J$ be measurable and $\nu : \mathbf{R}^J \to \mathbf{R}^n$ be continuous, the external inputs $x_i \in \mathbf{R}^I$, the inner states $s_i \in \mathbf{R}^J$, and the outputs $p_i \in \mathbf{R}$ (for $i = 1, \ldots, N$). Then, any open dynamic system of the form of Eq. (15) can be approximated by an RNN, with sigmoid activation function, to arbitrary accuracy.*

Given that RNNs approximate p_i, we use continuity condition on basis ϕ and in turn $\hat{u}$ to show how to universally approximate an intensity function with an RNN.

Theorem 4. *Let $\{t_i\}_{i=0}^N$ be a sequence of events with $t_i \in [0, T]$ and $\lambda^\star(t)$ be an intensity function. Given a parametric family of functions $\mathcal{F} = \{\hat{u}(\cdot\,; p) : p \in \mathcal{P}\}$ which is uniformly dense in $C([0, T], \mathbf{R}_{++})$ and $\hat{u}(x; p)$ continuous with respect to p for all $x \in [0, T]$. Then there exists a recurrent neural network*

$$\begin{aligned} h_i &= \sigma(W h_{i-1} + v t_{i-1} + b) & & \\ \hat{p}_i &= A h_i & &\text{for } t \in (t_{i-1}, t_i] \\ \hat{\lambda}(t) &= \hat{u}(\tau; \hat{p}_i) & &\text{and } \tau = t - t_{i-1}, \end{aligned} \tag{16}$$

where σ is a sigmoid activation function and $[W, v, b, A]$ are weights of appropriate shapes, such that $\hat{\lambda}(t)$ approximates $\lambda^\star(t)$ with arbitrary precision for all $(0, T]$.

Proof. Let $\varepsilon > 0$ be arbitrary. For any interval $(t_{i-1}, t_i]$, we know from the uniform density of $\mathcal{F}$ that there exists a p_i such that

$$\sup_{\tau \in [0,T]} |\hat{u}_i(\tau; p_i) - u_i(\tau)| \leq \frac{\varepsilon}{2}. \tag{17}$$

By the continuity conditions of $\hat{u}$, it follows that for each p_i and any $\tau \in [0, T]$ there exists δ_i such that

$$\|p_i - \hat{p}_i\| < \delta_i \implies |\hat{u}(\tau;\, p_i) - \hat{u}(\tau;\, \hat{p}_i)| < \frac{\varepsilon}{2} \tag{18}$$

by taking the minimum over δ_τ's in the $(\varepsilon/2,\, \delta_\tau)$-condition of continuity for all $\tau \in [0, T]$ (where the subscript emphasises the range of τ for fixed i).

The LHS of Eq. (18) is the precision needed in our RNN approximtor for each interval $(t_{i-1}, t_i]$. We take the minimum approximation discrepancy over the sequence of $\hat{p}_i$'s, $\delta := \min_i \delta_i$ and use an RNN with precision δ to bound the approximation quality due to $\hat{p}_i$'s using Theorem 3,

$$\sup_{\tau \in [0,T]} |\hat{u}(\tau;\, p_i) - \hat{u}(\tau;\, \hat{p}_i)| < \frac{\varepsilon}{2}. \tag{19}$$

Using the triangle inequality of the uniform metric, we can combine and bound the discrepancies due to $\hat{u}$ in Eq. (17) and those due to $\hat{p}_i$ in Eq. (19),

$$\sup_{\tau \in [0,T]} |u_i(\tau) - \hat{u}(\tau; \hat{p}_i)| < \varepsilon. \tag{20}$$

Eq. (20) holds for all $i \in \{1, \dots, N\}$. Thus uniform density condition for $\lambda^\star(t)$ also holds for the piece-wise approximator $\hat{\lambda}(t)$ given by Eq. (16) over the entire sequence. □

From Theorem 4 and Corollary 1, universal approximation with respect to the uniform metric follows immediately when using basis functions which are continuous with respect to their parameter space, for example Table 4.

Extensions and discussions. While the original work on learning the compensator function [46] does not provide theoretical backings for its proposal, we note that Theorem 4, combined with universal approximation capabilities of monotone neural networks [**?**], can be used to show that the class of monotonic (increasing) neural networks provide universal approximation for compensator functions. The guarantee described here does not explicitly account for additional dimensions or marks. To extend Theorem 4 in this manner, we consider replacing basis functions $\phi(x)$, which has domain $\mathbf{R}$, to basis functions with extended domain $\mathbf{R} \times K$ where K is a compact set. For example, K can be a set of discrete finite marks in the case of approximated marked temporal point processes. The universal approximation property would then generalise as long as $\sum(\phi)$ is dense in $C([0, T] \times K, \mathbf{R}_{++})$ and continuous in the parameter space of the basis functions. Likewise, if we want to approximate a spatial point process, we let $K = \mathbf{R}^2$ and find an appropriate set of basis functions with domain $\mathbf{R} \times \mathbf{R}^2$.

It is worth mentioning two distinctions from the intensity free approach [**?**]. First, although density approximation allows for direct event time sampling, the log-normal mixture representation assumes that an event will always occur on $\mathbf{R}_+$ — specifically, events cannot naturally stop. Instead, the intensity function representation allow for events to stop with probability $1 - \mathbf{P}(\tau < \infty) = \exp(-\Lambda^\star(\infty))$. In other-words, $1 - \mathbf{P}(\tau < \infty)$ is the probability of events not occurring in finite time, which is non-zero when the intensity function decays and stays at zero. Furthermore, the intensity free approach proposed one functional form (log-normal mixture) for approximating densities, whereas we show that a variety of basis functions all fulfil the goal of universal approximation.

5.3 Describing and predicting online items with reshare cascades

Our proposed model jointly models the unfolding of a heterogeneous set of cascades. When applied to cascades of the same online items, the model directly characterizes their spread dynamics and supplies interpretable quantities, such as content virality and content influence decay, as well as methods for predicting the final content popularities.

5.3.1 Separable Hawkes processes Fitting

In this section, we discuss jointly learning a single set of parameters from a collection of Hawkes realizations.

Parameter estimation. The parameters of a Hawkes process can be estimated by maximizing the likelihood function of a general point process [12]:

$$L(\Theta \mid \mathcal{H}_i(T)) = e^{-\int_0^T \lambda(\tau|\mathcal{H}_i(T))d\tau} \prod_{t_j \in \mathcal{H}_i(T)} \lambda(t_j \mid \mathcal{H}_i(T)) \tag{21}$$

Joint likelihood function. Let $\mathbb{H} = \{\mathcal{H}_1, \mathcal{H}_2, \dots\}$ be a set of independent Hawkes realizations, assumed to be generated from the same model parameterized by n^*, the branching factor, and Θ^g, the parameter set of $g(\cdot)$. It is then straightforward to estimate n^* and Θ^g by maximizing the joint log-likelihood function $\mathcal{L}(n^*, \Theta^g \mid \mathcal{H})$ defined as the sum of the individual log-likelihoods (i.e., the log of eq. (21)):

$$\mathcal{L}(n^*, \Theta^g \mid \mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \log L(n^*, \Theta^g \mid \mathcal{H}_i) \tag{22}$$

After plugging eq. (21) into eq. (22), we see that the joint log-likelihood function can be re-arranged as a sum of two functions with independent parameter sets given $\int_0^\infty g(\tau)d\tau = 1$ and $T \to \infty$:

$$\mathcal{L}(n^*, \Theta^g \mid \mathbb{H}) = \mathcal{L}_g(\Theta^g \mid \mathbb{H}) + \mathcal{L}_n(n^* \mid \mathbb{H}) \tag{23}$$

with $\mathcal{L}_g$ a function of Θ^g and $\mathcal{L}_n$ a function of n^*:

$$\mathcal{L}_g(\Theta^g \mid \mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \sum_{t_j \in \mathcal{H}_i, j \geq 1} \log \sum_{t_z < t_j} g(t_j - t_z \mid \Theta^g) \tag{24}$$

$$\mathcal{L}_n(n^* \mid \mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \log \left[(n^*)^{N_i - 1} e^{-N_i n^*} \right] \tag{25}$$

Regarding the assumption $T \to \infty$, most cascades are complete in practice given a large T. We also note that eq. (25) can be solved efficiently and analytically by setting its first derivative to 0.

The above results indicate that Θ^g and n^* can be learned independently in two separate phases, by maximizing $\mathcal{L}_g$ and $\mathcal{L}_n$. This amounts to fitting n^* from observed final cascade sizes only, and Θ^g from inter-arrival times between events.

We note that maximizing $\mathcal{L}_n$ is equivalent to the maximum likelihood estimation of the Borel distribution. One can see this by expanding both forms, as shown below:

$$\begin{aligned}
&\arg\max_{n^*} \sum_{\mathcal{H}_i \in \mathbb{H}} \log \mathbb{B}(N_i \mid n^*) \\
&= \arg\max_{n^*} \sum_{\mathcal{H}_i \in \mathbb{H}} \left[\log(n^*)^{N_i - 1} e^{-N_i n^*} + \log \frac{N_i^{N_i - 1}}{N_i!} \right] \\
&\overset{(a)}{=} \arg\max_{n^*} \mathcal{L}_n(n^* \mid \mathbb{H})
\end{aligned} \tag{26}$$

where we discard the log ratio of constants N_i at step (a).

To the best of our knowledge, this is the first work to discuss the separable form of Hawkes parameter estimations and its connection to the Borel distribution.

5.3.2 Dual Mixture Model

In practice, an online item is reshared across a set of diffusion cascades of diverse dynamics. In this section, we propose a dual mixture model that allows individual cascades to differ one from another. Given the separability of the log-likelihood functions (eqs. (24) and (25)), we introduce a Borel mixture model (BMM) and a kernel mixture model (KMM) to automatically uncover the latent clusters of models based on cascade sizes and time intervals. Finally, we employ the fitted dual mixture model to construct item level characterizations, such as $\hat{n}_v^*$, $\hat{\theta}_v$ and the diffusion embeddings with a distance measure.

Mixture models for Hawkes processes. We are given $\mathbb{H}_v$, a set of cascades relating to an online item v, and the number of components k_v — there exist k_v latent generative models with unknown relations to the cascades in $\mathbb{H}_v$. We seek to learn k_v groups of n^* and Θ^g, and their weights. As indicated in section 5.3.1, we model these two parameter sets separately using cascade sizes and inter-arrival times. We denote the obtained model as $M_v = \{M_v^B, M_v^K\}$ where $M_v^B = \{(n_1^*, p_1^B), \ldots, (n_{k_v}^*, p_{k_v}^B)\}$, $M_v^g = \{(\Theta_1^g, p_1^g), \ldots, (\Theta_{k_v}^g, p_{k_v}^g)$. $p_1^B, \ldots, p_{k_v}^B$ and $p_1^g, \ldots, p_{k_v}^g$ are the component weights for corresponding Borel models and kernel functions.

Given two mixture models, M_v^B and M_v^K, inferred separately from a group of cascades, we assume the intensity functions of the corresponding Hawkes processes are parameterized by the cartesian product of M_v^B and M_v^K, i.e.,

$$M_v^H = \{(n_i^*, \Theta_j^g, p_i^B p_j^g) \mid (n_i^*, p_i^B) \in M_v^B \text{ and } (\Theta_j^g, p_j^g) \in M_v^g\} \tag{27}$$

where $p_i^B p_j^g$ gives the component weight.

Borel mixture model (BMM). To learn the M_v^B for the online item v, we present an EM estimation algorithm [15]. A BMM can be fitted on $\mathbb{H}_v$ by maximizing the log-likelihood

$$\mathcal{L}_{BMM} = \sum_{\mathcal{H}_{v,i} \in \mathbb{H}_v} \log \sum_{k=1}^{k_v} \underbrace{p_k^B \mathbb{B}(N_{v,i} \mid n_k^*)}_{q^B(k, N_{v,i})} \tag{28}$$

As maximizing eq. (28) directly suffers from the identifiability issue [5], we apply the Expectation-Maximization (EM) algorithm commonly used for learning mixture models [61]. This algorithm optimizes an alternative lower bound Q_{BMM} defined as

$$Q_{BMM} = \sum_{\mathcal{H}_{v,i} \in \mathbb{H}_v} \sum_{k=1}^{k_v} p^B(k \mid N_{v,i}) \log q^B(k, N_{v,i}) \tag{29}$$

where $p^B(k \mid N_{v,i})$ is the probability of N_i being a member of the kth model and is updated during the E step. Next we give the update formulas for the E and M steps.

E-step: membership probabilities are updated

$$p^B(k \mid N_{v,i}) = \frac{q^B(k, N_{v,i})}{\sum_{j=1}^{k_v} q^B(j, N_{v,i})} \tag{30}$$

M-step: n_k^* and p_k^B are updated analytically

$$(n_k^*)^{new} = \frac{\sum_{N_{v,i}} p^B(k \mid N_{v,i})(N_{v,i} - 1)}{\sum_{N_{v,i}} p^B(k \mid N_{v,i}) N_{v,i}} \tag{31}$$

$$(p_k^B)^{new} = \sum_{N_{v,i}} \frac{p^B(k \mid N_{v,i})}{|\mathbb{H}_v|} \tag{32}$$

Parameters are updated iteratively by alternating these two steps until the convergence of $\mathcal{L}_{BMM}$.

Kernel mixture model (KMM). As we follow similar derivations for obtaining M_v^K, we note only two differences regarding the definition of $\mathcal{L}_{KMM}$ and the update of Θ_k^g

$$\mathcal{L}_{KMM} = \sum_{\mathcal{H}_{v,i} \in \mathbb{H}_v} \log \sum_{k=1}^{k_v} p_k^g f^g(\mathcal{H}_{v,i} \mid \Theta_k^g) \tag{33}$$

$$(\Theta_k^g)^{new} = \arg\max_{\Theta^g} \sum_{\mathcal{H}_{v,i} \in \mathbb{H}_v} p^g(k \mid \mathcal{H}_{v,i}) \log f^g(\mathcal{H}_{v,i} \mid \Theta^g)$$

where $f^g(\mathcal{H}_{v,i} \mid \Theta^g) = \prod_{t_j \in \mathcal{H}_{v,i}} \sum_{t_z < t_j} g(t_j - t_z \mid \Theta^g)$. The way $(\Theta_k^g)^{new}$ is solved depends on specific kernel functions. In our experiments, we solve this with a non-linear solver, Ipopt [62], where a power-law kernel function is employed.

eqs. (28) and (33) have respectively linear and quadratic computational complexity, however the EM algorithm allows an efficient implementation of the dual mixture model.

Determining the number of components. Prior literature uses a number of information criteria for choosing a component number of mixture models [8, 39], including the Akaike information criteria (AIC). In our experiments, we employ AIC defined as $2k_v - 2\mathcal{L}_{BMM}$ to select k_v with BMMs. Note that fitting BMM is computationally efficient — due to the analytical updates of the EM algorithm — which allows one to experiment various values for k_v. In our experiments, the numbers of components k_v given by AIC are generally between 2 and 5.

Characterizing items using the dual mixture model. We build item-level quantifications based on the dual mixture model fitted on all cascades relating to the given item. The diffusion embedding provides a fixed length vector describing the information in the components of BMM and KMM, while the content virality and influence decay provide single value summarizations of the two mixtures.

A diffusion embedding constructed from the fitted mixture models M_v is a vector of mixture component weights. Taking the power-law kernel function as an example, we build a diffusion embedding in two steps:

- Parameter discretization: we first discretize the continuous model parameters n^*, θ and c by separating them into fixed number of quantile bins. Given BMMs learned from all observed online items V, we obtain the value of the ith quantile $q_i^{n^*}$ from the weighted samples $\{(n_j^*, p_j^B) \mid j \in \{1, \ldots, k_v\}, \forall v \in V\}$. We use the algorithm provided in [26] to compute weighted quantiles. Similarity, we get q_i^c, q_i^θ from the fitted KMMs.

- Weight aggregation: we then convert M_v^B into a vector of weights for an online item v, $\boldsymbol{m}_v^{n^*} = [m_{v,1}^{n^*}, \ldots]^\mathsf{T}$ where each element is the sum of weights $m_{v,i}^B = \sum_{q_{i-1}^{n^*} < n_j^* \leq q_i^{n^*}} p_j^B$. Moreover, M_v^g can be encoded as $\boldsymbol{m}_v^c = [m_{v,1}^c, \ldots]^\mathsf{T}$ and $\boldsymbol{m}_v^\theta = [m_{v,1}^\theta, \ldots]^\mathsf{T}$.

In the end, three vectors $(\boldsymbol{m}_v^{n^*}, \boldsymbol{m}_v^c, \boldsymbol{m}_v^\theta)$ are provided for each online item as the diffusion embeddings and can be used with off-the-shelf supervised or unsupervised tools.

We also compute the single value summarizations as: $\hat{n}_v^* = \sum_{k=1}^{k_v} n_k^* p_k^B$, $\hat{c}_v = \sum_{k=1}^{k_v} c_k p_k^g$, $\hat{\theta}_v = \sum_{k=1}^{k_v} \theta_k p_k^g$. We denote $\hat{n}_v^*$ as content virality, and $\hat{\theta}_v$ as influence decay. These are two values of interest showing how viral and how long the influence of an online item stay in online discussions.

Distance between diffusion embeddings. Given two items described by their respective diffusion embeddings $(\boldsymbol{m}_1^{n^*}, \boldsymbol{m}_1^c, \boldsymbol{m}_1^\theta)$ and $(\boldsymbol{m}_2^{n^*}, \boldsymbol{m}_2^c, \boldsymbol{m}_2^\theta)$, we seek to measure their distance $D_{1,2}$. We note that the position of elements in the embeddings represents quantiles at an increasing order, but common distance measures, such as the Euclidean distance and the cosine distance, ignore such information. For example, given $\boldsymbol{m}_1^{n^*} = [1, 0, 0, \cdots]$, $\boldsymbol{m}_2^{n^*} = [0, 1, 0, \cdots]$ and $\boldsymbol{m}_3^{n^*} = [0, 0, 1, \cdots]$, $\boldsymbol{m}_1^{n^*}$ is intuitively closer to $\boldsymbol{m}_2^{n^*}$ than to $\boldsymbol{m}_3^{n^*}$ instead of equally close. To address this, we employ the Wasserstein distance [3] which accounts for positional information. The Wasserstein distance of order 1 for single dimensional histogram has a closed-form solution defined as $W_1(\boldsymbol{M}_1^{n^*}, \boldsymbol{M}_2^{n^*}) = \sum_i |M_{1,i}^{n^*} - M_{2,i}^{n^*}|$, where $\boldsymbol{M}_\cdot^{n^*} = [\sum_{j=1}^1 \boldsymbol{m}_{\cdot,j}^{n^*}, \sum_{j=1}^2 \boldsymbol{m}_{\cdot,j}^{n^*}, \sum_{j=1}^3 \boldsymbol{m}_{\cdot,j}^{n^*}, \cdots]$ represents the cumulative weights at increasing quantiles. We then define the distance of the pair of diffusion embeddings as

$$D_{1,2} = W_1(\boldsymbol{M}_1^{n^*}, \boldsymbol{M}_2^{n^*}) + W_1(\boldsymbol{M}_1^c, \boldsymbol{M}_2^c) + W_1(\boldsymbol{M}_1^\theta, \boldsymbol{M}_2^\theta) \tag{34}$$

5.3.3 Predicting the future of cascades

In this section, we show how fitted mixture models can be applied to future observations. We describe the evaluation of generalization performance on holdout parts of unseen cascades. Next, we derive predictions of final popularities.

Models for future content. We build mixture models for a newly published item by combining historical fitted models of items V_ρ from the same publisher ρ, i.e.,

$$M_\rho^B = \bigcup_{v \in V_\rho} \{(n_i^*, p_i^B / |V_\rho|), \cdots\}, \quad \forall (n_i^*, p_i^B) \in M_v^B \tag{35}$$

$$M_\rho^g = \bigcup_{v \in V_\rho} \{(\Theta_i^g, p_i^g / |V_\rho|), \cdots\}, \quad \forall (\Theta_i^g, p_i^g) \in M_v^B \tag{36}$$

and $M_\rho = \{M_\rho^B, M_\rho^g\}$, assuming the new item follows the dynamics of its predecessors. Following eq. (27), we obtain M_ρ^H from M_ρ. In our experiments, we limit V_ρ to the most recent published items.

Cascade holdout log-likelihood. When fitting a Hawkes process on a cascade $\mathcal{H}_i(T)$ until an observation time T, the log-likelihood value of the holdout part of this cascade, i.e., $HLL = \mathcal{L}(\Theta \mid \mathcal{H}_i) - \mathcal{L}(\Theta \mid \mathcal{H}_i(T))$, evaluates the model generalization performance to unseen events. For our proposed dual mixture model, we compute an expected holdout log-likelihood stemming from the posterior model probabilities given $\mathcal{H}_i(T)$, i.e.,

$$\mathbb{E}[HLL] = \sum_{(n_k^*, \Theta_j^g, p_k^B p_j^g) \in M_\rho^H} [\mathcal{L}(\Theta \mid \mathcal{H}_i) - \mathcal{L}(\Theta \mid \mathcal{H}_i(T))] \times \mathbb{P}[n_k^*, \Theta_j^g \mid \mathcal{H}_i(T)] \tag{37}$$

where we have: $\mathbb{P}[n_k^*, \Theta_j^g \mid \mathcal{H}_i(T)] = \frac{\mathbb{P}[\mathcal{H}_i(T) | n_k^*, \Theta_j^g] p_k^B p_j^g}{\sum_{M_\rho^H} \mathbb{P}[\mathcal{H}_i(T) | n^*, \Theta^g] p^B p^g}$

Table 5: Statistics of the two social media datasets.

	Start time	End time	#categories	#publishers	#items	#cascades	#tweets
ActiveRT2017-Fit	Jan 1, 2017	May 1, 2017	18 (*Music*, *Gaming*, ...)	11,297 channels	75,717 videos	30,535,891	85,334,424
ActiveRT2017-Test	Jun 1, 2017	Dec 31, 2017					
RNCNIX-Fit	June 30, 2017	Jan 1, 2019	2 (*RNIX*, *CNIX*)	73 domains	102,429 articles	8,129,126	56,397,252
RNCNIX-Test	Feb 1, 2019	Dec 31, 2019					

Cascade posterior size distribution. Given a pair of parameters n^* and Θ^g, we are able to derive the posterior size distribution given $\mathcal{H}_i(T)$ of a cascade i. The future events after time T are of two kinds: direct offspring of observed events (their count denoted as N_i^d) and indirect offspring (children of children, total count denoted as N_i^{ind}). The process generating direct offspring is an inhomogeneous Poisson process of conditional intensity $\lambda(t|\,\mathcal{H}_i(T)), t > T$ — note that this is not a stochastic function as only the history up to time T is accounted in the intensity function. Consequently, N_i^d follows a Poisson distribution of the intensity $\Lambda_i(T \mid n^*, \Theta^g) = \int_T^\infty \lambda(\tau|\,\mathcal{H}_i(T), n^*, \Theta^g)d\tau$.

Furthermore, each direct offspring initiated a Hawkes process and its total progeny number follows a Borel distribution. Given the number of direct offspring N_i^d, the total number of direct and indirect offspring follows a Borel-Tanner distribution (also known as the generalized Borel distribution) [25]: $\mathbb{B}(\kappa \mid n^*, N_i^d) = \frac{N_i^d(\kappa n^*)^{\kappa - N_i^d} e^{-\kappa n^*}}{\kappa(\kappa - N_i^d)!}$ for $\kappa = N_i^d, N_i^d + 1, \cdots$. Its mean, $\frac{N_i^d}{1-n^*}$, and variance, $\frac{N_i^d n^*}{(1-n^*)^3}$, are similar to those of a Borel distribution.

Finally, the posterior cascade size distribution is therefore

$$\begin{aligned}\mathbb{P}[N_i = n \mid \mathcal{H}_i(T)] &= N_i(T) \\ &+ \sum_{z=0}^{n-N_i(T)} Poi(z \mid \Lambda_i(T \mid n^*, \Theta^g))\mathbb{B}(n - N_i(T) \mid n^*, z)\end{aligned} \tag{38}$$

where $Poi(\cdot|\lambda)$ is the Poisson distribution given intensity λ. eq. (38) leads to a quadratic complexity in computing the final size distribution, which is intractable in most real-life scenarios. A numerical trick can be applied to reduce the complexity by introducing a threshold probability ϵ_p and summing until $Poi(z \mid \Lambda_i(T \mid n^*, \Theta^g)) < \epsilon_p$.

Online item popularity prediction. The final popularity of an online item consists of two parts in prediction: the final popularities of current observed cascades and new cascades created in future.

We first use past average cascade counts of the publisher ρ as an estimation of the new cascades that will emerge in future, denoted as $\hat{C}_\rho$. The final popularity of these is thus the mean of a Borel-Tanner distribution given $\hat{C}_\rho$ initial events, i.e.,$\frac{\hat{C}_\rho}{1-n^*}$. We then compute the mean values from a posterior distribution as the predicted final popularity $\hat{N}_{v,i}$ of the observed cascade i given n^* and Θ^g, i.e.,

$$
\begin{aligned}
&\hat{N}_{v,i}(n^*, \Theta^g) \\
&\quad = N_{v,i}(T) + \sum_{\kappa=0}^{\infty} \sum_{z=0}^{\kappa} \kappa \cdot Poi(z \mid \Lambda_i(T \mid n^*, \Theta^g)) \mathbb{B}(\kappa \mid n^*, z) \\
&\quad \overset{(a)}{=} N_{v,i}(T) + \sum_{z=0}^{\infty} Poi(z \mid \Lambda_i(T \mid n^*, \Theta^g)) \sum_{\kappa=z}^{\infty} \kappa \cdot \mathbb{B}(\kappa \mid n^*, z) \\
&\quad \overset{(b)}{=} N_{v,i}(T) + \frac{\sum_{z=0}^{\infty} z \cdot Poi(z \mid \Lambda_i(T \mid n^*, \Theta^g))}{1 - n^*} \\
&\quad \overset{(c)}{=} N_{v,i}(T) + \frac{\Lambda_i(T \mid n^*, \Theta^g)}{1 - n^*}
\end{aligned} \tag{39}
$$

where step (a) exchanges the order of two summations. Step (b) and step (c) follow the means of a Borel-Tanner distribution [25] and a Poisson distribution. Last, we add predictions of all cascades and future cascades relating to a new online item and take expectation over possible parameter sets from the mixture models

$$
\hat{N}_v = \mathbb{E}_{M_\rho^H} \left[\frac{\hat{C}_\rho}{1 - n^*} + \sum_{\mathcal{H}_{v,i}(T) \in \mathbb{H}_v(T)} \hat{N}_{v,i}(n^*, \Theta^g) \right] \tag{40}
$$

As the variance of Borel-Tanner distribution is also known [25], eq. (38) enables us to derive the variance of final popularities.

6 Selected results

In this section we present a selection of the results that we obtained during this project.

6.1 Results: Aggregate attention on YouTube and Twitter

We report the results as violin plots in this section. The outlines are kernel density estimates for the left-leaning (blue) and right-leaning (red) videos, respectively. The center dashed line is the median, whereas the two outer lines denote the interquartile range. To compare the distributions of each metric for the left- and right-leaning videos, we adopt the one-sided Mann–Whitney U test.
Total view count. Figure 8(a) shows the distribution of video views at day 120 after upload. Using the view count at the same day removes the effects of video age, so that the videos published for longer time are not taking an unfair advantage. In `Abortion` and `Gun Control`, the median, as well as 25^{th} and 75^{th} percentile of views of left-leaning videos are higher than that of right-leaning videos. The median views for left-leaning videos are 107,346 for `Abortion` and 153,482 for `Gun Control`, versus 62,780 and 103,373 for right-leaning ones. The differences in view distribution are statistically significant. For `BLM`, right-leaning videos have higher median and 75^{th} of views, but the effect is not significant.
Relative engagement. From Figure 8(b), we can see that videos in all three topics are highly engaging, with mean relative engagement at 0.834 for `Abortion`, 0.824 for `Gun Control`,

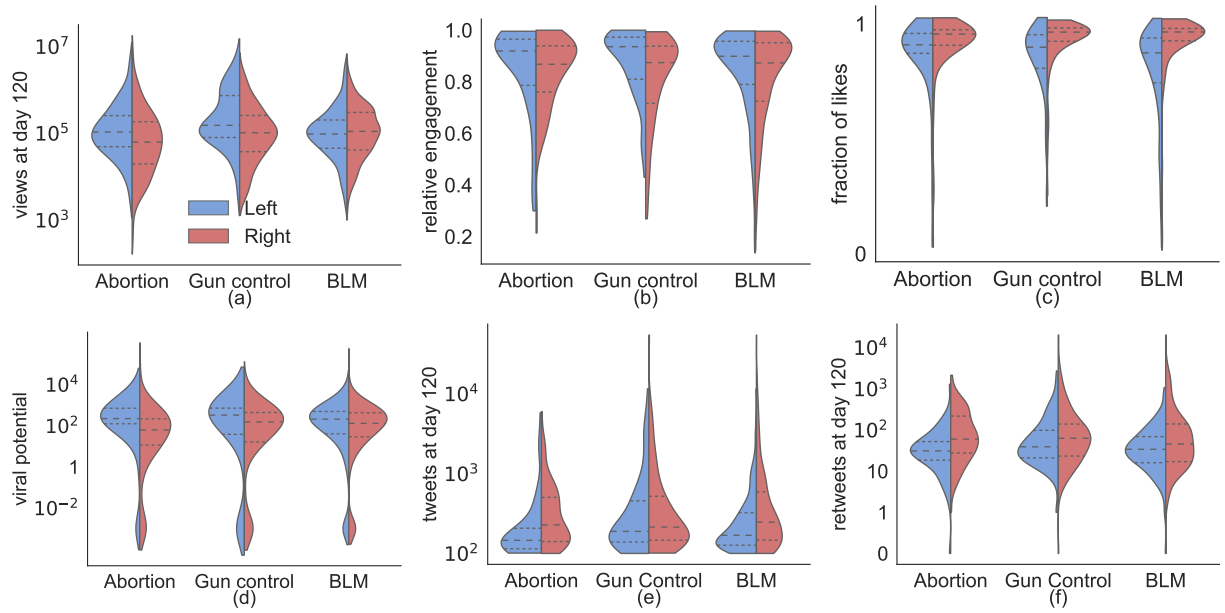


Figure 8: Violin plots comparing the left-leaning (blue) and right-leaning (red) videos in (a) total number of views at 120 days after upload, (b) relative engagement, (c) fraction of likes, (d) viral potential, (e) total number of tweets, and (f) total number of retweets at 120 days after the video upload. Left-leaning videos are more viewed, more engaging, having more uniform reactions, and more viral than right-leaning videos across all three topics (except views in `BLM`). In `Abortion` and `BLM`, right-leaning videos are significantly more tweeted, especially with more retweets.

and 0.831 for `BLM`. This is because our data processing procedure requires videos to have at least 100 tweets and 100 views, which tends to select videos with significant amount of interests. Left-leaning videos are significantly more engaging than their right-leaning counterparts across all three topics.

Fraction of likes. Figure 8(c) presents the proportion of likes in videos' reactions. We find that right-leaning videos across all topics have significantly larger fraction of likes than left-leaning videos.

Viral potential. Figure 8(d) shows the distributions of viral potential. We find that the left-leaning videos have significantly higher viral scores than the right-leaning videos across all three topics, meaning that given the same amount of tweets exposing the video on Twitter, an average left-leaning video can effectively attracts more views than an average right-leaning video. The difference is most notable in `Abortion`: a typical left-leaning video receives 224 views from an average tweet, whereas a typical right-leaning video receives only 63 views.

Tweet counts. Figure 8(e) and (f) shows the distributions of total tweets and retweets for each political leaning. Contrasting to the observation that left-leaning videos are more viewed, here we find that right-leaning videos are significantly more tweeted, especially with more retweets and more replies in `Abortion` and `BLM`. On the other hand, we do not observe a significant difference in original tweets and quotes, except for `BLM` where right-leaning videos have prevailing volume across all tweet types.

from \ to	left video	right video
liberal	15.73%	**16.31%**
conservative	**15.77%**	14.44%

Table 6: Percentage of root comments that are toxic. Bolded values are posts on opposite ideology videos.

from \ to	lib. on left video	cons. on left video	lib. on right video	cons. on right video
liberal	12.12%	**18.24%**	13.42%	**15.31%**
conservative	**15.24%**	11.11%	**17.15%**	10.18%

Table 7: Percentage of replies that are toxic. Bolded values are replies between two users of opposite ideologies.

6.2 Results: Toxicity as a measure of quality

We obtained the toxicity scores for YouTube comments by querying the Perspective API [30]. The returned score ranges from 0 to 1, where scores closer to 1 indicate that a higher fraction of raters will label the comment as toxic. We used 0.7 as a threshold as suggested in prior work [29]. For each cell of Table 6&7, we sampled 100K random comments that satisfy corresponding definitions, and then counted the fraction of comments deemed toxic. Margins of error were less than 0.24% with the sample size of 100K. The data subsampling was to avoid making excessive API requests.

Table 6 reports the frequency of toxic root comments. We find that liberals and conservatives' root comments had about the same toxicity when posting on left-leaning videos. However, conservatives posted fewer toxic root comments on right-leaning videos, and thus slightly fewer toxic root comments overall.

Table 7 reports the frequency of toxic replies. We find that replies to people of opposite ideology were much more frequently toxic, for both liberals and conservatives. There also appears to be a "defense of home territory" phenomenon. Conservatives were significantly more toxic in their replies to liberals on right-leaning videos (17.15%) than on left-leaning videos (15.24%) and analogously for liberals responding to conservatives (18.24% on left-leaning vs. 15.31% on right-leaning videos). Commenting on an opposing video generates more hostile responses than commenting on a same-ideology video. Interestingly, this holds true even for replies from people who share their ideology. For example, liberals received more toxic replies from liberals on right-leaning videos (13.42% toxic) than they did on left-leaning videos (12.12% toxic).

6.3 Results: Data sampling impacts on Twitter networks

We measure the effects of data sampling on two commonly studied networks on Twitter: the user-hashtag bipartite graph, and the user-user retweet network.

	complete	sample	ratio
#tweets with hashtags	24,539,003	13,149,980	53.59%
#users with hashtags	6,964,076	4,758,161	68.32%
avg. hashtags per user	9.23	7.29	78.97%
#hashtags	1,166,483	880,096	75.45%
avg. users per hashtags	55.09	39.40	71.51%

Table 8: Statistics of user-hashtag bipartite graph in CYBERBULLYING dataset. Ratio (rightmost column) compares the value of the sample set against that of the complete set, mean sampling rate $\bar{\rho}$=52.72%.

6.3.1 User-hashtag bipartite graph

The bipartite graph maps the affiliation between two disjoint sets of entities. No two entities within the same set are linked. Bipartite graphs have been used in many social applications, e.g., mining the relation between scholars and published papers [44], or between artists and concert venues [2]. Here we construct the user-hashtag bipartite graphs for both the complete and the sample sets. This graph links users to their used hashtags. Each edge has a weight – the number of tweets between its associated user and hashtag. The basic statistics for the bipartite graphs are summarized in Table 8.

Clustering techniques are often used to detect communities in such bipartite graphs. We apply spectral clustering [59] on the user-hashtag bipartite graph, with the number of clusters set at 6. The resulted clusters are summarized in Table 9, together with the most used 5 hashtags and a manually-assigned category. Apart from the cyberbullying keywords, there are significant amount of hashtags related to politics, live streaming, and Korean pop culture, which are considered as some of the most discussed topics on Twitter. We further quantify how the clusters traverse from the complete set to the sample set in Figure 9. Three of the complete clusters (CC1, CC2, and CC3) are maintained in the sample set (mapping to SC1, SC2, and SC3 respectively), since more than half of the entities preserve. The remaining three complete clusters disperse. Investigating the statistics for the complete clusters, the preserved ones have a larger average weighted degree, meaning more tweets between the users and hashtags in these clusters. Another notable observation is that albeit the entities traverse to the sample clusters differently, all complete clusters have similar missing rates (28% to 34%). It suggests that Twitter data sampling impacts the community structure. Denser structures are more resilient to sampling.

6.3.2 User-user retweet network

Retweet network describes the information sharing between users. We build a user-user retweet network by following the "@RT" relation.. Each node is a user, and each edge is a directed link weighted by the number of retweets between two users. The user-user retweet network has been extensively investigated in literature [55, 43, 22].

We choose to characterize the retweet network using the bow-tie structure. Initially proposed to measure the World Wide Web [7], the bow-tie structure was also used to measure the QA community [68] or YouTube video networks [65]. The bow-tie structure characterizes a network into 6 components: (a) the largest strongly connected component (LSCC) as the central part; (b) the IN component contains nodes pointing to LSCC but not reachable from LSCC; (c) the OUT component

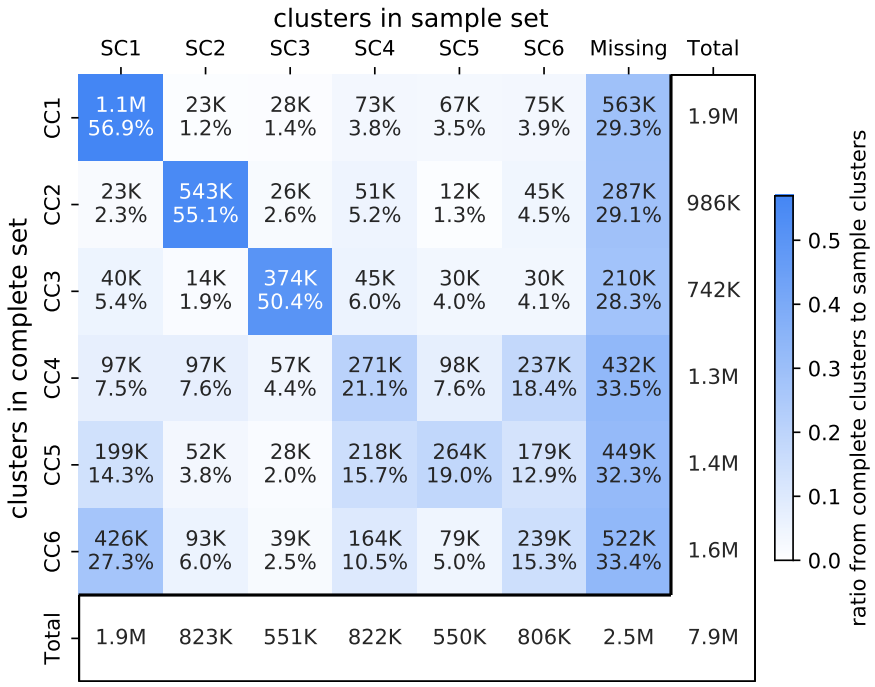


Figure 9: The change of clusters from complete set to sample set. Each cell denotes the volume (top number) and the ratio (bottom percentage) of entities (users and hashtags) that traverse from a complete cluster to a sample cluster. Clusters are ordered to achieve maximal ratios along the diagonal.

		CC1	CC2	CC3	CC4	CC5	CC6
complete set	size	**1,925,520**	**986,262**	**742,263**	1,289,086	1,389,829	1,562,503
	#users	**1,606,450**	**939,288**	**602,845**	1,080,359	1,227,127	1,390,276
	#hashtags	**319,070**	**46,974**	**139,418**	208,727	162,702	172,227
	avg. degree	**8.03**	**7.64**	**22.19**	3.46	4.74	4.07
	category	**politics**	**Korean pop**	**cyberbullying**	Southeast Asia pop	politics	streaming
	hashtags	**brexit**	**bts**	**gay**	peckpalitchoke(th)	kamleshtiwari	ps4live
		demdebate	**mamavote**	**pussy**	peckpalitchoke	standwithhongkong	bigolive
		afd	**blackpink**	**sex**	vixx	hongkong	10tv
		cdnpoli	**pcas**	**horny**	wemadeit	bigil	mixch.tv(ja)
		elxn43	**exo**	**porn**	mayward	lebanon	twitch
		SC1	**SC2**	**SC3**	SC4	SC5	SC6
sample set	size	**1,880,247**	**823,232**	**551,219**	822,436	549,589	805,852
	#users	**1,600,579**	**767,183**	**446,303**	686,609	465,339	688,922
	#hashtags	**279,668**	**56,049**	**104,916**	135,827	84,250	116,930
	avg. degree	**5.58**	**5.75**	**14.98**	3.06	3.51	3.28
	category	**politics**	**Korean pop**	**cyberbullying**	mixed	mixed	mixed
	hashtags	**ps4live**	**bts**	**gay**	mixch.tv(ja)	bigolive	Idolish7(ja)
		10tv	**mamavote**	**pussy**	bigil	kamleshtiwari	reunion
		brexit	**blackpink**	**sex**	peckpalitchoke(th)	bb13	Idolish7(ja)
		afd	**pcas**	**horny**	reality_about_islam(hi)	biggboss13	vixx
		demdebate	**bts(ko)**	**porn**	doki.live(ja)	execution_rajeh_mahmoud(ar)	vixx(ko)

Table 9: Statistics and the most used 5 hashtags in the 6 clusters of the user-hashtag bipartite graph. Three complete clusters maintain their structure in the sample set (**boldfaced**). The language code within brackets is the original language for the hashtag. ja: Japanese; ko: Korean; th: Thai; hi: Hindi; ar: Arabic.

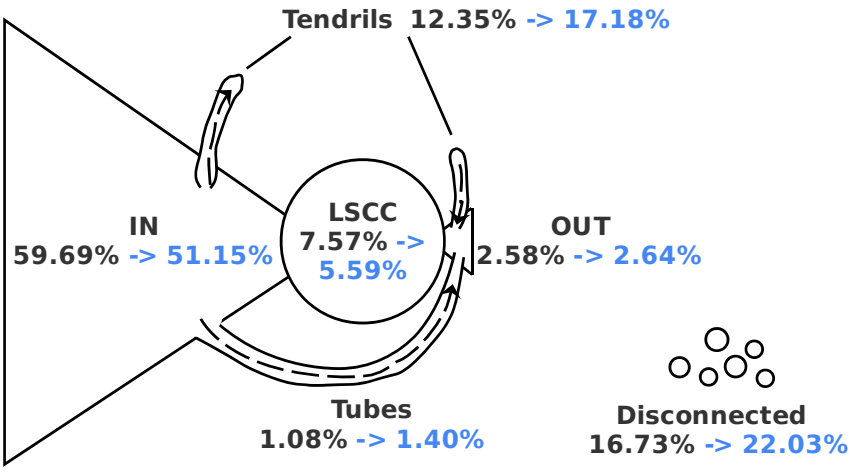


Figure 10: Visualization of bow-tie structure in complete set. The black number indicates the relative size of component in the complete set, blue number indicates the relative size in the sample set.

contains nodes that can be reached by LSCC but not pointing back to LSCC; (d) the Tubes component connects the IN and OUT components; (e) the Tendrils component contains nodes pointing from In component or pointing to OUT component; (f) the Disconnected component includes nodes not in the above 5 components. Figure 10 visualizes the bow-tie structure of the user-user retweet network, alongside with the relative size for each component in the complete and sample sets. The LSCC and IN components, which make up the majority part of the bow-tie, reduce the most in both absolute size and relative ratio due to sampling. OUT and Tubes are relatively small in both complete and sample sets. Tendrils and disconnected components enlarge 39% and 32% after sampling.

Figure 11 shows the node flow of each components from the complete set to the sample set. About a quarter of LSCC component shift to the IN component. For the OUT, Tubes, Tendrils, and Disconnected components, 20% to 31% nodes move into the Tendrils component, resulting in a slight increase of absolute size for Tendrils. Most notably, nodes in the LSCC has a much smaller chance of missing (2.2%, other components are with 19% to 38% missing rates).

6.4 Results: Modeling diffusions on Twitter

Generalization to unseen data. On each of the three dataset, we fit the parameters of all six models. We follow the experimental setup in [50]: 40% of the tweets in each cascade are used to fit model parameters, and we report the negative log-likelihood on the remaining 60% of the events normalized by the event count.

Fig. 12a and b show as boxplots the generalization performance on *NEWS*, without and with marks respectively. Two conclusions emerge. First, the power-law kernel for both Hawkes and HawkesN consistently outperforms other kernel functions. This emphasizes the importance of developing the generalized SIR model, as different types of parametric kernel function might fit different types of data better. Second, HawkesN outperforms Hawkes confirming results reported in [50].

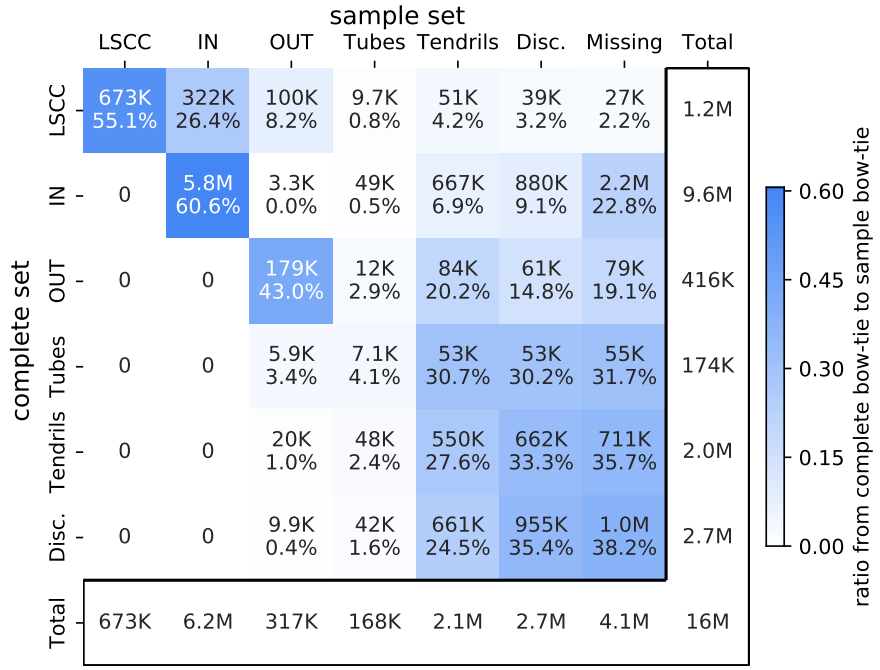


Figure 11: The change of bow-tie components from complete set to sample set. Each cell denotes the volume (top) and the ratio (bottom) of users that traverse from a component in complete set to a component in sample set.

Popularity prediction. We predict final retweet cascade popularity following the setup described in [42]. We observe each cascade for one hour and we fit model parameters; we predict final diffusion sizes (popularity) and test against the observed final cascade size. We measure performance using the Absolute Relative Error (ARE):

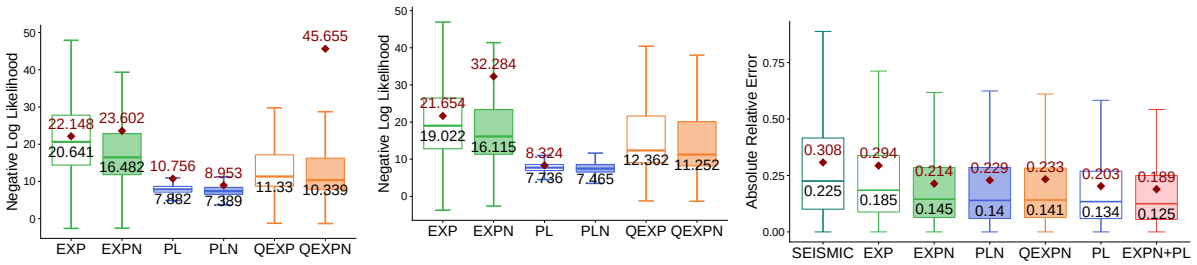
$$ARE = \frac{|\hat{N}_\infty - N_\infty|}{N_\infty}$$

where $\hat{N}_\infty$ and N_∞ are the predicted size and the true size, respectively. We compare HawkesN models to the Hawkes models (EXP, PL), and to SEISMIC [70]. Furthermore, we adopt the observation that EXPN and PL are two complementary models on *NEWS* to introduce a combined model by averaging EXPN and PL prediction outcomes [71]. Results are reported with 10-fold cross validation where 6 folds are used for testing after trained on 4 folds during each iteration.

fig. 12c shows the prediction performances on the *NEWS*. Among all the Hawkes and HawkesN models, PL delivers the best prediction performance, and EXPN predicts better than EXP. Overall, the combined model, EXPN+PL, consistently outperforms all other models, on all datasets. It provides a choice to deal with complementary modeling power of kernel functions on different cascades. This only reinforces the conclusion that there may exist more than one cascade dynamics, and that each model captures the best one of them.

6.5 Results: Evaluation of UNIPoint models

Table 10 reports log-likelihoods of all models across the three synthetic and three real world datasets.



(a) Holdout log-likelihood

(b) Holdout log-likelihood with marks

(c) Popularity prediction performance

Figure 12: Fig. *(a)-(b)* depict holdout negative log-likelihood per event of six models on *NEWS*, with and without additional mark information. Fig. *(c)* shows diffusion final popularity prediction performance on *NEWS*. The red diamond shows the mean value in each boxplot — lower is better.

Dataset		Synthetic			Real World		
Models		SelfCorrecting	ExpHawkes	DecayingSine	MOOC	Reddit	StackOverflow
Baseline	ExpHawkes	$-0.994 \pm .001$	$0.044 \pm .037$	$-0.838 \pm .019$	$3.578 \pm .060$	$-0.100 \pm .039$	$-1.031 \pm .025$
	PLHawkes	$-0.994 \pm .001$	$0.036 \pm .037$	$-0.845 \pm .019$	$0.532 \pm .070$	$-0.787 \pm .035$	$-0.918 \pm .024$
	RMTPP	$-0.776 \pm .003$	$0.054 \pm .038$	$-0.864 \pm .020$	$2.040 \pm .098$	$-0.336 \pm .031$	$-0.864 \pm .022$
	FullyNeural	$-0.789 \pm .003$	$0.059 \pm .037$	$-0.833 \pm .020$	$4.699 \pm .054^{\dagger}$	$0.206 \pm .046^{\dagger}$	$-0.810 \pm .022$
	NeuralHawkes	$-0.777 \pm .006^{\dagger}$	$0.066 \pm .037^{\dagger}$	$-0.821 \pm .021^{\dagger}$	$4.641 \pm .110$	$0.201 \pm .048$	$-0.801 \pm .023^{\dagger}$
UNIPoint	ExpSum	$-0.774 \pm .008^{\ddagger}$	$0.056 \pm .042$	$-0.828 \pm .020$	$3.114 \pm .125$	$0.151 \pm .045$	$-0.812 \pm .023$
	PLSum	$-0.779 \pm .006$	$0.064 \pm .038^{\ddagger}$	$-0.829 \pm .020$	$\mathbf{4.939 \pm .085}^{\ddagger}$	$0.162 \pm .046$	$-0.814 \pm .023$
	ReLUSum	$-0.780 \pm .007$	$0.059 \pm .039$	$-0.828 \pm .021$	$4.676 \pm .075$	$\mathbf{0.221 \pm .046}^{\ddagger}$	$-0.810 \pm .023$
	CosSum	$-0.777 \pm .008$	$0.062 \pm .039$	$-0.828 \pm .020$	$4.471 \pm .075$	$0.139 \pm .044$	$-0.814 \pm .023$
	SigSum	$-0.776 \pm .007$	$0.064 \pm .038$	$-0.827 \pm .020^{\ddagger}$	$4.346 \pm .076$	$0.170 \pm .045$	$-0.814 \pm .023$
	MixedSum	$-0.779 \pm .007$	$0.062 \pm .038$	$-0.828 \pm .020$	$4.928 \pm .085$	$0.201 \pm .047$	$-0.804 \pm .023^{\ddagger}$

Table 10: Averaged log-likelihood scores with corresponding 95% confidence intervals. A higher score is better; the best of the baselines are indicated by † and the best of the UNIPoint models are indicated by ‡. Bold indicates results when the difference between † and ‡ are *significantly better* (t-test $p = 0.05$).

Synthetic datasets. Contrasting the log-likelihood and total variation metrics reveal interesting insights about model performance. The SelfCorrecting dataset has a piece-wise monotonically increasing intensity function. Both metrics indicate that ExpHawkes, PLHawkes, and RMTPP under perform the other approaches by a large margin, since they are restricted to piece-wise monotone intensity functions. All UNIPoint variants perform well, achieving average likelihoods within 0.01 of each other. ExpSum is the best variant, possibly due to its exponential shape matching that of the ground-truth SelfCorrecting intensity function.

Real-world datasets. For all three real-world datasets, baselines ExpHawkes, PLHawkes, and RMTPP significantly under-perform in comparison to the rest of the approaches. This likely occurs due to their inability to support non-monotone intensity functions in inter-event intervals.

We observe that UNIPoint variants are significantly better than the baselines for MOOC and Reddit. UNIPoint is second best (to NeuralHawkes) on StackOverflow dataset, but the difference is not statistically significant. NeuralHawkes performs strongly on the StackOverflow dataset, potentially because it has the closest architecture to the UNIPoint ExpSum variant, while also being more complex. In particular NeuralHawkes has time decaying hidden states and LSTM recurrent units rather than the a perceptron recurrent unit and vector-formed hidden state of UNIPoint.

Using mixture of basis function, MixedSum provides good overall performance. Among the UNIPoint variants, it is either the best or a close second across all datasets, suggesting that using different types of basis functions improves model flexibility in practice even with a fixed parameter budget. We also observe an improvement in performance when more basis functions are used, see Appendix E.

Overall, our evaluations demonstrate the power of UNIPoint for modelling complex intensity function that are not piece-wise monotone. Results on real-world datasets show models with flexible intensity functions outperform Hawkes processes. Open questions remain on which neural architectures, among the ones with universal approximation power, strike the best balance of representational power, parsimony, and learnability.

6.6 Results: Forecasting for unseen content

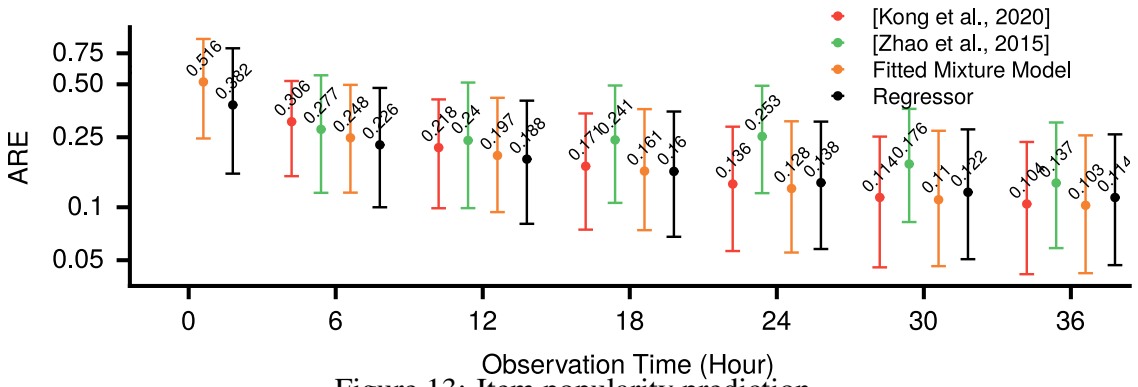


Figure 13: Item popularity prediction

Figure 14: Forecasting for unseen content on *ActiveRT2017-Test*. Item final popularity predictions using four models evaluated with Absolute Relative Error (ARE) — lower is better. Times at x axis are the observation times since an online item was published. The dots indicate the median values and error bars give the 25^{th}/75^{th} quantiles of the ARE values.

Prediction of final popularity. We compare the final popularity predictions on *ActiveRT2017-Test* with dual mixture models against a predictor built using *Seismic* [70], an ensemble model in [33] and a regressor trained using temporal features. *Seismic* and the ensemble model predictions are produced by their provided R packages. Since both models were designed to predict the final popularities of individual cascades, we build an item popularity predictor by following the same steps as in section 5.3.3 and using the predictions instead of $\hat{N}_{v,i}(n^*, \Theta^g)$ in eq. (39). We construct the regressor using the temporal features of six-point summaries (min, mean, median, max, 25^{th} and 75^{th} percentile) of inter-arrival times, cascade sizes, cascade durations and number of followers of Twitter users involved in cascades, and the tuples (observation times, online items) for the set of examples, and the item final popularity is the dependent variable to predict. We train a single regressor using the `GBM` package in R [23], and we obtain predictions for each tuple via 5-fold cross validation on *ActiveRT2017-Fit*. Finally, final popularity predictions of the dual mixture models are computed using eq. (40) and at each observation time T. We note that we re-fit the BMMs on cascades after the time T in historical cascades to capture changes of content virality in time.

We evaluate the prediction results using the Absolute Relative Error (ARE) — also used in [70] and defined as $\frac{|\hat{N}_v - N_v|}{N_v}$ where $\hat{N}_v$ and N_v are the predicted popularity and the actual final popularity.

fig. 13 summarizes the prediction results, with the ARE values in log scale. As *Seismic* and the ensemble models do not provide cold-start predictions, only results for the dual mixture models and the regressor are presented at $T = 0$ observation time. We see that both the dual mixture models and the temporal features regressor consistently outperform the other two baselines, *Seismic* and the ensemble model, up to the 18-hour observation time. Also, the regressor slightly outperforms the dual-mixture model for short observation times, after which the dual-mixture model delivers the best predictive performances.

References

[1] Linda J. S. Allen. An introduction to stochastic epidemic models. In *Mathematical Epidemiology*, chapter 3. Springer, 2008.

[2] Shushan Arakelyan, Fred Morstatter, Margaret Martin, Emilio Ferrara, and Aram Galstyan. Mining and forecasting career trajectories of music artists. In *Hypertext*, 2018.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv*, 2017.

[4] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 2015.

[5] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[6] Alexandre Bovet and Hernán A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Commun.*, 2019.

[7] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 2000.

[8] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 2004.

[9] Civiqs. Do you support or oppose the black lives matter movement? https://civiqs.com/results/black_lives_matter/, 2020. (Accessed on 09/20/2020).

[10] D. R. Cox and D Oakes. *Analysis of survival data*. Routledge, 1984.

[11] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[12] Daryl J Daley and David Vere-Jones. Conditional intensities and likelihoods. In *An introduction to the theory of point processes*, volume I, chapter 7.2. Springer, 2008.

[13] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. Social media participation in an activist movement for racial equality. In *ICWSM*, 2016.

[14] Chen Debao. Degree of approximation by superpositions of a sigmoidal function. *Approximation Theory and its Applications*, 9(3):17–28, 1993.

[15] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977.

[16] Jeff Diamant and Aleksandra Sandstorm. Do state laws on abortion reflect public opinion? — pew research center. https://www.pewresearch.org/fact-tank/2020/01/21/do-state-laws-on-abortion-reflect-public-opinion/, 2020. (Accessed on 04/30/2020).

[17] Christian Donner and Manfred Opper. Efficient bayesian inference of sigmoidal gaussian cox processes. *The Journal of Machine Learning Research*, 19(1):2710–2743, 2018.

[18] Elephrame. At least 5,200 black lives matter protests and other demonstrations have been held in the past 2,279 days. https://elephrame.com/textbook/BLM/chart, 2020. (Accessed on 09/30/2020).

[19] Centers for Disease Control, Prevention. Web based Injury Statistics Query, and Reporting System (WISQARS) [Online]. National center for injury prevention and control, centers for disease control and prevention. https://wisqars-viz.cdc.gov:8006/explore-data/home, 2003. (Accessed on 06/15/2020).

[20] R Kelly Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, 2009.

[21] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. The structure of online diffusion networks. In *EC*, 2012.

[22] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. Assessing the bias in samples of large online networks. *Social Networks*, 2014.

[23] Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and GBM Developers. *gbm: Generalized Boosted Regression Models*, 2019. R package version 2.1.5.

[24] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 2019.

[25] Frank A. Haight and Melvin Allen Breuer. The Borel-Tanner Distribution. *Biometrika*, 1960.

[26] Frank E Harrell Jr, Charles Dupont, et al. Hmisc: Harrell miscellaneous. r package version 4.0-3. *Online publication*, 2017.

[27] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 1971.

[28] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[29] Yiqing Hua, Thomas Ristenpart, and Mor Naaman. Towards measuring adversarial Twitter interactions against candidates in the US midterm elections. In *ICWSM*, 2020.

[30] Jigsaw. Perspective API. https://perspectiveapi.com, 2021. Accessed: 2021-04-14.

[31] Bindu Kalesan, Marcos D Villarreal, Katherine M Keyes, and Sandro Galea. Gun ownership and social gun culture. *Injury prevention*, 22(3):216–220, 2016.

[32] Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. Describing and predicting online items with reshare cascades via dual mixture self-exciting processes. In *International Conference on Information and Knowledge Management*, 2020.

[33] Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. Modeling information cascades with self-exciting processes via generalized epidemic models. In *WSDM*, 2020.

[34] Patrick J Laub, Thomas Taimre, and Philip K Pollett. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015.

[35] Phillip B Levine and Douglas Staiger. Abortion policy and fertility outcomes: the eastern european experience. *The Journal of Law and Economics*, 47(1):223–243, 2004.

[36] Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. Inside the right-leaning echo chambers: Characterizing Gab, an unmoderated social system. In *ASONAM*, 2018.

[37] Rafael Lima and Jaesik Choi. Hawkes process kernel structure parametric search with renormalization factors. *arXiv preprint arXiv:1805.09570*, 2018.

[38] Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts. Variational inference for gaussian process modulated poisson processes. In *International Conference on Machine Learning*, pages 1814–1822, 2015.

[39] Olga Lukočienė and Jeroen K Vermunt. Determining the number of components in mixture models for hierarchical data. In *Advances in data analysis, data handling and business intelligence*. Springer, 2009.

[40] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[41] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.

[42] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Feature driven and point process approaches for popularity prediction. In *CIKM*, 2016.

[43] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *ICWSM*, 2013.

[44] Mark EJ Newman. The structure of scientific collaboration networks. *PNAS*, 2001.

[45] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

[46] Takahiro Omi, Naonori Ueda, and Kazuyuki Aihara. Fully neural network based model for general temporal point processes. In *Advances in Neural Information Processing Systems 32*, pages 2120–2129. Curran Associates, Inc., 2019.

[47] Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.

[48] Kim Parker, Juliana Menasce Horowitz, Ruth Igielnik, J. Baxter Oliphant, and Anna Brown. Guns in america: Attitudes and experiences of americans — pew research center. https://www.pewsocialtrends.org/2017/06/22/americas-complex-relationship-with-guns/, 2017. (Accessed on 10/13/2020).

[49] Elizabeth Witwer Rachel K. Jones and Jenna Jerman. Abortion incidence and service availability in the united states, 2017 — guttmacher institute. https://www.guttmacher.org/sites/default/files/report_pdf/abortion-incidence-service-availability-us-2017.pdf, 2019. (Accessed on 04/30/2020).

[50] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. Sir-hawkes: on the relationship between epidemic models and hawkes point processes. In *WWW*, 2018.

[51] Marian-Andrei Rizoiu and Lexing Xie. Online popularity under promotion: Viral potential, forecasting, and the economics of time. In *ICWSM*, 2017.

[52] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be HIP: Hawkes intensity processes for social media popularity. In *TheWebConf*, 2017.

[53] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

[54] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[55] Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina. Correcting for missing data in information cascades. In *WSDM*, 2011.

[56] Anton Maximilian Schäfer and Hans-Georg Zimmermann. Recurrent neural networks are universal approximators. *International journal of neural systems*, 17(04):253–263, 2007.

[57] Alexander Soen, Alexander Patrick Mathews, Daniel Grixti-Cheng, and Lexing Xie. Unipoint: Universally approximating point processes intensities. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[58] Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

[59] X Yu Stella and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, 2003.

[60] Behzad Tabibian, Isabel Valera, Mehrdad Farajtabar, Le Song, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Distilling information reliability and source trustworthiness from digital traces. In *Proceedings of the 26th International Conference on World Wide Web*, pages 847–855, 2017.

[61] Carlo Tomasi. Estimating gaussian mixture densities with em–a tutorial. *Duke University*, 2004.

[62] A Wächter and L T Biegler. On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming. *Mathematical Programming*, 2006.

[63] Siqi Wu and Paul Resnick. Cross-partisan discussions on youtube: Conservatives talk to liberals but liberals don't talk to conservatives. In *ICWSM*, 2021.

[64] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond views: Measuring and predicting engagement in online videos. In *ICWSM*, 2018.

[65] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Estimating attention flow in online video networks. In *CSCW*, 2019.

[66] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Variation across scales: Measurement fidelity under twitter data sampling. In *ICWSM*, 2020.

[67] Ping Yan. Distribution theory, stochastic processes and infectious disease modelling. In *Mathematical Epidemiology*, chapter 10. Springer, 2008.

[68] Jun Zhang, Mark S Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *WWW*, 2007.

[69] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *KDD*, 2015.

[70] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *KDD*, 2015.

[71] Zhi-Hua Zhou. Combination methods. In *Ensemble methods: foundations and algorithms*, chapter 4. Chapman and Hall/CRC, 2012.

[72] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. *arXiv preprint arXiv:2002.09291*, 2020.