



AFRL-AFOSR-JP-TR-2019-0002

154002 The Anatomy of Social Media Popularity

Lexing Xie
AUSTRALIAN NATIONAL UNIVERSITY RESEARCH OFFICE ACTON
(AUSTRALIA)
CHANCELRY BLDG 10B LG EAST RD
ACTON, 200
AU

01/15/2019
Final Report

DISTRIBUTION A: Distribution approved for public release.

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				Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Executive Services, Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.					
1. REPORT DATE (DD-MM-YYYY) 15-01-2019		2. REPORT TYPE Final		3. DATES COVERED (From - To) 15 Sep 2015 to 14 Sep 2018	
4. TITLE AND SUBTITLE 154002 The Anatomy of Social Media Popularity				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER FA2386-15-1-4018	
				5c. PROGRAM ELEMENT NUMBER 61102F	
6. AUTHOR(S) Lexing Xie				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AUSTRALIAN NATIONAL UNIVERSITY RESEARCH OFFICE ACTON (AUSTRALIA) CHANCELRY BLDG 10B LG EAST RD ACTON, 200 AU				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AOARD UNIT 45002 APO AP 96338-5002				10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/AFOSR IOA	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-AFOSR-JP-TR-2019-0002	
12. DISTRIBUTION/AVAILABILITY STATEMENT A DISTRIBUTION UNLIMITED: PB Public Release					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The Anatomy of Social Media Popularity project studied popularity as a state of being liked or supported by many people. The approaches primarily stem from the computer science and mathematics disciplines. Examples include the development of the Hawkes Intensity Process (HIP) which factored three intuitions: magnitude of user influence, content quality, and decaying social memory to develop a predictive model. This model was then refined into Recurrent Neural Network models to forecast (verified with actual data) or predict (infer from past data only) popularity from heterogeneous streams to quantify average response to unit promotion and relative influence amongst users of varying fame levels. These models identified four sets of additional metrics to quantify the likelihood of an online event going viral, requiring the development of an algorithm to estimate user influence. The RNN-MAS model outperformed YouTubes popularity prediction system by 17% and also captured seasonal trends of unseen influence, though performance was sensitive to content type (e.g. superusers in Gaming videos, cohorts of regular users in Activism videos). The experimental portion of this work also created five new video datasets. Code developed under this project is openly hosted on GitHub. The contributions of this work advances the science of understanding social influences in the cyber domain.					
15. SUBJECT TERMS Viral Messages					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			LIN, ALAN
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER (Include area code) 315-227-7009

AOARD 15IOA018 – Final Report

Project title: The Anatomy of Social Media Popularity

Principle investigator:

Lexing Xie, lexing.xie@anu.edu.au

Co-principle investigators:

Pascal Van Hentenryck, Georgia Institute of Technology, pascal.vanhentenryck@isye.gatech.edu

Manuel Cebrian, Massachusetts Institute of Technology, cebrian@media.mit.edu

Key personnel:

Marian-Andrei RizoIU, marian-andrei.rizoIU@anu.edu.au

PhD students:

Swapnil Mishra (Sept 2015 - Sept 2018)

Siqi Wu (July 2016 - Jan 2020)

Quyu Kong (March 2018 - Nov 2021)

Organization:

Research School of Computer Science, Hanna Neumann building

The Australian National University, Acton, ACT 2601, Australia.

Computational Media Lab: <http://cm.cecs.anu.edu.au>

Project period of performance:

15 Sept 2015 - 14 Sept 2018

Project abstract

Popularity is a social phenomenon, a state of being liked or supported by many people. Popularity can be measured for an individual (e.g., in elections), a unit of work (e.g., by awards for books and films), a physical item (e.g., in online sales ranking), or an online item (e.g., in views and shares to a photo or video). Popularity captures people's attention, as it directly links to political outcome, monetary awards and various notions of fame. Yet popularity is known to be very difficult to predict, in part because the mechanisms that drive popularity is not well understood.

This project develops methods and tools for describing and predicting the popularity of online items. In particular, we aim to describe the complex evolution of popularity over time, estimate the extent that popularity is driven by external promotions versus word-of-mouth diffusion, and quantify the strength of different influencing factors. The research issued from this project allowed us to design several novel metrics to quantify the expected popularity gain per unit promotion [[RizoIU and Xie, 2017](#)], the relative effect of celebrity versus grass-root users [[Mishra et al., 2018](#)], and the effect of latent influences [[RizoIU et al., 2018a](#)].

Project Highlights

We begin by summarizing four project highlights during the projects' performance period. These contributions are further detailed in the **Progress summary** and in the **Technical Sections** of this report.

Four new models, capturing popularity dynamics at different granularities (event level and aggregate level) and across platform boundaries:

- The Hawkes Intensity Process, or HIP, a novel model that describes volumes of attention, or popularity series directly (Tech. Sec. 1.2, paper [Rizoiu et al., 2017]).
- A Hawkes point processes model explicitly tuned for prediction. The point process part capturing three key intuitions: magnitude of user influence, content quality and decaying social memory; the predictive parts take into account aggregate features. (Technical Sec. 1.1 of this report, paper [Mishra et al., 2016a]).
- Uncovering the underlying relationship between self-exciting stochastic processes and epidemic models, namely that a Hawkes point process has the same event intensity as an SIR model in which the recovery events are unobserved (Technical Sec. 1.3 of this report, paper [Rizoiu et al., 2018c]).
- A Recurrent Neural Network model capable of capturing popularity dynamics at different granularities (event level and aggregate level) and across platform boundaries. (Technical Sec. 1.4 of this report, paper [Mishra et al., 2018]).

Four sets of metrics to quantify the relation between users and online cultural items:

- Two novel metrics (based on HIP model [Rizoiu et al., 2017]) to quantify online popularity under promotion: the *viral potential* measures popularity gain per unit of promotion, and *maturity time* is the time it takes for such effects to unfold (Tech. Sec. 3.1, paper [Rizoiu and Xie, 2017]).
- A set of engagement metrics for online videos, including watch time, average watch percentage, and the novel *relative engagement*, calibrated with respect to video length, stable over time, and strongly correlated with video quality (Tech. Sec. 3.2, paper [Wu et al., 2018]).
- Two new metrics to capture the effects of particular types of users tweeting about Youtube videos: the promotion score and the loudness level (Tech. Sec. 3.3, paper [Mishra et al., 2018]).
- An estimation algorithm and a measure of user influence starting from online Twitter diffusion, where the network structure between users is latent (Paper [Rizoiu et al., 2018a]).

Nine new applications and insights in large-scale Twitter and YouTube data.

- Explaining popularity with a novel visualization called the *endo-exo map*. With this map one can describe the most popular videos and detect unpromotable videos (Tech. Sec. 4.2, paper [Rizoiu et al., 2017]);
- Predicting the total size of a diffusion cascade, by making better use of both feature-driven and generative model approaches. (Tech. Sec. 4.1, paper [Mishra et al., 2016a]);
- Forecasting popularity dynamics under external promotions using the Hawkes Intensity Process model. Prediction made via HIP are more accurate, and can uniquely identify videos that may go viral (Tech. Sec. 4.3, paper [Rizoiu et al., 2017]);
- Analyse factors influencing the forecasts of popularity, and predict popularity jumps (Tech. Sec. 4.4, paper [Rizoiu and Xie, 2017]);
- Quantify economic value versus time in promotion, and provide guidelines for constructing cost-effective promotion schedules (Tech. Sec. 4.5, paper [Rizoiu and Xie, 2017]);
- Explain why constant promotion is desirable – that this schedule increases the perceived *viral*

potential (Tech. Sec. 4.5.2, paper [Rizoiu and Xie, 2017]);

- Explain user engagement using a novel visualization called the *engagement map*. This map allows to visualize higher quality videos, as identified by independent external sources (Tech. Sec. 4.6, paper [Wu et al., 2018]);
- Study the relative effect of tweeting by either celebrities or grass-root users. This allows us to select which category is more effecting in promoting particular types of online content (Tech. Sec. 4.7, paper [Mishra et al., 2018]);
- A possible explanation toward the unpredictability of online popularity, drawn from the theoretical distribution of final size of a diffusion cascade which exhibits two modes. (Tech. Sec. 4.8, paper [Rizoiu et al., 2018c]);

Five new datasets, curated for tracking longitudinal user engagement in a collection of online videos:

(1) The TWEETTED VIDEOS dataset [Rizoiu et al., 2017], which captures the daily number of views, shares and the length of watch time for more than 5 Million videos, posted by 1.2 Million channels during a period of two months (July 1st, 2016 to August 31st, 2016) (Tech. Sec. 2, paper [Wu et al., 2018]).

(2) The tweeted news dataset, which contains tweets about news articles in ten major news outlets. This dataset provides a common benchmark for the feature-driven and generative models in popularity prediction, as well as for classification and regression problem settings (Tech. Sec. 2, paper [Mishra et al., 2016a]).

(3–5) QUALITY VIDEOS datasets, containing videos – and their viewing and engagement information – which were rated as high quality by domain experts. Two of these datasets are on Music (VEVO and BILLBOARD) and one is on News (Tech. Sec. 2, paper [Wu et al., 2018]).

Publication list

The publications from this project are listed below. In summary, there has been 9 full papers in competitive high-impact venues, 1 tutorial book chapter, and 1 conference demo. Note that WWW, ICWSM, WSDM and CIKM are the most prestigious Computer Science conferences (with full-length papers) for publishing work on social media, their acceptance rates are between 15% and 18%.

- [Yu et al., 2015] Honglin Yu, Lexing Xie, and Scott Sanner. The Lifecycle of a Youtube Video: Phases, Content and Popularity. In *Ninth International AAAI Conference on Web and Social Media*, page 10, 2015. <http://users.cecs.anu.edu.au/~xlx/papers/icwsm15-phase.pdf>
Summary: This paper proposes a novel phase representation that extends the well-known endogenous growth and exogenous shock model for video popularity. We find that videos go through not one, but multiple stages of popularity increase or decrease over many months;
- [Mishra et al., 2016a] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Feature Driven and Point Process Approaches for Popularity Prediction. In *Proceedings of International Conference on Information and Knowledge Management - CIKM '16*, page 10, Indianapolis, USA, 2016. 10.1145/2983323.2983812. <https://arxiv.org/pdf/1608.04862.pdf>
Summary: This work proposes a Hawkes point process for modeling information diffusion

in Social Media, as well as a thorough benchmark for predicting the total volume of attention – between different classes of approaches (i.e. generative and feature-driven) and across different problem settings, i.e. as classification and regression problems.

Technical sections: 1.1 and 4.1;

- [Rizoiu et al., 2016] Marian-Andrei Rizoiu, Lexing Xie, Tiberio Caetano, and Manuel Cebrian. Evolution of Privacy Loss on Wikipedia. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, WSDM '16, San Francisco, CA, USA, 2016. 10.1145/2835776.2835798. <http://arxiv.org/pdf/1512.03523.pdf>
Summary: The digital traces left behind by the users in the online environment reveal more about them than they might like. We show that machine learning algorithms can be used to uncover hidden links between an user's past activity and her private traits, like gender, education level or religious views, even for retired users.
- [Rizoiu et al., 2017] 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 *26th International Conference on World Wide Web - WWW '17*, pages 735–744, Perth, Australia, 2017. ACM Press. ISBN 9781450349130. 10.1145/3038912.3052650. <http://arxiv.org/pdf/1602.06033.pdf>.
Summary: We propose the Hawkes Intensity Process (HIP), a novel generative model to describe popularity as the result of endogenous network reactions constant exogenous influence. We construct the endo-exo map to visualize the propensity of videos to react to external stimulation, and we identify viral and un-promotable videos.
Technical sections: 1.2, 4.2 and 4.3;
- [Rizoiu and Xie, 2017] Marian-Andrei Rizoiu and Lexing Xie. Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *11th International AAAI Conference on Web and Social Media ICWSM'17*, pages 182–191, 2017. <https://arxiv.org/pdf/1703.01012.pdf>.
Summary: We propose two novel metrics based on the HIP model: to describe popularity gain per unit of promotion, and to quantify the time it takes for such effects to unfold. We study the factors that influence the accuracy popularity forecasts and we illustrate the interplay between the unfolding of popularity over time, and the time-sensitive value of resources.
Technical sections: 3.1, 4.4 and 4.5.
- [Rizoiu et al., 2018b] Marian-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. A Tutorial on Hawkes Processes for Events in Social Media. In Shih-Fu Chang, editor, *Research Frontiers of Multimedia*, pages 1–26. ACM Books, 2017. <https://arxiv.org/pdf/1708.06401.pdf>.
Summary: The chapter provides an accessible introduction for point processes, and especially Hawkes processes, for modeling discrete, inter-dependent events over continuous time. We review definitions and key concepts, and we describe a practical example drawn from social media data – we show how to model retweet cascades using Hawkes self-exciting processes.
- [Kong et al., 2018] Quyu Kong, Marian-Andrei Rizoiu, Siqi Wu, and Lexing Xie. Will this video go viral? explaining and predicting the popularity of youtube videos. In *WWW*

2018 Companion: The 2018 Web Conference Demos, Lyon, France, 2018. URL <https://arxiv.org/pdf/1801.04117.pdf>.

Summary: In this demo paper we introduce HIPie – an interactive visualization system – created to fill visualize the HIP popularity model [Rizoiu et al., 2017] and to enable users to reason about the virality and the popularity of online videos.

Technical sections: 1.2, 2.2 and 3.2.

- [Mishra et al., 2018] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Conference on Web and Social Media – ICWSM’18*, pages 1–10, Stanford, CA, USA, 2018. <https://arxiv.org/pdf/1804.02101.pdf>.

Summary: This paper proposes RNN-MAS, a recurrent neural network for modeling asynchronous streams, which not only outperforms the current state-of-the-art Youtube popularity prediction systems, but also captures complex dynamics, such as seasonal trends of unseen influence. We also define two new metrics: the promotion score and the loudness level, and we use them to compare the effects of a video being promoted by a single highly-followed user against a group of mid-followed users.

Technical sections: 1.4, 3.3 and 4.7.

- [Wu et al., 2018] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond Views: Measuring and Predicting Engagement in Online Videos. In *International AAAI Conference on Web and Social Media – ICWSM’18*, Stanford, CA, USA, 2018. <https://arxiv.org/abs/1709.02541>.

Summary: We study a set of engagement metrics, including watch time, percentage of video watched, and relate them to views and video properties such as length and content category. We propose a new metric, *relative engagement*, which is calibrated over video duration, stable over time, and strongly correlated with video quality.

Technical sections: 3.2 and 4.6.

- [Rizoiu et al., 2018a] Marian-Andrei Rizoiu, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie. #DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate. In *International AAAI Conference on Web and Social Media – ICWSM’18*, pages 1–10, Stanford, CA, USA, 2018a. <https://arxiv.org/abs/1802.09808>.

Summary: In this paper, we define a measure of user influence based on the user’s active contributions to information diffusions, i.e. their tweets and retweets. Given that Twitter does not expose the retweet structure - it associates all retweets with the original tweet - we model the latent diffusion structure using only tweet time and user features, and we implement a scalable novel approach to estimate influence over all possible unfoldings.

- [Rizoiu et al., 2018c] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference, WWW ’18*, Lyon, France, 2018c. <https://arxiv.org/pdf/1711.01679.pdf>.

Summary: In this paper, we establish a novel connection between self-exciting point processes and epidemic models. Namely, the rate of events in an extended Hawkes model is identical to the rate of new infections in the Susceptible-Infected-Recovered (SIR) model

after marginalizing out recovery events – which are unobserved in a Hawkes process. This result leads to HawkesN, a generalization of the Hawkes model which accounts for a finite population size and we derive the distribution of cascade sizes for HawkesN, inspired by methods.

Technical sections: 1.3 and 4.8.

Awards and Honours

- Lexing Xie was awarded **The Chris Wallace Award for Outstanding Research** in Jan 2018. One recipient per year (at associate professor or below) is selected by the Computing Research and Education Association of Australasia (CORE) annually.
- Lexing Xie held **IEEE Circuits and Systems Society (CAS) Distinguished Lecturership**, 2016-2017.

Publicity and outreach

The research issued from this project enjoyed quite some attention from the news media. Here below we list such coverage:

- Our Wikipedia privacy work was featured in the March 2016 Wikimedia Research showcase, https://www.mediawiki.org/wiki/Wikimedia_Research/Showcase#March_2016.
- A dedicated profile of our YouTube video lifecycle work appeared in [ANU Reporter](#) and [NCI News](#).
- In April 2017, our group was prompted by an inquiry from the Polish online news www.press.pl, asking about the impact of emergent sources of information versus traditional media. We used our extensive TWEETED VIDEOS dataset to study how YouTube videos posted by two traditional news sources (BBC news, ABC news) and two emergent online news sources (The Alex Jones Channel, The Young Turks) are viewed. The results were covered by the Polish news agency (<http://www.press.pl/magazyn-press/artykul/46939,informacje-z-youtube>) and detailed in our blogpost **Escape from broadcasting: Talk shows as news sources** (post: http://cm.cecs.anu.edu.au/post/youtube_news/).
- ANU College of Engineering and Computer Science featured our Google PhD Fellowship recipient Siqi Wu and his work on Understanding Human Behavior in Online Videos (post: <http://cs.anu.edu.au/news/anu-student-awarded-google-phd-fellowship>).
- The [Business Insider](#), the [Sydney Morning Herald](#) and the [ANU Reporter](#) wrote about our findings concerning the bot influence in the 2016 US Elections [[Rizoïu et al., 2018a](#)].

During the performance period of this project, the members of the ComputationalMedia group presented the results of our research for the broader audience on the group homepage (<http://cm.cecs.anu.edu.au>), both for publicity and for outreach within the general public. The following blog posts are relevant to the research in this project:

- **Feature Driven and Point Process Approaches for Popularity Prediction** explains the modeling and the prediction of popularity of retweet cascades (post: <http://cm.cecs.anu.edu.au/post/fdhawkesforpopularity/>, paper [Mishra et al., 2016a]);
- **Expecting to be HIP (I) - Deriving Stochastic Expectations** provides a gentle introduction to the math underlying the HIP model (post: http://cm.cecs.anu.edu.au/post/hawkes_intensity/, paper [Rizoïu et al., 2017]);
- **Expecting to be HIP (II) – Could this video go viral?** explains the problem and the main results in online video popularity modeling using HIP and introduces the HIP Demo (post: http://cm.cecs.anu.edu.au/post/expecting_to_be_HIP/, paper [Rizoïu et al., 2017]);
- **Expecting to be HIP (III) – Examining the role of promotions** examines popularity under promotion and the construction of promotion schedules (post: <http://cm.cecs.anu.edu.au/post/promotability/>, paper [Rizoïu and Xie, 2017]);
- **Expecting to be HIP (IV) - HIPsters Unite!** acknowledges the people that involved in the development of the HIP framework and it links publications and resources (post: http://cm.cecs.anu.edu.au/post/hipsters_unite/).

Software, Tutorials, Demo and Datasets

This project led to the following software artifacts, tutorials and datasets.

Software artifacts

- A **web-based, interactive visualizer system** was constructed based on the HIP model. A screenshot of the system is presented in Fig. 1. The system was presented in the 2017 Trust and Influence Program Review in Colorado, at the WWW’17 conference in Perth Australia and at the ICWSM’17 conference in Montréal Canada. It is further detailed in [Kong et al., 2018], a demo paper presented at WWW’18 in Lyon, France. It is publicly accessible at <http://hipie.ml>, and the source code is available at <https://github.com/computationalmedia/hipie>.
- The **code for simulating and fitting the HIP model** is published in the GitHub repository <https://github.com/andrei-rizoïu/hip-popularity>. The code is available both in R and in Python. The code corresponds to the paper [Rizoïu et al., 2017] and can compute the measures for online popularity under promotion [Rizoïu and Xie, 2017].
- The **code for fitting and simulating the Hawkes point-process** for social media is publicly available in the GitHub repository <https://github.com/s-mishra/featured-driven-hawkes>. This software is associated with the paper [Mishra et al., 2016a].
- The **code for computing the user engagement** is published in the GitHub repository <https://github.com/avalanchesiqi/youtube-engagement>. The code corresponds to the paper [Wu et al., 2018].

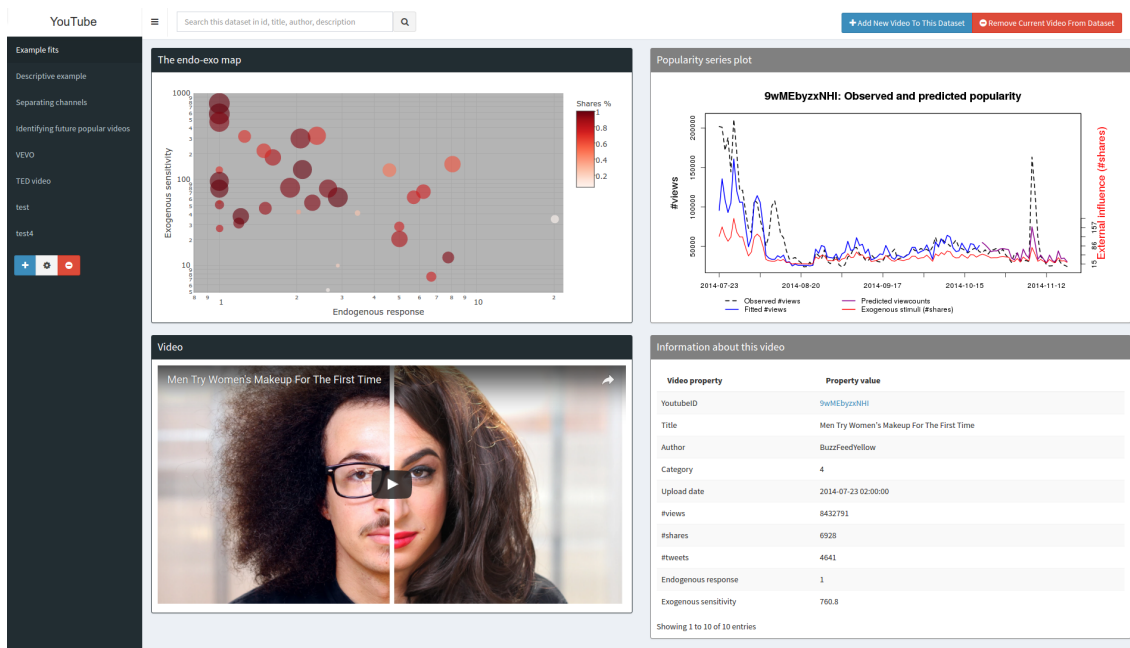


Figure 1: Screenshot of the HIPie [Kong et al., 2018] – our interactive HIP visualization system. The top-left panel of the interface shows the selected videos projected on the endo-exo map [Rizoiu et al., 2017]. The diameter of circles is proportional to video popularity, their color intensity is proportional to the promotions they received. Once a video is selected, its popularity over time – both fitted and observed – as well as its metadata are shown on the right side panels. It features visualizations of TED videos and VEVO musicians. Furthermore, it allows users to add and compare their own videos. It is publicly accessible at <http://hipie.ml>.

- The **code for estimating user influence from diffusion cascades with latent structure** is published in the GitHub repository <https://github.com/computationalmedia/cascade-influence>. The code corresponds to the paper [Rizoiu et al., 2018a].
- The **code for fitting and simulating HawkesN and the SIR epidemic model** is publicly available in the GitHub repository <https://github.com/computationalmedia/sir-hawkes>. This software is associated with the paper [Rizoiu et al., 2018c].

Tutorials

- We constructed a short tutorial for the using the ACTIVETWEETS dataset and the HIP popularity modeling system. It is publicly available online at <https://github.com/andrei-rizoiu/hip-popularity#using-hip-and-the-active-dataset-a-short-tutorial>, and it is associated with the paper [Rizoiu et al., 2017].
- We provide a step-by-step programming tutorial for researchers and practitioners who would like to start modeling with Hawkes point processes https://github.com/s-mishra/featured-driven-hawkes/blob/master/code/marked_hawkes_point_process.ipynb. This tutorial is associated with the paper [Mishra et al., 2016a].
- We constructed a hand-on tutorial showing how to simulate a stochastic SIR epidemic model, and how to fit and SIR and a HawkesN model on the synthetic data, showing their parameter equivalence. It also shows how to compute the size distribution for a diffusion cascade. It is publicly available online at <https://github.com/computationalmedia/sir-hawkes>, and it is associated with the paper [Rizoiu et al., 2018c].

Datasets

- TWEETED VIDEOS dataset is continuously collected since May 2014, from all the tweets that link to a YouTube video. We constructed and we employed several subsets of this dataset:
 - A six-month subset (called ACTIVETWEETS dataset) contains 13,738 videos, which were uploaded between 2014-05-29 and 2014-08-09, is used in [Rizoiu et al., 2017] and [Mishra et al., 2018] to model and predict online popularity (Tech. Sec. 1.2 and 1.4, paper [Rizoiu et al., 2017] and [Wu et al., 2018]).
 - A two-month subset containing more than 5 Million videos during 2016-07-01 and 2016-08-31, is used in [Wu et al., 2018] to measure user engagements with online videos (Tech. Sec. 2, paper [Rizoiu et al., 2017] and [Wu et al., 2018]).
- QUALITY VIDEOS datasets containing videos – and their viewing and engagement information – which were rated as high quality by domain experts. Two of these datasets are on Music (VEVO and BILLBOARD) and one is on News. These datasets are used in [Wu et al., 2018] to validate the connection between the relative engagement and video quality (Tech. Sec. 2, paper [Wu et al., 2018]).

- NEWS DATASET was constructed as a domain specific dataset of tweets dealing with news articles. It was collected over a period of four months from 2015-04 to 2015-07. We tracked the official twitter handles of ten popular news outlets, namely NewYork Times, Associated Press, CNN, Washington Post, Reuters, Yahoo News, The Guardian, BBC, Huffington Post and Google News. (paper: [Mishra et al., 2016a]).

Progress summary

We framed the research undertaken in this project into three successive topics focusing on estimation and modeling, prediction, and optimization. During the first year, we mainly concentrated on the profiling online popularity of Youtube videos [Yu et al., 2015] and on the theoretical modeling on online popularity. We modeled online information diffusion cascades using Hawkes Point Processes [Mishra et al., 2016a] and we developed of a novel mathematical model, the Hawkes Intensity Process (HIP) [Rizoiu et al., 2017] for the popularity of Youtube videos. Most of the research in the second year was dedicated to constructing measures of content virality and promotability [Rizoiu and Xie, 2017] and user engagement [Wu et al., 2018]. In the third year we branched out by linking two different classes of popularity models [Rizoiu et al., 2018c], developing neural network-based solution for popularity prediction and interpretation Mishra et al. [2018] and inferring latent user influence Rizoiu et al. [2018a].

Research undertaken during year 1. We base our observations on the popularity of videos on YouTube and their discussions on Twitter, both of which are publicly available – while the methods and insights apply more generally to social media and collective attention. We first conducted a large scale longitudinal measurement of the popularity change of YouTube videos over several years, and we studied their complex lifecycles, with multiple rises and falls [Yu et al., 2015]. We design novel methods to estimate and segment power-law shaped phases from popularity history of YouTube videos, and correlate phase statistics with video popularity.

We next turned to modeling the diffusion of information within social media using theoretical models. We proposed a Hawkes process for information diffusion [Mishra et al., 2016a], which uses stochastic point-processes to model the word-of-mouth diffusion pattern. The proposed model captures three key intuitions in a social network: *decaying magnitude of influence*, tweets by users with many followers tend to get retweeted more; *memory over time*, that most retweeting happens when the content is *fresh*; and *content quality*. The model is capable of outputting predictions about the total size of the diffusion, making it an accurate tool for predicting the total popularity of an item. We compare our generative approach to another class of methods: feature-driven, which use descriptive features and classical machine learning predictors. Furthermore, we bridge the gap between these solutions with a new hybrid approach and a new performance benchmark. We find that a Hawkes process with a predictive overlay outperform recent feature-driven and generative approaches on existing tweet data and a new public benchmark on news tweets.

We then developed a novel mathematical model, the Hawkes Intensity Process (HIP) [Rizoiu et al., 2017], that systematically explains the complex multi-phase popularity dynamics identified in [Yu et al., 2015] as a continuous influence of exogenous excitation and endogenous response. The key idea is to regard popularity as externally-driven, with exogenous events attracting attention to a video from outside the social network, as shown schematically in the upper flow of Fig. 2a. The

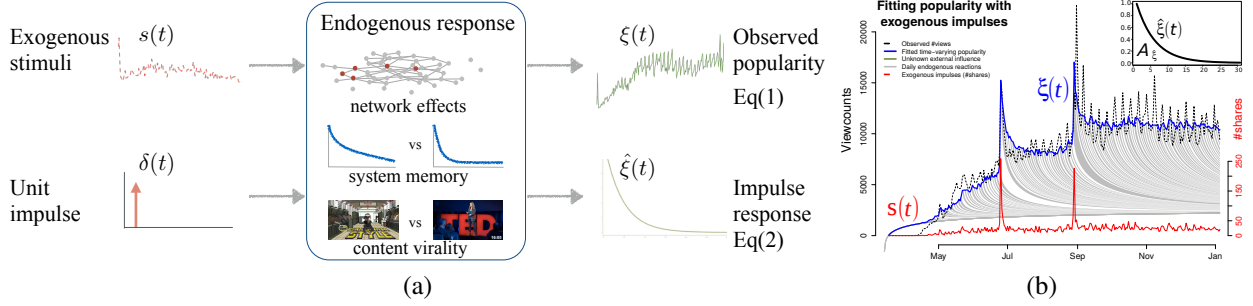


Figure 2: (a) Conceptual schema of popularity as a continuous interplay between exogenous stimuli and endogenous reactions, accounting for content, network and memory effects. Top flow: popularity series $\xi(t)$ is a result of external stimuli $s(t)$ driving endogenous response involving network effects, system memory and content virality. Bottom flow: the impulse response $\hat{\xi}(t)$ is a result of a stimuli of unit impulse driving the same endogenous response. (b) The decomposition of the fitted viewcount series into scaled and temporally shifted *impulse responses* (inset), for the Music video id 0bR4L0Y94AQ.

exogenous events activate an endogenous response inside the network which may, or may not, amplify the exogenous signal. The endogenous response is modeled using stochastic point processes and its size and shape are dependent on content and network parameters, which are fit by observing part of the popularity series. An example of this continuous interplay is shown in Fig. 2b (and furthermore in Fig. 3), we show that this model explains complex multiphase popularity lifecycles well. We also show in [Rizoiu et al., 2017] that it outperforms the state-of-the-art non-parametric model when used to forecast the effect of given promotions.

Research progress during year 2. Based on the modeling advancement made during the first year, we turned to studying online popularity under promotion. Unlike the HIP [Rizoiu et al., 2017] which models popularity under naturally occurring external attention, in [Rizoiu and Xie, 2017] we first study the reaction of online videos to promotion: we describe popularity gain per unit of promotion – dubbed the *viral potential* –, and we quantify the time it takes for such effects to unfold – the *maturity time*. Second, we investigate the factors that influence the forecast of future popularity under promotion: the item’s viral potential, the amount of and the timing of promotions, and the non-linear relationship between attention and popularity. We set out to detect videos with a sudden rise in popularity – i.e. future popular videos – and we find that jointly taking into account the viral potential and the non-linearity of popularity performs best. The third question we investigated was how to schedule promotion to maximize effectiveness for different content. We find that the optimal schedule depends on the items maturity time: the most cost-effective way to promote fast unfolding content is to introduce promotions late; the trade-off between maturity and cost becomes increasingly important for content with slow unfolding.

Next, we approach another angle of online popularity: that of user engagement, or the watching behavior in online videos. Engagement has been measured in reading behavior of news and web pages, click-through-rate for online ads, but not for online videos. In [Wu et al., 2018], we focus on measuring engagement with online videos via large-scale, longitudinal measurements. In particular, we aim to derive video-specific engagement metrics, link them with established notions of content quality. We study a set of engagement metrics, including watch time, percentage of video watched, and relate them to views and video properties such as length and content category. We

propose a new metric, *relative engagement*, which is calibrated over video duration, stable over time, and strongly correlated with video quality.

In the meanwhile, we have also concentrated on completely formalizing the theoretical link between the Hawkes diffusion modeling and the HIP popularity modeling. This amounted in a publication in WWW'17 – one of the top conference venues in the field of Databases & Information Systems¹ – and a book chapter which introduces stochastic point processes to computer scientists [Rizoiu et al., 2018b]. We also produced a number of software packages, tutorials, a demo visualization system and datasets, as described at an earlier section, all freely available for usage and download.

Research progress during year 3. Going back to the theoretical frameworks for modeling information diffusion, we studied the link between two classes of models: epidemic models and Hawkes point processes. The former originate from epidemiology, and consider information as a viral contagion which spreads into a population of online users. The latter has roots in geophysics and finance, view individual actions as discrete events in continuous time, and modulate the rate of events according to the self-exciting nature of event sequences (as in our previous paper [Mishra et al., 2016a]). In [Rizoiu et al., 2018c], we established a novel connection between these two frameworks. Namely, the rate of events in an extended Hawkes model is identical to the rate of new infections in the Susceptible- Infected-Recovered (SIR) model after marginalizing out recovery events – which are unobserved in a Hawkes process. This result paves the way to apply tools developed for SIR to Hawkes, and vice versa. It also allowed us to propose HawkesN, a generalization of the Hawkes model which accounts for a finite population size. Finally, we derive the distribution of cascade sizes for HawkesN, inspired by methods in stochastic SIR. Such distributions provide nuanced explanations to the general unpredictability of popularity: the distribution for diffusion cascade sizes tends to have two modes, one corresponding to large cascade sizes and another one around zero.

In parallel with developing generative models, we have also modeled popularity with neural network architectures. The challenge is that the existing methods typically focus on a single source of external influence, whereas for many types of online content such as YouTube videos or news articles, attention is driven by multiple heterogeneous sources simultaneously - e.g. microblogs or traditional media coverage. In [Mishra et al., 2018], we propose RNN-MAS, a recurrent neural network for modeling asynchronous streams. It is a sequence generator that connects multiple streams of different granularity via joint inference. We show RNN-MAS not only to outperform the current state-of-the-art Youtube popularity prediction system (i.e. our HIP algorithm [Rizoiu et al., 2017]) by 17%, but also to capture complex dynamics, such as seasonal trends of unseen influence. We defined two new metrics: *the promotion score* quantifies the gain in popularity from one unit of promotion for a Youtube video; *the loudness level* captures the effects of a particular user tweeting about the video. We use the loudness level to compare the effects of a video being promoted by a single highly-followed user (in the top 1% most followed users) against being promoted by a group of users with a median number of followers. We found that results depend on the type of content being promoted: superusers are more successful in promoting Howto and Gaming videos, whereas the cohort of regular users are more influential for Activism videos. This work is significant as it provides more accurate and explainable popularity predictions, as well as computational tools for content producers and marketers to allocate resources for promotion campaigns.

¹source: https://scholar.google.com/citations?view_op=top_venues&vq=eng_databasesinformationsystems

We also applied some of the insights gained from developing theoretical modeling to addressing timely societal issues relating to social media and its impact. One problem we address in [Rizoiu et al., 2018a] was estimating user influence from a set of retweet diffusions, without observing the network in which these diffusions occur. We define a measure of user influence based on the user's active contributions to information diffusions, i.e. their tweets and retweets. Given that Twitter does not expose the retweet structure - it associates all retweets with the original tweet - we model the latent diffusion structure using only tweet time and user features, and we implement a scalable novel approach to estimate influence over all possible unfoldings.

Broader impact The broader impact of our work includes answering three types of questions. The first is explanatory - *how did this become popular?* The second is predictive - *how to identify quality content?* The third is actionable - *how should I construct my promotion schedule to maximize impact?* In the current environment where information is abundant but attention is a scarce resource, understanding popularity dynamics and predicting future popularity will help both content consumers to cope with information overload, and content producers to disseminate information within limited resources.

We are actively exploring opportunities for technical transfer and public outreach. We have engaged with AFSOR Trust and Influence performer Graphika Inc (in New York), Google/YouTube, Netflix and Facebook.

Technical Section

1	Modeling contributions	15
1.1	Hawkes point process for social events	15
1.1.1	Modeling information diffusion	16
1.1.2	The branching factor	16
1.1.3	Estimating the Hawkes process	17
1.1.4	The expected number of future events	18
1.2	The Hawkes Intensity Process (HIP) model for popularity	18
1.2.1	The HIP model	19
1.2.2	Estimating HIP from data	20
1.2.3	Properties of the HIP	21
1.3	Linking epidemic models and Hawkes processes	22
1.3.1	Prerequisites – the SIR model	22
1.3.2	HawkesN: a process in finite population	23
1.3.3	Linking HawkesN and SIR	25
1.4	RNN for Volume and Point Processes	26
1.4.1	Prerequisites – Recurrent Neural Networks	26
1.4.2	Recurrent Point Process (RPP)	27
1.4.3	Volume RNN	28
1.4.4	RNN-MAS for asynchronous streams	28
1.4.5	Model Learning	29
2	Datasets	30
2.1	Datasets	30
2.2	The popularity scale	31
2.3	Time series of attention and engagement	32
3	Popularity measures	33
3.1	Quantifying popularity under promotions	33
3.2	Measures of user engagement and <i>relative engagement</i>	35
3.3	Popularity measures via (deep) neural networks	37
4	Selected results	39
4.1	Results: predicting retweet cascade size	39
4.2	Results: explaining popularity	41
4.3	Results: forecasting popularity	43
4.3.1	A video that will go viral	43
4.3.2	Evaluation of forecast	44
4.3.3	Forecasting results	44
4.4	Result: future popularity under promotion	45
4.4.1	What-if scenarios for two videos	46
4.4.2	Predicting future views	46
4.4.3	Predicting popularity jumps	48
4.5	Result: profiling promotion schedules	49

4.5.1	The economics of time for promotions	49
4.5.2	Promotion and perceived potential	51
4.6	Results: analyzing <i>relative engagement</i>	51
4.6.1	<i>Relative engagement</i> and video quality	51
4.6.2	Temporal dynamics of <i>relative engagement</i>	52
4.7	Result: grassroots users vs. celebrities	53
4.7.1	Understanding unseen influence	53
4.7.2	Average unit promotion score	54
4.7.3	User Influence	55
4.8	Result: explaining popularity unpredictability	56
4.8.1	Epidemic size distributions in SIR	56
4.8.2	Cascade size distribution in HawkesN	58
4.8.3	Explaining popularity unpredictability	58

In the following technical sections of this report, we detail the main technical contributions achieved during this AOARD project. We detail the four categories of technical contributions outlined in the **Project Highlights** section – four modelling contributions are described in Sec. 1; five datasets, on which we report most of the results, are described in Sec. 2; three sets of metrics are described in Sec. 3, and some of the main results are shown in Sec. 4. Full technical details are in our papers [Mishra et al., 2016a, Rizoïu et al., 2017, 2018b,c, Rizoïu and Xie, 2017, Kong et al., 2018, Mishra et al., 2018, Wu et al., 2018].

1 Modeling contributions

Here, we detail our proposed models for the diffusion of information in online media and popularity dynamics under external influence. First, we first review some key concepts of the Hawkes process, and we use this self-exciting point-process to model the word-of-mouth spread of attention (Sec. 1.1). The full construction of the Hawkes-based information diffusion model, together with its use in predicting the total popularity can be found in our paper [Mishra et al., 2016a]. Next, we describe HIP, the model that we propose in [Rizoïu et al., 2017] to explain the observed popularity history from daily volumes when the underlying point process is unobserved (Sec. 1.2). Two key metrics can be derived from the HIP model, the endogenous response and exogenous sensitivity, to quantify the viral potential of a video (Sec. 1.2.3). Next, we show our constructed link between epidemic models and Hawkes point processes (Sec. 1.3) and finally we present a Deep Neural Network solution for popularity prediction (Sec. 1.4).

1.1 Hawkes point process for social events

In self-exciting processes the occurrences of past events make future events more probable. In particular, we model each retweet as a stochastic event, and model three key intuitions in a social network: *magnitude of influence*, tweets by users with many followers tend to get retweeted more; *memory over time*, that most retweeting happens when the content is *fresh* [Wu and Huber-

man, 2007]; and *content quality*. This modeling of information diffusions has been presented and extensively discussed in our paper [Mishra et al., 2016a]

1.1.1 Modeling information diffusion

The marked Hawkes process [Hawkes, 1971] is a well-known self-exciting process that captures these properties. Its event rate is written as:

$$\lambda(t) = \sum_{t_i < t} \phi_{m_i}(t - t_i) , \quad (1)$$

where $\lambda(t)$ is the arrival rate of new events. It depends on each previously occurred event (m_i, t_i) of magnitude m_i at time t_i through the triggering kernel $\phi_m(\tau)$. For the problem of estimating the size of a retweet cascade, there are no exogenously generated events (as in a general Hawkes process [Hawkes and Oakes, 1974]) apart from the initial tweet, which is accounted for with magnitude m_0 at time $t_0 = 0$. The effects of m_i and t_i are separable in our triggering kernel:

$$\phi_m(\tau) = \kappa m^\beta (\tau + c)^{-(1+\theta)} . \quad (2)$$

κ describes the *virality* – or quality – of the tweet content and it scales the subsequent retweet rate; β introduces a warping effect for the user influence and it is related to the observed long-tail distribution of user influence in social networks; and $1 + \theta$ ($\theta > 0$) is the power-law exponent, describing how fast an event is *forgotten*, parameter $c > 0$ is a temporal shift term to keep $\phi_m(\tau)$ bounded when $\tau \simeq 0$; Here κm^β accounts for the magnitude of influence, and the power-law kernel $(\tau + c)^{-(1+\theta)}$ models the memory over time. We approximate user influence m using the number of followers obtained from Twitter API.

Throughout literature, three families of functions have been mainly used to model temporal decay with Hawkes point-processes [Rodriguez et al., 2011]: power-law functions $\phi^p(\tau) = (\tau + c)^{-(1+\theta)}$, used in geophysics [Helmstetter and Sornette, 2002] and social networks [Crane and Sornette, 2008a]; exponential functions $\phi^e(\tau) = e^{-\theta\tau}$, used in financial data [Filimonov and Sornette, 2013]; Reyleigh functions $\phi^r(\tau) = e^{-\frac{1}{2}\theta\tau^2}$, used in epidemiology [Wallinga and Teunis, 2004]. We experimented with all three kernels and found that the power-law kernel consistently outperform the other two. Consequently, we only discuss detailed results from the power law kernel in the rest of this paper.

1.1.2 The branching factor

One of the key quantities that describe a Hawkes process is the branching factor n^* , defined as the expected number of children events directly spawned by an event. Note that the total number of expected events from one initial event is the sum of a geometric series $\sum_{k=1}^{\infty} (n^*)^k$. For $n^* < 1$, the process is in a *subcritical regime*: the total number of retweets is bounded, the event rate $\lambda(t)$ decays to zero over time and the retweet cascade dies out. When $n^* > 1$, the process is in a so-called *supercritical regime* with the total number of retweets being unbounded. We compute the branching factor by integrating over time and event magnitudes.

$$n^* = \int_1^{\infty} \int_0^{\infty} P(m) \phi_m(\tau) d\tau dm . \quad (3)$$

Here we assume m_i are *i.i.d.* samples from a power law distribution of social influence [Kwak et al., 2010]: $P(m) = (\alpha - 1) m^{-\alpha}$. Here α is an exponent which controls the heavy tail of the distribution. We extract the number of followers m for a large sample of users from our dataset described in Sec. 2 and fit it to a power-law distribution following the method detailed in [Clauset et al., 2009]. We obtain $\alpha = 2.016$ and we use it throughout the experiments. Substituting Eq. (2) and $P(m)$ into (3), we obtain the closed-form expression of the branching factor:

$$n^* = \kappa \frac{\alpha - 1}{\alpha - \beta - 1} \frac{1}{\theta c^\theta}, \text{ for } \beta < \alpha - 1 \text{ and } \theta > 0. \quad (4)$$

Not only the branching factor n^* is an intuitive descriptor of the *virality* of a cascade, it is also used for predicting the total popularity in Sec 1.1.4 and as a constraint in model estimation in Sec 1.1.3.

1.1.3 Estimating the Hawkes process

The log-likelihood function. Our marked Hawkes process is completely defined by the four parameters $\{\kappa, \beta, c, \theta\}$. We now describe a maximum-likelihood estimation procedure of the model from a set of observed events, consisting of $\{(m_i, t_i), i = 0, \dots, n\}$ for time $t < T$. The log-likelihood function of a point-process described by Eq. (1) and (2) is written as ([Daley and Vere-Jones, 2008], Ch. 7.2):

$$\begin{aligned} \mathcal{L}(\kappa, \beta, c, \theta) &= \log P(\{(m_i, t_i), i = 1, \dots, n\}) \\ &= \sum_{i=1}^n \log(\lambda(t_i)) - \int_0^T \lambda(\tau) d\tau \\ &\stackrel{\text{cf. (1)\&(2)}}{=} \sum_{i=2}^n \log \kappa + \sum_{i=2}^n \log \left(\sum_{t_j < t_i} \frac{(m_j)^\beta}{(t_i - t_j + c)^{1+\theta}} \right) \\ &\quad - \kappa \sum_{i=1}^n (m_i)^\beta \left[\frac{1}{\theta c^\theta} - \frac{(T + c - t_i)^{-\theta}}{\theta} \right]. \end{aligned} \quad (5)$$

The first two terms in Eq. 5 are from the likelihood computed using the event rate $\lambda(t)$, the last term is a normalizing factor from integrating the event rate over the observation window $[0, T]$.

Finding the maximum-likelihood solution. Eq. 5 is a non-linear objective that we optimize to find the set of parameters that maximizes data likelihood. Directly maximizing this objective is more efficient than expectation-maximization approaches [Ding et al., 2015] that try to estimate the branching structure (of which event triggers which) and that have a quadratic number of latent variables. There are a few natural constraints for each of model parameter, namely: $\theta > 0$, $\kappa > 0$, $c > 0$, and $0 < \beta < \alpha - 1$ for the branching factor to be meaningful (and positive). Furthermore, while the supercritical regimes $n^* > 1$ are mathematically valid, they will not lead to a finite cascade size prediction. We note that in reality no retweet cascade can grow indefinitely, due to finite network sizes. We further incorporate $n^* < 1$ as a non-linear constraint for the maximum likelihood estimation. We use the Ipopt [Wächter and Biegler, 2006], the large-scale interior point solver that handles both non-linear objectives and non-linear constraints, with pre-programmed gradient functions to find the solution.

1.1.4 The expected number of future events

Having observed a retweet cascade until time T and fitted the parameters of the point process, one can simulate a possible continuation of the cascade, until it dies out – assuming $n^* < 1$. In this section, we derive the expected number of events in the cascade, over all possible continuations. We group events based on the generation of their parents (i.e. the preceding event they are triggered by), and estimate the number of events in each generation. We denote as $Generation_1$ the set of simulated events spawned by the observed events. Similarly, $Generation_2$ groups events spawned by events in $Generation_1$, and so on. Let A_i be the expected number of events in $Generation_i$. The expected number of total events in the cascade, N_∞ , is defined as:

$$N_\infty = n + \sum_{i=1}^{\infty} A_i, \quad (6)$$

where n is the number of observed events. For generations after the first one, the best estimate about $A_i, i \geq 2$ is constructed using the average number of children events n^* and the number of events in the previous generation, i.e. $A_i = A_{i-1}n^*$. Assuming A_1 known, we derive:

$$A_i = A_{i-1}n^* = A_{i-2}(n^*)^2 = \dots = A_1(n^*)^{i-1}, i > 1 \quad (7)$$

Therefore, the second term in (6) is the sum of a converging geometric progression (assuming $n^* < 1$):

$$\sum_{i=1}^{\infty} A_i = \frac{A_1}{1 - n^*} \text{ where } n^* < 1 \quad (8)$$

A_1 could also be calculated in a fashion similar to Eq. (7). However, a more precise estimation can be obtained, given that magnitudes of the observed events – parents for events in $Generation_1$ – are known:

$$\begin{aligned} A_1 &= \int_T^\infty \lambda(\tau) d\tau = \int_T^\infty \sum_{t > t_i} \phi_{m_i}(t - t_i) dt \\ &= \kappa \sum_{i=1}^n \frac{m_i^\beta}{\theta (T + c - t_i)^\theta} \end{aligned} \quad (9)$$

We obtain the estimate of the total number of events in the cascade by introducing (9) and (8) into (6):

$$N_\infty = n + \frac{\kappa}{(1 - n^*)} \left(\sum_{i=1}^n \frac{m_i^\beta}{\theta (T + c - t_i)^\theta} \right), n^* < 1 \quad (10)$$

1.2 The Hawkes Intensity Process (HIP) model for popularity

The Hawkes point process, as described in the previous section, faces a few modeling challenges in large-scale applications. In terms of data source, what we often observe is the volume of total attention in a given interval (e.g. daily views on YouTube), rather than the times and properties of individual actions, due to constraints in user privacy and data volume. In terms of computation, full estimation of Hawkes process is quadratic in the number of events. We have also observed that

omitting the non-linear components tend to hurt modeling power [Mishra et al., 2016a]. But the full estimation quickly becomes expensive when the number of events is in the hundred thousands or millions – this is where the most popular videos are (see Sec 2). It is very desirable if one could estimate video popularity with daily data, which is typically a few dozens to a few hundred data points.

1.2.1 The HIP model

To this end, in our work [Rizoiu et al., 2017] we introduced the Hawkes intensity $\xi(t)$, the expectation of the event rate $\lambda(t)$ over event history $\mathcal{H}_t$, consisting of the set of (random) event times and magnitudes up to time t .

Theorem 1.1 HAWKES INTENSITY PROCESS (HIP) *Given a marked Hawkes process described in Equations (1) and (2). Its event history*

$$\mathcal{H}_t = \{(t_1, m_1), \dots, (t_n, m_n)\}_{t_n < t}$$

contains all event times and marks before time t , where each mark m is drawn iid from a power-law distribution $p(m) = (\alpha - 1)m^{-\alpha}$. Define event intensity as the expectation of the event rate over the event history $\xi(t) = \mathbb{E}_{\mathcal{H}_t} [\lambda(t)]$, then $\xi(t)$ follows the following self-consistent integral equation:

$$\xi(t) = \mu s(t) + C \int_0^t \xi(t - \tau)(\tau + c)^{-(1+\theta)} d\tau . \quad (11)$$

Here constant $C = \frac{\kappa(\alpha-1)}{\alpha-\beta-1}$, and κ and β are as in Eq (2).

Intuitively, this expression of event intensity $\xi(t)$ at time t is determined by the external stimulus $s(t)$, and a convolution of its own history with a power-law memory kernel $(\tau + c)^{-(1+\theta)}$. Theorem 1.1 can be intuitively understood by breaking down the expectation into several parts. Note $\mu s(t)$ is non-random and does not change after expectation. The expectation over stochastic history with a random number of events at random times, is equivalent to taking expectations over $N(t)$, the counting process. This in turn decompose into expectations of binary variables dNt that indicates whether or not there is an event, or jump in $N(t)$, in a small interval around t . Note that expectation of the user influence warping term m^β over power-law distribution of mark m has an analytical form, leading to the constant C . The full proof involves careful treatment of the above quantities with stochastic calculus, which we fully detail in our paper and its appendixes [Rizoiu et al., 2017]

Here (μ, θ, C, c) are video-dependent parameters estimated from the popularity history of each video. Here $\alpha > 0$ is the power-law exponent of user influence distribution, estimated as $\alpha = 2.016$ from a large Twitter sample using standard fitting procedures [Clauset et al., 2009]. Compared to existing models of data volume, HIP captures the ongoing interactions of exogenous and endogenous effects, and hence able to explain complex popularity series with multiple rises and falls (as shown in Figure 3). Helmstetter and Sornette [2002] fit the observed event rate after an initial shock, and Crane and Sornette [2008a] produce a curve fit on the long-term approximation of the endogenous decay with no exogenous input. SpikeM [Matsubara et al., 2012] models volumes of events both prior and after a single considered shock, without accounting for external influences.

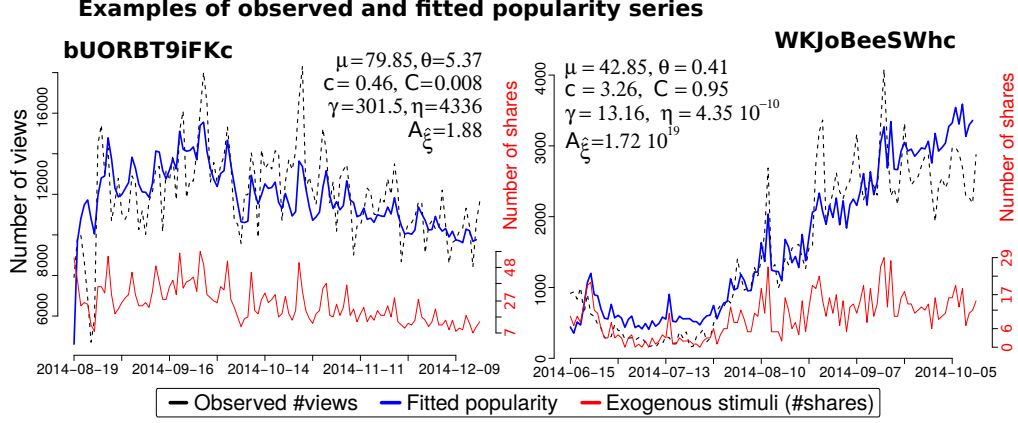


Figure 3: Explaining popularity dynamics using the Hawkes intensity model. (a) Number of shares (red), observed popularity history (black dashed) and popularity as explained by the HIP (blue) on two examples videos: a music video *bUORBT9iFKc* and a News & Politics video *WKJoBeeSWhc*. The multi-phased popularity history cannot be explained by current models such as the ones proposed by Crane and Sornette [2008a], while the HIP tracks the complex dynamics well.

1.2.2 Estimating HIP from data

We discuss key steps for estimating HIP from observed data stream of views and external discussions.

Discretizing over time. We often observe that behavioral statistics are aggregated over fixed and discrete intervals – in the case of YouTube, we have the daily history of views $\bar{\xi}[t]$ and shares $\bar{s}[t]$ for $t = 1, \dots, T$. Expressing HIP (Eq 11) over discrete time gives:

$$\xi[t] = \mu s[t] + C \sum_{\tau=1}^t \xi[t-\tau] (\tau + c)^{-(1+\theta)} . \quad (12)$$

Here we use square brackets to denote discrete time, e.g. $\xi[t]$ and round brackets to denote continuous time, e.g. $\xi(t)$.

The loss function For each video, we find an optimal set of models parameters (μ, θ, C, c) . This is done by minimizing the square error between the observed viewcount series $\bar{\xi}[t]$ and the model $\xi[t]$, $t = 1 : T$. The corresponding optimization problem is as follows:

$$\min_{\mu, \theta, C, c, \gamma, \eta} J = \frac{1}{2} \sum_{t=0}^T (\xi[t] - \bar{\xi}[t])^2 \quad (13)$$

We use L-BFGS [Liu and Nocedal, 1989] with analytical gradients and random restarts to minimize this non-linear loss function.

Three example fits are shown in Fig. 2b and Fig. 3. Visibly, the event intensity model in Eq 11 links the exogenous and endogenous effects of the social system, resulting in a tight fit between the model and the observed popularity history. For video *bUORBT9iFKc* (in Fig. 3 (left)) the memory kernel decays fast ($\theta = 5.37$), and the resulting intensity series tracks the temporal dynamics of the stimuli closely. For video *WKJoBeeSWhc* (in Fig. 3 (right)), the memory kernel decays slowly ($\theta = 0.41$), hence the delayed accumulation of exogenous promotion via the memory kernel results

in an overall rising trend. We can see that only by capturing the non-obvious joint effects from within and outside a social network can a model produce both fine-grained short-term dynamics and accurate long-term trends.

1.2.3 Properties of the HIP

We examine one key property of the HIP, which leads to two important metrics for measuring two distinct aspects of a video’s viral potential.

Exogenous sensitivity μ . As shown in Eq 11, the total attention that a video receives consists of two parts: input from the exogenous stimuli, and the endogenous response corresponding to non-linear effects accumulated through the integral equation. The scaling parameter μ quantifies a video’s sensitivity to external stimuli $s(t)$. When $\mu \rightarrow 0$, external promotion would have no effect; when μ is large, each unit of external promotion leads to a large number of new views.

HIP as an LTI system. We observe that the HIP model in Eq 12 (and also (11)) is a linear time-invariant (LTI) system for $t > 0$ [Oppenheim et al., 1997]. That is to say that if $\xi[t]$ is the event intensity function for input $s[t]$, then (from the same video) the event intensity function for a shifted and scaled version of the input $as[t - t_0]$ is $a\xi[t - t_0]$ for $a > 0, t_0 \geq 0$, i.e., scaled and shifted by the same amount. It is easy to see linearity by multiplying both sides of Eq 11 by the same constant. For time invariance, change of variable and then using the fact that $\xi[t] = 0$ when $t < 0$.

Impulse response function $\hat{\xi}[t]$. One important descriptor of an LTI system is the impulse response function, the response to the unit impulse function $\mathbb{1}[t]$, which takes value 1 when $t = 0$, and 0 otherwise. Define $\hat{\xi}[t]$ as the impulse response of the HIP model, it is the solution to the following self-consistent equation:

$$\hat{\xi}[t] = \mathbb{1}[t] + C \sum_{\tau=0}^T \hat{\xi}[t - \tau](\tau + c)^{-(1+\theta)} d\tau, \quad (14)$$

For each video, $\hat{\xi}[t]$ completely characterizes the endogenous response of the HIP model. The external input $s[t]$ can be expressed as a sum of shifted and scaled impulses $s[t] = \sum_{\tau=0}^T s[\tau]\delta[t - \tau]$. Due to the linearity property, the intensity function $\xi[t]$ can be written as the sum of impulse responses, scaled and shifted by the corresponding external input.

$$\xi[t] = \sum_{\tau=0}^T s[\tau]\hat{\xi}[t - \tau] \quad (15)$$

Fig 2b illustrates this property using a *sliced* and *stacked* popularity graph. The inset shows an example of impulse response. In each discrete time point t' , each white or gray *slice* corresponds to a version of the impulse response shifted by t' and scaled by the external stimuli $s(t')$. Adding all these responses together recovers the overall intensity $\xi(t)$ as in Eq 11 (blue line), which tracks closely the long-term dynamics of the observed popularity (dashed line).

Endogenous response $A_{\hat{\xi}}$. We define the total *endogenous response* generated from a single unit of exogenous excitation, computed as $A_{\hat{\xi}} = \sum_{t=0}^{\infty} \hat{\xi}[t]$. Other HIP-derived quantities, such as scaling parameter C or memory exponent θ could potentially serve to describe video virality. We find, however, that despite being related, the non-linear interactions among HIP parameters render

them inaccurate in explaining popularity, compared to A_ξ . In this work, we compute A_ξ by taking the sum over 10,000 time steps. Together with exogenous sensitivity μ , this is the second key quantity for measuring video virality.

1.3 Linking epidemic models and Hawkes processes

First, we briefly review a few key concepts of the SIR epidemic model and its bivariate process formulation (in Sec. 1.3.1). Next, we propose *HawkesN*, a generalization of the Hawkes process with finite population (in Sec. 1.3.2) and we show the connection between HawkesN and the SIR epidemic model (in Sec. 1.3.3).

1.3.1 Prerequisites – the SIR model

The Susceptible-Infected-Recovered (SIR) model defines three classes of individuals (also known as compartments): those *susceptible* to infection, those currently *infected* (and therefore infectious) and those *recovered* from the infection and no longer infective. SIR models the following process: when a susceptible individual meets an infectious individual, the former becomes infected at the rate β ; infected individuals recover from the infection at a constant rate γ .

Deterministic SIR. In the deterministic SIR, the individuals and their assignment to each of the three compartments are not observed. The temporal dynamics of the sizes of each of the compartments are governed by the following ordinary differential equations [Allen, 2008]:

$$\frac{dS(t)}{dt} = -\beta \frac{S(t)}{N} I(t) \quad (16)$$

$$\frac{dI(t)}{dt} = \beta \frac{S(t)}{N} I(t) - \gamma I(t) \quad (17)$$

$$\frac{dR(t)}{dt} = \gamma I(t). \quad (18)$$

$S(t)$, $I(t)$ and $R(t)$ are deterministic functions, denoting the sizes at time t of the susceptible, infected and recovered populations, respectively; $N = S(t) + I(t) + R(t)$ is the total population size.

There are a number of assumptions made by the SIR model. Firstly, it assumes that the population is homogeneous and individuals meet any other individual uniformly at random. Secondly, it assumes that all rates are constant: the infection rate β and the recovery rate γ . Thirdly, it assumes that the population has no births and no deaths – i.e. N is constant throughout the unfolding of the epidemic. The last assumption holds when the speed of the epidemic outpaces considerably the speed of change in the population – e.g., an average retweet diffusion only lasts minutes, compared to years of expected activity of a user on Twitter.

Stochastic SIR. Several stochastic formulations of the SIR model have been proposed [Allen, 2008], which model the behavior of independent and identically distributed agents. The actions of the agents are described by the same set of holistic rules defined in Eq. (16)-(18) and the same assumptions detailed above [Bobashev et al., 2007].

One such stochastic formulation of SIR is the bivariate point process representation [Yan, 2008], in which two types of events occur: *infection events* and *recovery events*. The j -th infected individual in the SIR process gets infected at time t_j^I and recovers at time t_j^R . Therefore, to each infection

event corresponds a recovery event. S_t , I_t and R_t are discrete random variable taking integer values. They are the stochastic counterparts of $S(t)$, $I(t)$ and $R(t)$ respectively. Eq. (16)-(18) can be written as stochastic differential equations, but for ease of following we will keep referring to Eq. (16)-(18) in the rest of this paper. We define the *time to recovery* (i.e. the time the individual j is infectious) as: $\tau_j = t_j^R - t_j^I$. From Eq. (18) results that the times to recovery are distributed exponentially with parameter γ , and infections last on average $\frac{1}{\gamma}$ units of time.

Let C_t be the counting process of the *infection process* and R_t the counting process of the *recovery process*. Note that $C_t = N - S_t$ is total number of occurred infections (regardless if they are still infectious) and it is distinct from I_t (number of infectious at time t). Let $\mathcal{H}_t$ be the history of the bivariate epidemic process up to time t , i.e. $\mathcal{H}_t = \{t_1^I, t_2^I, \dots, t_1^R, t_2^R, \dots\}$. It can be shown that the rate of *new infections* $\lambda^I(t)$ and the rate of *new recoveries* $\lambda^R(t)$ are:

$$\lambda^I(t) = \beta \frac{S_t}{N} I_t; \quad \lambda^R(t) = \gamma I_t. \quad (19)$$

We sketch the proof for the previous statement. Yan [2008] derives the probability of a new infection at time t given $\mathcal{H}_t$ as:

$$\begin{aligned} \mathbb{P}(C_{t+\delta t} - C_t = 1 | \mathcal{H}_t) &= \beta \frac{S_t}{N} I_t \delta t + o(\delta t) \\ \mathbb{P}(C_{t+\delta t} - C_t > 1 | \mathcal{H}_t) &= 0 \\ \mathbb{P}(C_{t+\delta t} - C_t = 0 | \mathcal{H}_t) &= 1 - \beta \frac{S_t}{N} I_t \delta t + o(\delta t), \end{aligned}$$

It follows that the new infections process is a temporal point process of intensity $\beta \frac{S_t}{N} I_t$; $\lambda^R(t)$ is derived similarly.

Fig. 4a illustrates an SIR realization as a bivariate point process: five infection events occur at times $t_1^I, \dots, t_5^I$ (shown in red); five recovery events occur at $t_1^R, \dots, t_5^R$ (shown in blue). The middle panel of Fig. 4a shows the size of the infectious population I_t over time. Each new infection event increments I_t , and each new recovery decreases I_t by one. The bottom panel of Fig. 4a shows the corresponding new infection and new recovery rates. Initially, $\lambda^I(t)$ is significantly higher than $\lambda^R(t)$. As the number of susceptible individuals S_t gets depleted, the term $\frac{S_t}{N}$ in Eq. (19) inhibits $\lambda^I(t)$ which becomes zero after the fifth infection ($S_t = 0, t \geq t_5^I$). The new recovery rate also becomes zero after the last infected individual recovers ($I_t = 0, t \geq t_5^R$).

The connection between deterministic and stochastic SIR is that the mean behavior of the stochastic process asymptotically approaches that of the deterministic process [Allen, 2008, Yan, 2008]. The connection between the deterministic and the stochastic population sizes is $S(t) = \mathbb{E}_{\mathcal{H}_t} [S_t]$, $I(t) = \mathbb{E}_{\mathcal{H}_t} [I_t]$ and $R(t) = \mathbb{E}_{\mathcal{H}_t} [R_t]$ [Allen, 2008]. Our own results simulating the two variants are presented in an online supplement of the paper [Rizoiu et al., 2018c].

As will be elaborated in Sec. 1.3.3, the bivariate point process SIR formulation provides the link to the Hawkes point processes.

1.3.2 HawkesN: a process in finite population

We generalize the Hawkes model [Hawkes, 1971] to account for finite population sizes. Intuitively cascades not only follow self-exciting word of mouth diffusions, but they are also limited by the size of the relevant community. The effect of introducing the finite population size N is that the

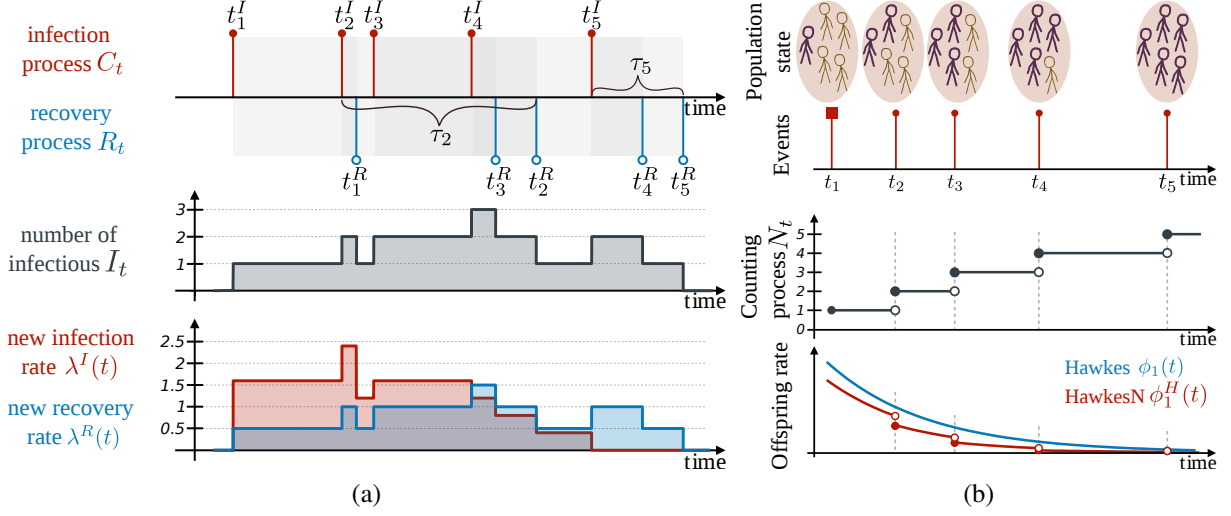


Figure 4: **(a)** An illustration of SIR as a bivariate point process: the infection process and the recovery process. (*top panel*) The j^{th} individual gets infected at time t_j^I and recovers at t_j^R . The *time to recovery* $\tau_j = t_j^R - t_j^I$ is the period the individual stays infectious. (*middle panel*) The size of the infectious population $I(t)$ over time. (*lower panel*) The infection rate $\lambda^I(t)$ and the recovery rate $\lambda^R(t)$ for the SIR parameters: $N = 5$, $\beta = 2$, $\gamma = 0.5$. **(b)** An example diffusion illustrating the finite population effects in self-exciting process. (*top panel*) An event refers to the j^{th} user taking an action at time t_j (e.g. posting a tweet). The state of the user population is shown at each time t_j : **purple users** have performed the past observed actions; **orange users** are yet to perform any action. (*middle panel*) The counting process N_t increases by one with each event; (*lower panel*) The *offspring rate* $\phi_1(t)$ – the rate of events generated by this first event at time t_1 , modeled by Hawkes and by HawkesN (denoted by $\phi_1^H(t)$).

event rate at time t is modulated by the available population. To the best of our knowledge, no prior work on modeling social processes using Hawkes models had accounted for a finite underlying population.

The event rate function in HawkesN is defined as:

$$\lambda^H(t) = \left(1 - \frac{N_t}{N}\right) \left[\mu + \sum_{t_j < t} \phi(t - t_j) \right], \quad (20)$$

where $\phi(t - t_j)$ can be the same kernel function used with Hawkes, and N_t is the counting process associated with the point process. Both $\lambda^H(t)$ and N_t are right-continuous functions. The term $1 - \frac{N_t}{N}$ scales the event rate at time t with the proportion of the events which can still occur after time t . When $t = 0$, we have $\lambda^H(t) = \mu$. When $N_t = N$, we have $\lambda^H(t) = 0$, i.e., there will be no more new events when the pool of users who can act is exhausted. When $N \rightarrow \infty$, Eq. (20) simplifies to Eq. (1). In other words, the Hawkes process is a special case of HawkesN with infinite population.

Fig. 4b illustrates the HawkesN process for an information diffusion in a population of five users ($N = 5$). Each user takes an action at most once, represented as event time $t_j, j = 1..5$. The corresponding counting process N_t is shown in the middle plot. Events $t_2..t_5$ are considered to have been triggered by event t_1 . The bottom panel compares the *offspring rates* – the rate of events

generated by the first event at t_1 – for Hawkes (denoted as $\phi_1(t)$) and HawkesN (denoted as $\phi_1^H(t)$). In HawkesN, the population modulates the event rate by decreasing it after each new event and the event rate becomes zero after t_5 . The Hawkes process does not take into account the population size, i.e. it will have $\phi(t) > 0$ in Eq. 1 even after t_5 .

We use the exponential kernel function for HawkesN:

$$\phi(\tau) = \kappa\theta e^{-\theta\tau} \quad (21)$$

κ is a scaling factor and θ is the parameter of the exponential function which models the decay of *social memory*. The exponential kernel is a common choice in literature [Mishra et al., 2016a, Zarezade et al., 2017, Zhao et al., 2015, Shen et al., 2014, Bao et al., 2015, Ding et al., 2015, Gao et al., 2015]. Other kernels have been used with Hawkes, including power-law functions [Helmstetter and Sornette, 2002, Crane and Sornette, 2008a, Mishra et al., 2016a, Kobayashi and Lambiotte, 2016] and Rayleigh functions [Wallinga and Teunis, 2004]. Using HawkesN with non-exponential kernel functions is left for future work.

1.3.3 Linking HawkesN and SIR

We now present our main result, Theorem 1.2, which links stochastic SIR and the HawkesN process.

Intuition. When modeling information diffusion, both SIR and HawkesN model the same phenomenon: users come into contact with the diffused content, which they further broadcast to other users. Each new broadcast is modeled as a new event in HawkesN, and as a new infection in SIR. The key to linking HawkesN and SIR models is the conceptual similarity between an event in HawkesN and a new infection in SIR. In HawkesN, past events generate new events at the rate $\phi(t)$, which is exponentially time-decaying in Eq. (21). In SIR, an infectious individual j infects susceptible individuals at a rate of $\frac{\beta S_t}{N}$ during the time it is infectious τ_j , which is exponentially distributed with parameter γ (discussed in Sec. 1.3.1).

Theorem 1.2 *Suppose the new infections in a stochastic SIR process of parameters $\{\beta, \gamma, N\}$ follow a point process of intensity $\lambda^I(t)$. Suppose also the events in a HawkesN process with parameters $\{\mu, \kappa, \theta, N\}$ have the intensity $\lambda^H(t)$ (Eq 20). Let $\mathcal{T} = \{\tau_1, \tau_2, \dots\}$ be the set of the times to recovery of the infected individuals in SIR. The expectation of $\lambda^I(t)$ over $\mathcal{T}$ is equal $\lambda^H(t)$:*

$$\mathbb{E}_{\mathcal{T}}[\lambda^I(t)] = \lambda^H(t),$$

when $\mu = 0, \beta = \kappa\theta, \gamma = \theta$.

Note that both $\mathbb{E}_{\mathcal{T}}[\lambda^I(t)]$ and $\lambda^H(t)$ are random functions, as they depend on the random infection times t_j^I (for SIR) and the random event times t_j (for HawkesN). The expectation only removes the randomness from the recovery times t_j^R in SIR.

The rest of this section proves this theorem.

The expected new infection rate. We express S_t and I_t in Eq. (19) using indicator functions of the infection event times and the times to recovery:

$$\begin{aligned} S_t &= N - C_t = N - \sum_{j \geq 1} \mathbb{1}(t_j^I < t) \\ I_t &= C_t - R_t = \sum_{j \geq 1} \mathbb{1}(t_j^I < t, t_j^R > t) = \sum_{t_j^I < t} \mathbb{1}(t_j^I + \tau_j > t). \end{aligned} \quad (22)$$

We examine a point process consisting only of the infection events $\{t_j^I\}$. The event rate in this process is obtained by marginalizing out times of recovery:

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}} [\lambda^I(t)] &\stackrel{(19),(22)}{=} \mathbb{E}_{\mathcal{T}} \left[\beta \frac{S_t}{N} \sum_{t_j^I < t} \mathbb{1}(t_j^I + \tau_j > t) \right] \\
&= \sum_{t_j^I < t} \mathbb{E}_{\mathcal{T}} \left[\beta \frac{S_t}{N} \mathbb{1}(t_j^I + \tau_j > t) \right] \\
&= \sum_{t_j^I < t} \int_0^\infty \beta \frac{S_t}{N} \mathbb{1}(t_j^I + \zeta > t) r(\zeta) d\zeta \\
&= \sum_{t_j^I < t} \beta \frac{S_t}{N} \int_{t-t_j^I}^\infty r(\zeta) d\zeta,
\end{aligned} \tag{23}$$

where $r(\zeta)$ is the exponential probability distribution function for the time to recovery (cf. Sec. 1.3.1). Knowing $S_t = N - C_t$, we obtain:

$$\mathbb{E}_{\mathcal{T}} [\lambda^I(t)] = \left(1 - \frac{C_t}{N}\right) \sum_{t_j^I < t} \beta e^{-\gamma(t-t_j^I)}. \tag{24}$$

We can see that Eq. (20) (with the exponential kernel in Eq. (21)) and Eq. (24) are identical when $N_t = C_t$ and under the parameter equivalence in Theorem 1.2. That is to say, the new infection point process in an SIR model and the HawkesN point-process with no background event rate are described by the same conditional intensity. This completes the proof of Theorem 1.2. We also demonstrate the equivalence empirically, through simulation and subsequent parameter fitting, in the online supplement of the paper [Rizoiu et al., 2018c].

1.4 RNN for Volume and Point Processes

In this section, we first introduce background notions relating to Recurrent Neural Networks (Sec. 1.4.1). Next, we introduce a recurrent model for modeling event series in Section 1.4.2, followed by a recurrent model for a volume series in Section 1.4.3. Finally we present a joint model for predicting attention volumes in Section 1.4.4. Our notations for all the model are summarized in Table 1.

1.4.1 Prerequisites – Recurrent Neural Networks

Recurrent Neural Networks (RNN) [Elman, 1990] are common sequence models where the same feed forward structure is replicated at each time step. They have additional connections from the output of previous the time step to the input of the current time step – therefore creating a recurrent structure. Their hidden state vector $\mathbf{h}_t$ can be defined recursively as:

$$\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1})$$

where f is the feed forward network, $\mathbf{x}_t$ is the current input, $\mathbf{h}_{t-1}$ is the output from previous time step. Long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997, Graves, 2013] units

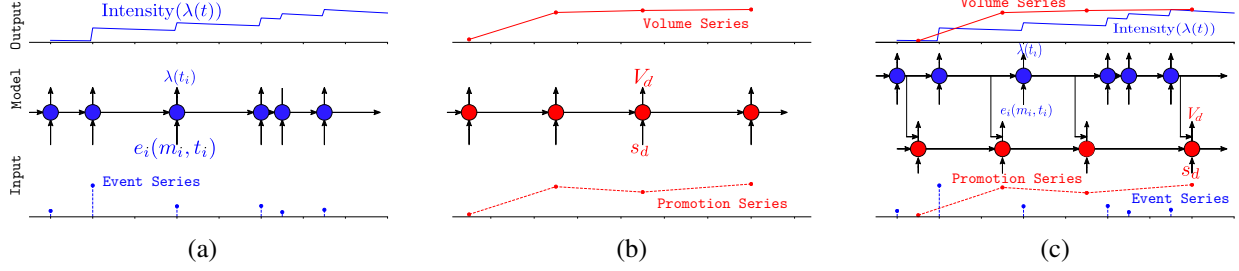


Figure 5: Our LSTM based models, (a) RPP, (b) VoIRNN-S/TS and (c) RNN-MAS. Each individual figure has three parts with model realization shown in middle along with output on top and input at bottom.

are essentially recurrent networks with additional gated structure, defined as:

$$\begin{aligned}
 \mathbf{i}_t &= \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{V}_i \mathbf{c}_{t-1} + \mathbf{b}_i) \\
 \mathbf{f}_t &= \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{V}_f \mathbf{c}_{t-1} + \mathbf{b}_f) \\
 \mathbf{c}_t &= \mathbf{f}_t \mathbf{c}_{t-1} \odot \tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \\
 \mathbf{o}_t &= \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{V}_o \mathbf{c}_t + \mathbf{b}_o) \\
 \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t)
 \end{aligned} \tag{25}$$

where $\mathbf{x}_t$ is input at time t , σ is the logistic sigmoid function and $\odot$ denotes element-wise multiplication. The notations $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{c}_t$, $\mathbf{o}_t$ and $\mathbf{h}_t$ stand for the input, forget, cell-state, output and hidden state at time t . We use the following short-hand notation for the LSTM set of equations in Equation 25:

$$(\mathbf{h}_t, \mathbf{c}_t) = LSTM(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}) \tag{26}$$

LSTM and its variants have been successfully used for modeling time series and predicting sequence, due to their ability to capture the effects of past data in their hidden state [Hochreiter and Schmidhuber, 1997, Graves, 2013, Chung et al., 2014]. We build on this intuition to construct based on LSTM the models proposed in Section 1.4.2.

1.4.2 Recurrent Point Process (RPP)

We present a model to frame events data as point processes in Fig. 5a, where the influence of history on future events is learned as parameters of LSTM unit.

The equation for event rate at time t for the model is written as:

$$\lambda(t) = \exp(W_o^P \mathbf{h}_t^P + \alpha(t - t_i)) \tag{27}$$

such that $\mathbf{h}_t$ is :

$$\mathbf{h}_t^P = LSTM((m_i, t_i), \mathbf{h}_{t_i}^P, \mathbf{c}_{t_i}) \tag{28}$$

where t_i is the time for last event before time t , m_i is the magnitude of last event. In point process literature magnitude of an event is generally referred to as mark of an event, where mark represent attributes of the event, e.g. magnitude of the earthquake, or the type of a trade - bond or stock.

	Symbol	Quantity
Event	$e_i = (m_i, t_i)$	i^{th} event at time t_i of magnitude m_i
Data	$\lambda(t_i)$	Intensity at i^{th} event
Volume	s_d	Exogenous stimulus at fixed interval d
Data	V_d	Attention Volume at fixed interval d

Table 1: Summary of notations for RNN models

In our work for modeling tweets we consider mark as the magnitude of user tweeting. $\mathbf{h}_{t_i}$ and $\mathbf{c}_{t_i}$ denotes hidden and cell state respectively at time t_i .

Our model is similar to the one proposed by [Du et al., 2016]. In their work they assume discrete values of marks as they model different event types rather than magnitude. We generalize their model to work for continuous marks by modeling mark generation separately from history, as done in [Rizoiu et al., 2017, Mishra et al., 2016a].

1.4.3 Volume RNN

RNN can also model volume streams. HIP can be seen as generalization of autoregressive models, with a history size as long as the sequence itself. Intuitively in HIP, the current value of volume intensity can be seen as a convolution of a memory kernel with its past values. Hence we formulate a LSTM model in Fig. 5b, for predicting attention volume series as follows, using Equation (26):

$$V_d = W_o^{Vol} \mathbf{h}_d \quad (29)$$

such that,

$$\mathbf{h}_d = LSTM(\mathbf{s}_d, \mathbf{h}_{d-1}, \mathbf{c}_{d-1}) \quad (30)$$

where V_d is the attention volume at time d , $\mathbf{h}_d$ is the hidden state time d and input to the system is $\mathbf{s}_d$ at time d , where s_d is the exogenous stimuli for the attention volume at time d . For volume RNN model we have time points as discrete values, spread over fixed time intervals (e.g. daily, hourly).

1.4.4 RNN-MAS for asynchronous streams

In real world scenarios, we have multiple streams of data promoting a single attention series. For example, being tweeted and being shared both lead to more views on a YouTube video. In this section we present a combined model to take into account the volume promotions and individual events for predicting an attention volume series.

RNN-MAS model is shown in the Fig. 5c. It has two components: a volume RNN model as shown in Fig. 5b predicting the desired attention volume series at fixed time intervals by taking into account the promotional volume series; a second component as RPP model as shown in Fig. 5a, responsible for modeling the individual promotion series. Intuitively we would like RPP to modulate the response of volume RNN, to make better predictions for attention series. We achieve it by combining the hidden states for individual models just before making the prediction for attention series.

We combine individual models as follows:

$$V_d = W_o^{MAS} [\mathbf{h}_d, \mathbf{h}_d^P] \quad (31)$$

where, $\mathbf{h}_d$ and $\mathbf{h}_d^P$ are the hidden state for volume RNN and RPP on day d , calculated as per Equation 30 and Equation 28 respectively, shown with arrows coming from RPP model towards the volume RNN model at regular intervals in Fig. 5c.

We note another plausible setup for joint modeling is attention volumes affecting the event rate of RPP. As our main concern is predicting popularity under promotion we don't utilize this setting.

Though there has been some recent work [Wang et al., 2017] and [Xiao et al., 2017] in combining asynchronous streams, a major limitation in them is that model aggregates exogenous individual events at fixed intervals rather than considering a separate exogenous volume stream.

1.4.5 Model Learning

RPP: For fitting our model to a sequence of events $S = \{e_i\}$ where, e_i is the i^{th} event such that m_i and t_i stands for magnitude and time of the i^{th} event, we maximize the log-likelihood of the observed sequence. Using Equation 27 and Equation 28, we can write log-likelihood for sequence as:

$$\begin{aligned} LL(S) &= \sum_{i=1}^{|S|} \log(\lambda(t_i)) - \int_0^{t_{|S|}} \lambda(\tau) d\tau \\ \implies LL(S) &= \sum_{i=1}^{|S|} \left[W_o^P \mathbf{h}_{t_i}^P + c\tau_i + \frac{1}{c} \exp(W_o^P \mathbf{h}_{t_i}^P) \right. \\ &\quad \left. - \frac{1}{c} \exp(W_o^P \mathbf{h}_{t_i}^P + c\tau_i) \right] \end{aligned} \quad (32)$$

For predicting the time for the next event in sequence we can utilize the relationship between conditional density function of time, $f(t)$, and conditional intensity function (event rate), $\lambda(t)$ [Daley and Vere-Jones, 2008]:

$$f(t) = \lambda(t) \exp\left(-\int_{t_i}^t \lambda(\tau) d\tau\right) \quad (33)$$

In order to predict the next time step, $\hat{t}_{i+1}$, where t_i is event time for the last observed event, we take the expectation in Equation 33 as follows:

$$\implies \hat{t}_{i+1} = \int_{t_i}^{\infty} t \cdot f(t) dt \quad (34)$$

Volume RNN: For training volume models, we use *RMSE* as the loss function and each fixed interval data is considered as separate time steps of the given variables. This can be seen as modeling a time-series regression with LSTM.

RNN-MAS: For training the joint model, we first train both RPP and volume RNN models independently. After fitting both models, we calculate the hidden state from RPP model for day d using

Equation 28 and hidden state for volume RNN for day d is calculated using Equation 30. Now the combined model is trained using *RMSE* for the attention volume series.

For all models, we jointly learn network parameters over 500 random sequences from ACTIVETWEETS described in Section 2.1. Now for individual fittings we use the jointly learned parameters as the initialization point. We observe faster convergence and better prediction results when compared to running with random initializations. For more details, see our paper [Mishra et al., 2018].

2 Datasets

We describe, in Sec. 2.1, the five datasets on which we report results in this report. We also describe three daily series available with each of the video datasets (Sec. 2.3): shares, views and watch time.

2.1 Datasets

THE NEWS dataset. We constructed a domain specific dataset of tweets dealing with news articles, by collecting data using the *free* Twitter Streaming API², over a period of four months from April 2015 to July 2015. We track the official twitter handles of ten popular news outlets, namely New York Times, Associated Press, CNN, Washington Post, Reuters, Yahoo News, The Guardian, BBC, Huffington Post and Google News. We also capture all tweets containing either i) user mentions of the above users or ii) a commonly used shortened url towards any of the ten news websites. This leads to a collection of 49,735,271 tweets. Most of the cascades in the NEWS dataset are of length one or two and, for compatibility with the TWEET-1MO, we present the statistics for cascades of length over 50: 110,954 cascades, mean cascade length is 158 and median length is 90. **TWEETED VIDEOS.** Similar to our previous work [Rizoiu et al., 2017], [Yu et al., 2014], we use Twitter to sample a large collection of YouTube videos. We use the Twitter Streaming API to collect tweets, by tracking the expression "YOUTUBE" OR ("YOUTU" AND "BE"). This covers textual mentions of YouTube, YouTube links and YouTube's URL shortener (youtu.be). We collected 244 million tweets over a two month period, between July 1st, 2016 and August 31st, 2016. In each tweet, we search in the `extended_urls` field provided by the Twitter API to extract the associated YouTube video id. This results in 36 million unique video IDs and over 206 million associated tweets. For each video, we extract its metadata and three attention-related series, as described in Sec. 2.3. A non-trivial number of the videos (45.82% of all videos) have either been deleted or their statistics are private and cannot be retrieved. This leaves a total of 19.5 million usable videos, which we further pre-filter on recency and level of activity. We remove any video that has been uploaded prior to the two month crawling period to avoid introducing bias towards older videos (since being twitted a while after being posted might be indicative of higher user engagement). We also filter out videos that received less than 100 views during the first 30 days after posting, so that we can construct accurate estimates of user engagement. This results in the TWEETED VIDEOS dataset, containing 5,327,689 videos and 1,256,646 channels. Table 2 shows a detailed breakdown per category, in addition to summarizing all of the constructed datasets.

QUALITY VIDEOS. We collect three QUALITY VIDEOS datasets that meets the standard of domain experts. Two of the datasets are on Music, one on News.

²<https://dev.twitter.com/streaming/overview>

Dataset	#Videos	#Channels
TWEETED VIDEOS	5,327,689	1,256,646
VEVO VIDEOS	72,488	14,342
BILLBOARD VIDEOS	63	47
TOP NEWS VIDEOS	29,732	91
ACTIVETWEETS	13,738	6,970

Table 2: Overview of the five used datasets.

- **VEVO VIDEOS.** Vevo is a multinational video hosting service which syndicates on YouTube licensed music clips from three major record companies (Universal Music Group, Sony Music Entertainment and Warner Music Group) [vev, 2017]. VEVO artists usually come from a professional music background, and their music videos are professionally produced and edited. We consider VEVO VIDEOS to be of higher quality than the average Music videos in the TWEETED VIDEOS dataset. The dataset contains all videos posted by verified Vevo artists on the YouTube platform, as of February 2017. We construct the list of all the YouTube channels which contain in their name the keyword “Vevo”, and which had at least one video tweet during the period May 2014 to February 2017. We filter out non-authentic channels by preserving only those with a “verified” YouTube status badge. We obtain 14,342 official Vevo artists channels, having posted 72,488 music clips.

- **BILLBOARD VIDEOS.** Billboards act as canonical rankings in the music industry, tracking music sales, radio airtime and other popularity metrics into a yearly *Hot 100* music chart. The songs that appear in this chart are usually perceived as having vast success and as being of high quality. We collected 63 videos from 47 artists based on the 2016 Billboard Hot 100 chart³.

- **TOP NEWS VIDEOS** is a set of quality News videos, which features the top 100 most viewed News channels, as reported by an external ranking source⁴. Popular news talk shows (such as “The Young Turks”), as well as traditional news broadcasting companies like “CNN” are part of this list. For each channel, we obtain up to their last 500 videos published before Aug 31st, 2016. The dataset contains 91 YouTube channels and 29,732 videos.

ACTIVETWEETS DATASET was collected in 2014, from all tweets that link to a YouTube video, similarly to the TWEETED VIDEOS dataset. It contains 13,738 videos, which were uploaded between 2014-05-29 and 2014-08-09, which were tweeted at least 100 times between 2014-05-29 to 2014-12-26, and were viewed at least 100 times and shared 100 times (through the Youtube share mechanism). Each video has recorded its metadata (title, author, upload date, category) and the daily views, shares and tweets series. All series are at least 120 days long. We augment each video with corresponding individual tweets data, emitted during same period. There are in total 30.2 million tweets in the dataset. On average each video is tweeted 2151 times, median video was tweeted 327 times.

2.2 The popularity scale

Popularity scale is one way to visualize the non-linear relationship between a video’s popularity rank (denoted in percentiles) and the number of views it receives [Rizoiu et al., 2017]. This scale

³<http://www.billboard.com/charts/year-end/2016/hot-100-songs>

⁴<https://vidstatsx.com/youtube-top-100-most-viewed-news-politics>

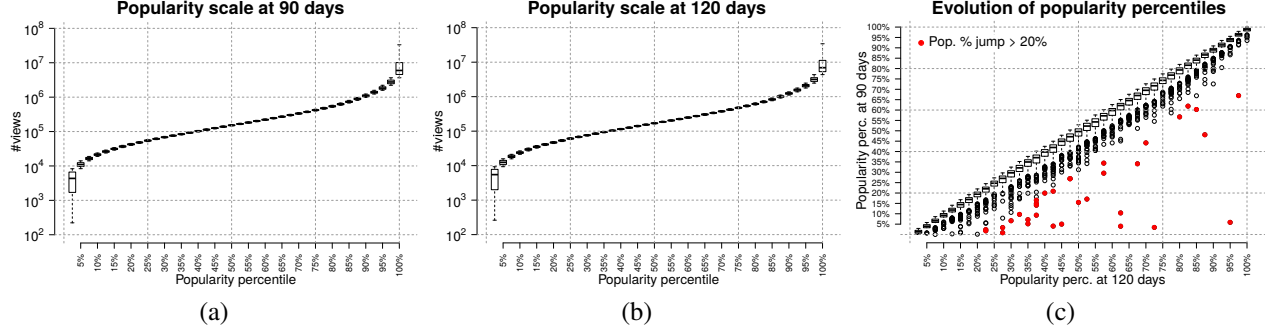


Figure 6: Popularity scale on the ACTIVETWEETS dataset at 90 days (a) and 120 days (b), each box in the box plot contains 2.5%, or 343 videos. (c) Change in popularity between 90 days and 120 days. Outliers are videos whose popularity jumped. The outliers, whose popularity increased by at least 20%, are highlighted in red.

is estimated from a video collection by sorting videos by their number of views, at a given age. For compactness we group all videos into 40 equal-sized bins, each containing 343 videos in the ACTIVETWEETS set. Fig. 6a and 6b plot the boxplots of views in each bin, at 90 and 120 days of age, respectively. Fig. 6c explores the change of popularity percentile between day 90 and 120, denoted as $\Psi_{90}(\cdot)$ and $\Psi_{120}(\cdot)$, respectively. $\Psi_T(v) : \mathbb{R} \rightarrow [0, 1]$ is a function that returns the popularity percentile corresponding to v views, on the popularity scale constructed at age T (i.e. T days after video upload). Most videos stay in or near the same percentile bin (i.e. the high density around the diagonal). The outliers are videos whose popularity “jumped” between day 90 and 120. There are no outliers in the upper-left part of the graphic, as the number of views cannot decrease over time. A video can fall into a lower popularity bin if it accumulates views more slowly than the other videos in his original bin. However, the popularity percentile can increase considerably – Fig. 6c highlights the 33 videos whose popularity jumped by more than 20%. Note that one effect of the non-linear mapping from views to rank is that similar amounts of newly acquired views can have very different impacts on rank: 40,000 additional views between day 90 and 120 may increase a video’s popularity by 15% when starting from lower 17.5% percentile, increase by 2.5% when starting from middle 50% percentile or even decrease by 2.5% when starting from upper 90% percentile – this is because in order to continue to rank in the top 10% at 120 days, a video requires an additional 144,740 views, more than the 40,000 being considered here.

2.3 Time series of attention and engagement

For each video, we use the YouTube Data API to obtain its metadata information – video id, title, description, publish time, category, duration, definition, channel id, channel name and associated freebase topic ids [Figueiredo et al., 2014] (which we resolve to entity names using a Freebase data dump). Next, we use the YTCrawl tool [Yu et al., 2015] to extract three fine-grained traces of video attention dynamics: for each video we record its daily watch time, the daily volume of views and the daily volume of shares it receives. Throughout this report, we denote as $x_v[i]$ and $s[i]$ the number of views and shares, respectively, that a video has received on the i^{th} day after upload. Similarly, $x_w[i]$ is the amount of time the video has been watched during the i^{th} day. Each of the attention series is observed for at least four months, i.e., $i = 1, 2, \dots, 120$. Most prior work on video

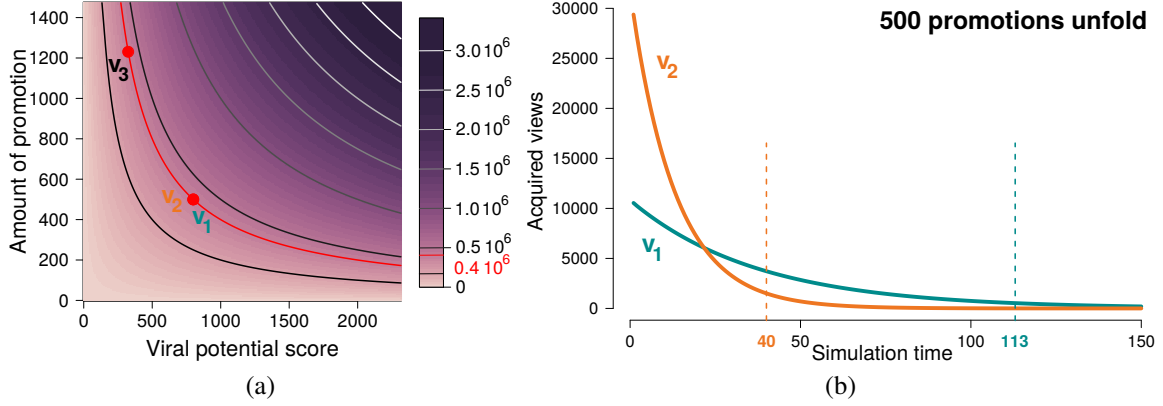


Figure 7: The relation between viral potential, volume of promotions and gained attention. (a) 3 videos placed on the equal-attention hyperbola $\nu S = 400,000$: v_1 (an Entertainment video, id 5TQAnGjyN9A), v_2 (Pets & Animals, id 00ATf2HR-FA) and v_3 (Music, id bx8O62Tuzno). (b) 400,000 views are generated by 500 promotions (for v_1 and v_2) and by 1231 promotions (for v_3), placed at $t = 0$. The corresponding maturity time is shown for each video.

popularity dynamics [Cha et al., 2007, Szabo and Huberman, 2010b, Figueiredo et al., 2016] record and study only video view counts. To the best of our knowledge, this is the first work to perform large-scale measurements of the dynamics of watch time, and their implications in understanding user engagement.

3 Popularity measures

In this section, we detail three sets of measures constructed during this project. First, in Sec. 3.1, we propose two new measures to quantify a video’s total *viral potential* and the time it takes to (almost) reach such potential. We reveal one novel insight about the effect of promotions – that timing does not matter in infinite time. Next, in Sec. 3.2, we study the interplay between viewcounts, watch time, watch percentage and video duration, and we propose the relative engagement, a novel metric to quantify YouTube user engagement. Finally, in Sec. 3.3 we propose two metrics based on RNN-MAS to quantify average response to unit promotion, and the relative influence among users of different *fame*.

3.1 Quantifying popularity under promotions

In this section, we propose two new measures to quantify a video’s total *viral potential* and the time it takes to (almost) reach such potential. We reveal one novel insight about the effect of promotions – that timing does not matter in infinite time. Finally, we show how to more accurately estimate attention resulted from promotions, in finite time.

Viral potential ν is designed to capture the total amount of attention generated by one unit of promotion. It is defined as the integral of the impulse response $\hat{\xi}(\tau)$ over infinite time, namely:

$$\nu = \int_0^\infty \hat{\xi}(\tau) d\tau . \quad (35)$$

ν exists and is finite when the branching factor of the HIP $\frac{C}{\theta c^\theta} < 1$. Throughout this paper, we use estimates of ν by simulating the impulse response series over 10,000 timesteps and numerically integrating it. ν can be used to compare how promotable videos are relative to each other. A corollary of HIP's LTI property shows a way of using ν to estimate the return of promotion campaigns.

Corollary 3.1 *Any promotion series $s(t), t \in [0, T]$ so that $\int_0^T s(\tau) d\tau = S$, applied to a HIP model with the same parameter will generate νS views, given infinite time.*

That is to say that the total amount of attention driven by a given amount of external promotion is fixed, no matter when the promotions happen. This is easy to see with the sliced graph [Rizoiu et al., 2017], as moving one unit of promotion from any time t to another time $t + \Delta$ results in moving one unit of impulse response by Δ . This does not change the total amount of resulting attention the when $t \rightarrow \infty$.

Maturity time t^* . Having measured the total impact of one unit of promotion, we would also like to quantify how fast a given video achieves such effects. Inspired by the definition of half-life in nuclear physics, we introduce *maturity time* t^* , defined as the time it takes to achieve a significant fraction ω of the total popularity of a unit impulse:

$$t^* = \min \left\{ t \geq 0 \left| \int_0^t \hat{\xi}(t) dt \geq \omega \nu \right. \right\} . \quad (36)$$

t^* is finite when ν exists and it captures the speed of the decay of $\hat{\xi}(t)$. In this work, we use $\omega = 0.95$, and numerically estimate t^* by performing a linear search in time: we sum the increasingly longer prefixes of $\hat{\xi}(t)$ and we stop when the obtained sum is greater than $\omega \nu$. The timestep where we stop serves as t^* .

The virality and timing of example videos. We illustrate in Fig. 7 the relation between viral potential, promotion and gained view as a result of promotion. From Corollary 3.1 follows that the hyperbolas $\nu S = \text{const}$ define equal-attention lines – i.e. all combinations of potential and promotion volume that amount to the same effect –, which can be used to study videos comparatively. Fig. 7a illustrates this with three example videos, which lay on the same hyperbola corresponding to 400,000 target views. Videos v_1 and v_2 have the same viral potential of $\nu \sim 800$. They would each require 500 units of promotions to achieve the target. v_3 on the other hand, has $\nu \sim 325$ (mainly due to its low exogenous sensitivity $\mu_3 = 5.76$, compared to $\mu_1 = 21.09$ and $\mu_2 = 58.76$), and would require 1,231 units of promotion to attain the same goal. Note that despite having a similar ν , videos v_1 has a much longer maturity time $t_1^* = 113$ days than v_2 , with $t_2^* = 40$ days. This is illustrated in Fig. 7b, which shows generation of the 400,000 views over time, for the three videos.

Estimating attention in finite time. Many applications are time-critical, and it is desirable to estimate the expected attention in finite time. Given a promotion schedule $s(t), t > 0$, and the beginning and the ending of the time period t_b and t_e , respectively, we derive from Eq. (14) the expected attention in the time interval $[t_b, t_e]$ as:

$$\begin{aligned} \Xi[t_b, t_e] &= \Xi^{old}[t_b, t_e] + \Xi^{new}[t_b, t_e] , \text{ where} \\ \Xi^{old}[t_b, t_e] &= \int_0^{t_b} s(t) \int_{t_b}^{t_e} \hat{\xi}(\tau - t) d\tau dt \\ \Xi^{new}[t_b, t_e] &= \int_{t_b}^{t_e} s(t) \int_0^{t_e - t} \hat{\xi}(\tau) d\tau dt . \end{aligned} \quad (37)$$

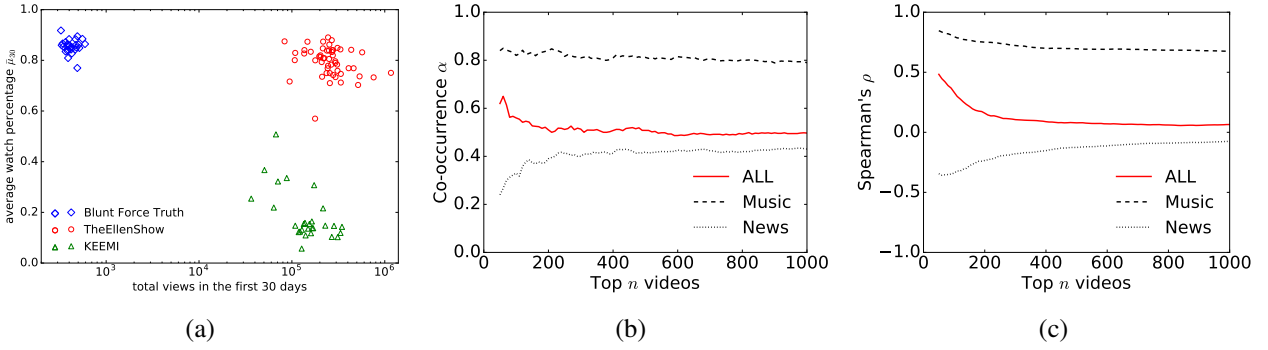


Figure 8: Divergence results from view count and watch time. v_1 : *TkC2bJJUNYs*, entertainment, 6M12S; v_2 : *xw6tDtIhFwQ*, entertainment, 5M37S. (a) Agreement between the “top n ” videos in TWEETED VIDEOS by viewcounts and by watch time, measured with intersection size (b) and Spearman’s ρ (c).

$\Xi^{old}[t_b, t_e]$ corresponds to the views generated by the unfolding of the *old promotions* that occurred before the studied interval. $\Xi^{new}[t_b, t_e]$ corresponds to the incomplete realization of the *current* promotions that occur during $[t_b, t_e]$.

3.2 Measures of user engagement and *relative engagement*

In this section, we study the interplay between viewcounts, watch time, watch percentage and video duration, and we propose the *relative engagement*, a novel metric to quantify YouTube user engagement. One key insight is summarized in the *engagement map*, that calibrates watch behavior for videos of different lengths.

Comparing viewcounts and watch time Fig. 8a shows a map of total views (x-axis) versus the average watch percentage $\bar{\mu}_{30}$ (y-axis) over the first 30 days. Videos from the comedy channel *TheEllenShow* have 100K to 1 million views, with $\bar{\mu}_{30}$ around 80%. Videos from the cooking vlog *KEEMI* on Korean food have similar amount of views, but their $\bar{\mu}_{30}$ is around 20%. The political channel *Blunt Force Truth* has 1K views on average, but a $\bar{\mu}_{30}$ of 80%. This example reveals that views are insufficient proxies for measuring the actual time users spend with an online video, and prompts us to investigate a richer set of engagement metrics.

Here, we systematically measure the agreement between views and another engagement measure - total watch time - by examining sets of top n videos at 30 days. Fig. 8b measures this overlap using the co-occurrence rate (size of intersection over n , left plot) and Spearman’s ρ coefficient (between -1 and 1, right plot). Fig. 8b shows that viewcount and watch time agree concerning only 50% of *ALL* videos (and the agreement decreases slowly as n increases). Furthermore, the agreement is dependent of video category: is it higher for Music ($\alpha \sim 0.8$) and particularly low for News ($\alpha \sim 0.4$), for which ρ shows no, or slightly negative correlation.

These result indicate that the viewcounts and the watch time provide different views of how users interact with the YouTube videos. Next, we analyze their interplay to construct more robust measures of user engagement.

Engagement map and *relative engagement*. Recent studies [Salganik et al., 2006], [Krumme

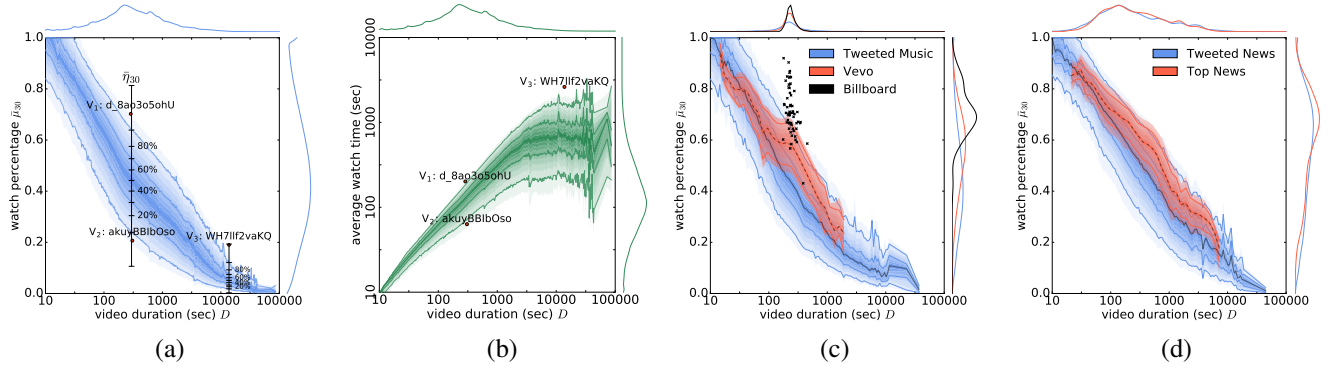


Figure 9: User engagement on the TWEETED VIDEOS dataset. **(a)** the *engagement map*; **(b)** video duration vs average watch time; *Relative engagement* and video quality for sets of videos. Videos in the QUALITY VIDEOS dataset are shifted towards higher *relative engagement* compared to the TWEETED VIDEOS videos in the categories Music **(c)** and News **(d)**.

et al., 2012] show that the quality of a digital item is linked to the user’s conscientious decision to continue watching or listening after first opening it. Therefore, the amount of time users spend watching it in a single viewing session should be indicative of video quality. For a given video, we compute the *average watch percentage* $\bar{\mu}_t$ up to day t as the total watch time divided by the total number of views and normalized by video duration D

$$\bar{\mu}_t = \sum_{i=1}^t x_w[i] / [D \cdot \sum_{i=1}^t x_v[i]] \quad (38)$$

Engagement map. As observed by Park et al. [2016], video duration has a significant impact on the *watch percentage* – in the TWEETED VIDEOS dataset, duration alone explains more than 45% of the variance of *watch percentage*. Intuitively, longer videos are less likely to be fully watched compared to shorter videos due to limited attention span. Consequently, we construct a two dimensional map, for which the x-axis shows the video duration D in log scale, and the y-axis shows the watch percentage $\bar{\mu}_{30}$ over the first 30 days. In Fig. 9a, we project all the videos in the TWEETED VIDEOS dataset onto this *engagement map*. The x-axis is split into 500 equally wide bins, and each video is placed in one of the bins. The color shades correspond to percentiles of *watch percentage* in each bin: the darkest color corresponds to the median value and the lightest correspond to the extremes (0% and 100%). This *engagement map* calibrates *watch percentage* against different video durations: highly-watched videos are projected towards the top of each duration bin, while barely-watched videos are on the bottom.

Note that a logically identical map could be constructed using the average watch time as the y-axis – shown in Fig. 9b. However, the duration-time map has several shortcomings: first, watch time confounds the effects of viewcount and of the watch duration in each session, which cannot be separated unless directly observing the individual watch sessions; second, the average watch time has no upper bound, making this map difficult to read.

Relative engagement η . Based on the *engagement map*, we propose the *relative engagement* $\eta \in (0, 1)$, defined as the rank percentile of each video in its duration bin. We denote the mapping from watch percentage μ to *relative engagement* η as lookup table f (constructed from data), and

its an inverse lookup f^{-1} :

$$\bar{\eta}_t = f(\bar{\mu}_t) \Leftrightarrow \bar{\mu}_t = f^{-1}(\bar{\eta}_t) \quad (39)$$

Fig. 9a illustrates the relation between video duration, *watch percentage* and *relative engagement* for three example videos. Videos v_1 ([d.8ao3o5ohU](#)) is on kids playing karate and v_2 ([akuyB-BibOso](#)) is on teaching toddler colors. They are of similar duration, but different watch percentages $\bar{\mu}_{30}(v_1) = 0.70$ and $\bar{\mu}_{30}(v_2) = 0.21$. This amounts to very different values of the *relative engagement*: $\eta_{30}(v_1) = 0.96$, while $\eta_{30}(v_2) = 0.07$. Video v_3 ([WH7llf2vaKQ](#)) is a much longer video ($D = 3$ hours 49 minutes) containing a live fighting show. It has a relatively low *watch percentage*, similar to v_2 ($\bar{\mu}_{30}(v_3) = 0.19$). However, its *relative engagement* $\eta_{30}(v_3) = 0.99$, positioning it among the most engaged videos in its peer group.

While watch time has been studied in the literature [Park et al., 2016, Dobrian et al., 2011], to the best of our knowledge, our paper [Wu et al., 2018] is the first to measures the average watch percentage in large-scale dataset, and to map its non-linear relation with video duration.

3.3 Popularity measures via (deep) neural networks

RNN-MAS describes popularity under influence of heterogeneous streams of shares and tweets. We propose two metrics based on RNN-MAS to quantify average response to unit promotion, and the relative influence among users of different *fame*. We also utilize the model to estimate a response series to unseen influence for a video.

Simulation

We can use our learned volume prediction models to simulate a series of views $V_d^{P(\cdot)}$ from $d = 1$ to $d = T$ days, where $V_d^{P(\cdot)}$ stands for views on d^{th} day under a promotion function $P(\cdot)$. For computing views we evolve hidden states as per the Equation 28 and Equation 30 for our models, where the input parameter is defined by the value of function $P(\cdot)$ at various time steps. At the end of each time step, views are computed as per Equation 31 and Equation 29 for RNN-MAS and Volume RNN respectively. We define, $\nu(P(\cdot))$, the cumulative views generated by promotion function $P(\cdot)$ from $d = 1$ to $d = T$ days as:

$$\nu(P(\cdot)) = \sum_{d=1}^T V_d^{P(\cdot)} \quad (40)$$

For all our simulations we choose $T = 10,000$ days, as this represents ≈ 27 years, longer than lifetime of any YouTube video.

Response to unseen influence

Despite taking into account multiple sources of external influence, there are influence signals that our models, RNN-MAS and volume RNN, do not capture explicitly. Examples include seasonality, or discussions in forums that are known to be an important factor for *gaming* videos — widely promoted on forum www.minecraftforum.net.

For understanding this response to unseen influence we set the promotion function $P(\cdot) = 0$ in our simulation setup described in Section 3.3 and obtain a series of $V_d^{P(\cdot)=0}$.

The response series to unseen influence here generalizes previous definitions. In HIP [Rizoiu et al., 2017] external influence $u[t]$, is assumed to be an initial impulse plus a constant (Equation 5 in Section 2.4). In this work the shape of response series to latent promotions is unconstrained. In our paper [Mishra et al., 2018], we present various case studies to illustrate how the response series to unseen influence for our models is able to capture richer temporal trends, such as seasonality.

Response to unit promotion.

In linear time variant (LTI) systems such as HIP, the total gain per unit promotion, ν , is well defined. It's calculated by computing the impulse response [Rizoiu and Xie, 2017] of the system. In our models, RNN-MAS and volume RNN, the notion of impulse response does not readily apply as it's a non-linear time variant system.

We compute the average response to unit impulse in three steps: compute response of our models to p units of promotion to volume RNN at $d = 0$ days; subtract response to unseen influences, calculated as per Section 3.3; normalize it by p . Response to p units of promotion is calculated using function $P(d) = p\mathbb{1}[d]$ in simulation setup of Section 3.3, where $\mathbb{1}[d]$ takes a value of 1 at $d = 0$, and 0 otherwise. The average response to unit promotion from promotion $P(d) = p\mathbb{1}[d]$, $\varrho(p)$ named as promotion score, is calculated as:

$$\varrho(p) = \frac{\nu(p\mathbb{1}[d]) - \nu(0)}{p} \quad (41)$$

In this definition, promotion score can be negative, capturing cases where being promoted decreases popularity (e.g. spam or paid tweets). Also note that $\varrho(p)$ is a function of p , i.e. for different simulation parameter p , the average response to unit promotion is different. This allows us to describe phenomena of diminishing returns in many real-world marketing scenarios [Jones, 1990]. In our paper [Mishra et al., 2018], we examine the effect of p and found that $\varrho(p)$ tend to decrease as p grows.

Loudness level

Using simulation metrics similar to ϑ and $\varrho(p)$, we can measure the influence of an individual or a group of users by the responses generated by their tweeting events.

Loudness level is a relative measure of total influence for one or a group of users. It is measured in decibels (dB). It is computed as log-ratio of the response of the target group $\psi(S)$ versus that of a comparison value ψ_0 .

We compute $\psi(S)$ in two steps: calculate $\nu(P(\cdot))$ such that $P(\cdot) = \{s_i\}$ in simulation setup of Section 3.3, where $\{s_i\}$ is the set of tweets generated by user group S ; subtract response to unseen influences, calculated as per Section 3.3. It can be written as:

$$\psi_{dB}(S) = \log_{10} \left(\frac{\psi(S)}{\psi_0} \right) dB \quad (42)$$

such that:

$$\psi(S) = \nu(s_i) - \nu(0)$$

where ψ_0 , the comparison value, is set to 1 for experiments in Section 4.7.3. As loudness level accounts for all activity generated by a user group, it can be used to compare the relative effects of promotion between users of different *fame* in Twitter.

4 Selected results

In this section, we present a selection of the applications and the insights obtained in this project. Firstly, we explore the performances of the Hawkes process for social events described in Sec. 1.1, when tasked with the prediction of the final size of a diffusion cascade after observing its initial unfolding. Next, in Sec. 4.2, we introduce the visualization tool that we constructed in [Rizoiu et al., 2017] and that we used to explain popularity, describe most popular videos and detect a particular class of items: *the unpromotable videos*. Next, in Sec. 4.3 we explore HIP’s power of predicting a video’s expected popularity under external promotions: we first illustrate the setting for popularity forecasts using video examples, and then present a quantitative evaluation. Next, in Sec. 4.4, we explore the relationship between viral potential, promotion and popularity. Next, Sec. 4.4 illustrates the relationship between promotion and popularity using two example videos that have similar viral potential. Next, We quantify the relative influence of four factors – the video’s viral potential, the amount of promotion, the timing of the promotion and the video’s prior position on the popularity scale – in two predictive exercises: predicting future views (Sec. 4.4.2) and predicting significant rises in popularity (Sec. 4.4.3). Next, in Sec. 4.5, we focus on time as a key variable for deploying promotions and observing their outcome, and present two novel observations. The first one explores the trade-off between maturity time of a video and the time-sensitive value of promotions under the economic setting of compounding interest. In the second, we explain the common practice of advertising at regular intervals as an increase in perceived potential. Next, in Sec. 4.6, we study the behavior of the *relative engagement* over a large collections of videos and we show that it demonstrates correlation to external notions of video quality (Sec. 4.6), and is stable over time (Sec. 4.6.2). Next, in Sec. 4.7 we study the relative effect of tweeting by either celebrities or grass-root users, which allows us to select which category is more effecting in promoting particular types of online content. Finally, in Sec. 4.8 we detail a technique for computing the final size distribution of a partially observed cascade, and we provide a new result concerning the unpredictability of online popularity.

4.1 Results: predicting retweet cascade size

In this section, we present some of the results of predicting the total diffusion cascades sizes in Twitter. The complete results and their interpretation are in our paper [Mishra et al., 2016a].

We compare our point processes model described in Sec. 1.1, denoted as HAWKES, with SEISMIC, the current state of the art in predicting cascade sizes [Zhao et al., 2015], on two datasets, TWEET-1MO and NEWS. Note that TWEET-1MO is a publicly available dataset on which Zhao et al. [2015] reported results, while NEWS dataset – described in Sec 2 – supports extracting a richer set of features that is not available in TWEET-1MO. In addition, Zhao et al. [2015] showed that SEISMIC compared favorably against a number of baselines: from auto-regressive and generative models such as dynamic Poisson model [Agarwal et al., 2009, Crane and Sornette, 2008a] and reinforced Poisson model [Shen et al., 2014], to linear regression and its variants [Szabo and Huberman, 2010a].

Experimental setup. We run our experiments on TWEET-1MO and a subset of NEWS: 20,093 cascades from the month of July, which have the total length of at least 50. Both algorithms observe the initial part of each cascade for a limited extent of time and predict its final popularity. We estimate separate sets of parameters for HAWKES and SEISMIC after observing the cascades for

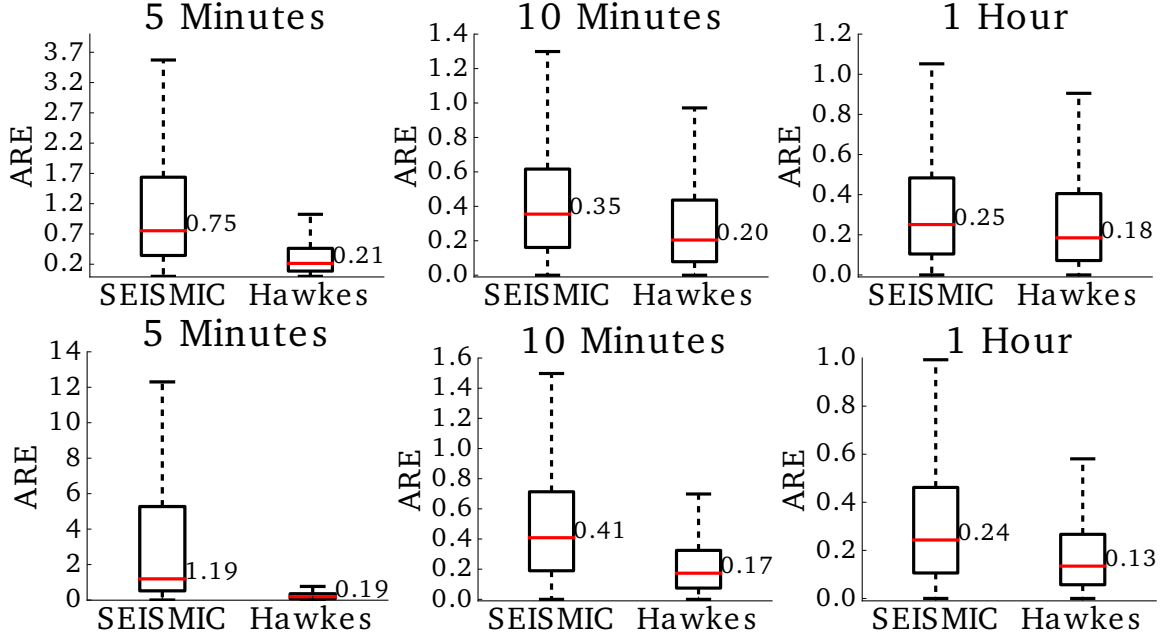


Figure 10: Distribution of Absolute Relative Error (ARE) on the TWEET-1MO dataset (top row) and on the NEWS (bottom row), for SEISMIC and HAWKES. A part of the diffusion cascade is observed before making the predictions: 5 min (left column), 10 min (middle column) and 1 hour (right column). The red line and the numeric annotations denotes the median values of the distributions.

5 minutes, 10 minutes and 1 hour, respectively. After the fitting is finished, we train the Random Forest regressor in the predictive layer of the HAWKES model. Performance of this regression layer is reported using ten fold cross-validation. 40% of cascades are used for training, 60% for testing. All experimental protocol choices – minimum cascade length threshold, random training/testing split and length of time prefixes – are made to mimic closely the original experimental setup [Zhao et al., 2015].

We use the Absolute Relative Error (ARE) to measure prediction performance on a given cascade w :

$$ARE^w = \frac{|\hat{N}_\infty^w - N_{real}^w|}{N_{real}^w},$$

where $\hat{N}_\infty$ and N_{real} are the predicted and observed popularities of cascade w .

Results. The prediction results are summarized in Fig. 10 as boxplots of ARE for each combination (dataset, approach, observed time). Note that generative models do not output a prediction for all cascades. For example, when branching factor $n^* \geq 1$ for HAWKES, the estimate for cascade size is infinite, similar failures also occur with SEISMIC when the infectiousness parameter $p \geq \frac{1}{n^*}$ ([Zhao et al., 2015] Eq(7)). We only report the ARE on cascades on which both approaches produce a prediction. We observe that HAWKES is able to produce meaningful predictions for more cascades than SEISMIC. For example, on NEWS data with 5 minutes, we output prediction for 873 (or 0.7%) more cascades than SEISMIC.

We can see that HAWKES consistently outperforms SEISMIC, as it achieves lower ARE on both

datasets. Mean ARE for HAWKES is at least 40% better than SEISMIC on TWEET-1MO and 33% on NEWS. We observe that NEWS dataset after 5 minutes of observation seems the most difficult to predict: HAWKES produces a mean ARE of 0.42, while the mean ARE for SEISMIC is 11.13. Fig. 10 (top row) shows that the median of HAWKES is at least 25% better than SEISMIC for any time prefix, suggesting that our approach predicts better for most of the cascades in TWEET-1MO. Similar conclusions can be drawn from the bottom row of Fig. 10 (NEWS data): the median ARE of HAWKES is at most 0.19 at 5 minutes, while the best of SEISMIC models is a median ARE of .24 at 1 hour.

We are glad to see that the HAWKES model achieves better predictive performance than SEISMIC, by exploiting full flexibility in its non-linear parameters, and a predictive layer that systematically optimizes for future prediction using information from other cascades.

4.2 Results: explaining popularity

In this section, we present a selection of the results concerning explaining popularity, fully described in our paper [Rizoiu et al., 2017]. Using the two quantities defined in Sec. 1.2.3, we construct a 2-dimensional map with *endogenous response* A_ξ as the x-axis and *exogenous sensitivity* μ as the y-axis. We call this plot the *endo-exo map*. This section presents example uses of this map for explaining video popularity, and identifying videos that are not promotable.

Explaining popularity. Intuitively, a video with a large endogenous response A_ξ and a high exogenous sensitivity μ has high potential to become viral. Specifically, each unit of exogenous excitation will generate μA_ξ events through the Hawkes intensity process. On the endo-exo map, videos in close proximity have similar potentials to become popular and the differences in their popularity would be due solely to the difference in exogenous attention. Fig 11(a) illustrates this phenomena using four videos. Videos v_1 and v_2 are very similar in both A_ξ and μ ; the fact that v_1 has 4.61x more views is explained by it receiving 3.22x more exogenous promotions. On the same map, v_4 received a similar amount of promotion as v_1 and their differences in popularity are explained by v_4 being less endogenously responsive (smaller A_ξ) than v_1 . Moreover, v_3 has a similar endogenous response and sees similar amounts of promotion as v_1 ; their differences in popularity are explained by v_3 being less exogenously sensitive, with a lower μ . The endo-exo map provides two distinct aspects from which a video’s popularity can be attributed to, detailed in the rest of this section.

What describes the most popular videos? One may wonder whether higher popularity can be attributed to higher exogenous sensitivity, higher endogenous response or a combination of both. We examine a collection containing diverse video categories and find that the explanation varies. We draw on the endo-exo map all the videos that belong to the same category in the ACTIVETWEETS dataset and we visualize them as two-dimensional density plots. Fig. 11 (c) and (d) compares the plots for the videos in Gaming and Film & Animation, to that of the top 5% in popularity from these two categories, respectively. Visibly, while most popular videos in Film & Animation are described by higher exogenous sensitivity (shifting upwards), the most popular Gaming videos have higher endogenous response – their density mass is shifted to the right of the endo-exo map. Other categories such as Comedy or News & Politics present two dense regions, one for higher A_ξ and one for higher μ . These observations show that the most popular videos in different categories differ in terms of the two main factors that drive popularity.

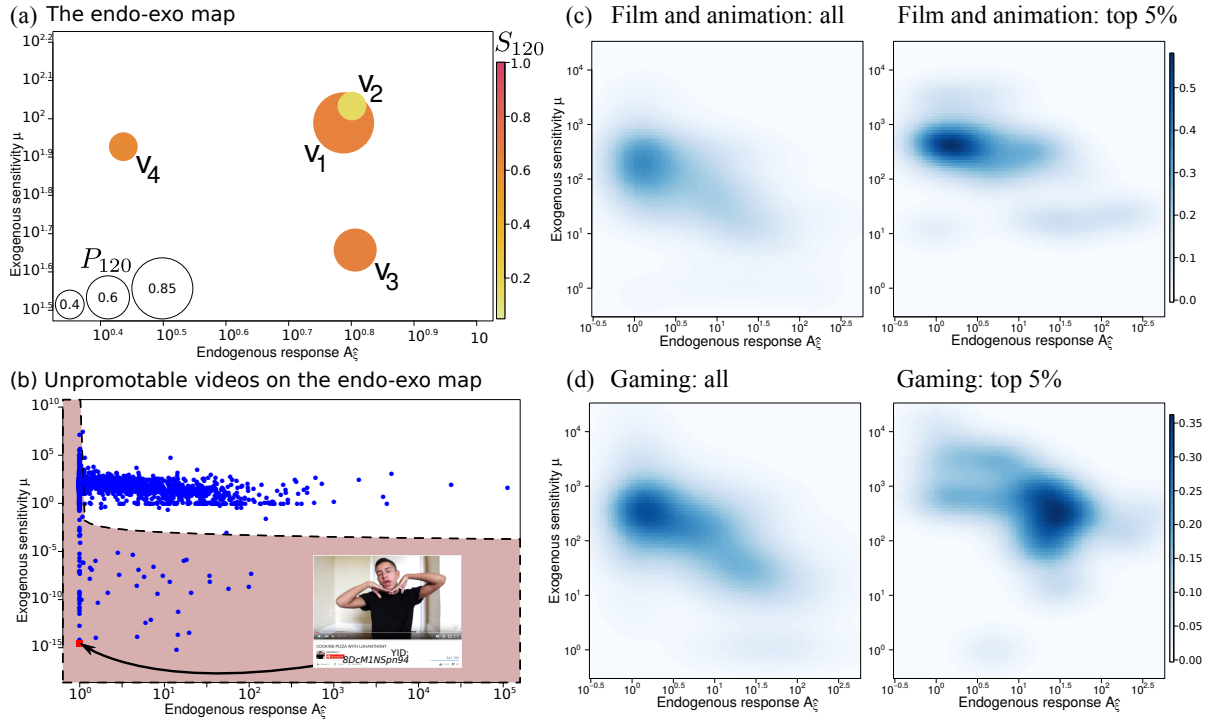


Figure 11: Visualizing video virality and video popularity using the *endo-exo map*. (a) Four example videos on the endo-exo map. X-axis A_ξ : the magnitude of endogenous reaction; Y-axis μ : sensitivity to exogenous stimuli. The radius of each circle is proportional to the *popularity percentile* $P_t(\cdot)$ of each video after $t = 120$ days, with values between 0.0 (least popular) and 1.0 (most popular). The color represents the amount (percentile) of total shares received, denoted as $S_t(\cdot)$, with values between 0.0 (no promotion) and 1.0 (receiving the most promotions). v_1 and v_2 present similar endogenous reaction and exogenous sensitivity, being at the same position on the endo-exo map. The difference in their popularity (size) is explained by the fact that v_1 received 3.22 times more promotions than v_2 . Both v_3 and v_4 receive similar amounts of promotion (color) as v_1 , but they achieve lower popularity (smaller size) due to their less privileged position on the endo-exo map: v_3 is less sensitive to external stimuli than v_1 and v_2 , while v_4 has a smaller endogenous reaction than v_1 and v_2 . Information about the four example videos are as follows, with their popularity percentile P_{120} and shares percentile S_{120} : v_1 is a short Gaming video, YoutubeID *0lTTWeavllc*, $P_{120}(634, 370 \text{ views}) = 85\%$, $S_{120}(351 \text{ shares}) = 65\%$; v_2 is a collection of “ALS ice bucket challenge” videos, YoutubeID *3hSIh-tbiKE*, $P_{120}(137, 481) = 40\%$, $S_{120}(109) = 10\%$; v_3 is a funny science video, explaining types of infinity in math, YoutubeID *23I5GS4JiDg*, $P_{120}(193, 052) = 60\%$, $S_{120}(356) = 65\%$; v_4 is from a Portuguese youtuber, YoutubeID *0ndmJzElcgU*, $P_{120}(93, 959) = 40\%$, $S_{120}(311) = 60\%$. (b) A zoomed-out scatter plot of the endo-exo map of the videos in the People & Blogs category. The shaded portion of this map consists of videos with low values of total response $\mu A_\xi < 10^{-3}$ and hence dubbed *unpromotable* videos. Thumbnail of an example video *8DcMINSpn94* is included, with $\mu = 2.88 \times 10^{-15}$ and $A_\xi = 1$. (c) Density plot for all (left) vs the most popular 5% (right) Film & Animation videos. (d) Density plot for all (left) vs the most popular 5% (right) Gaming videos. Popular Film and Animation videos tend to have a higher exogenous sensitivity, while those for Gaming have mainly a higher endogenous response.

Identifying unpromotable videos. The endo-exo map can be used to readily identify an interesting class of videos, which are very difficult to promote. Given that the quantity μA_ξ describes

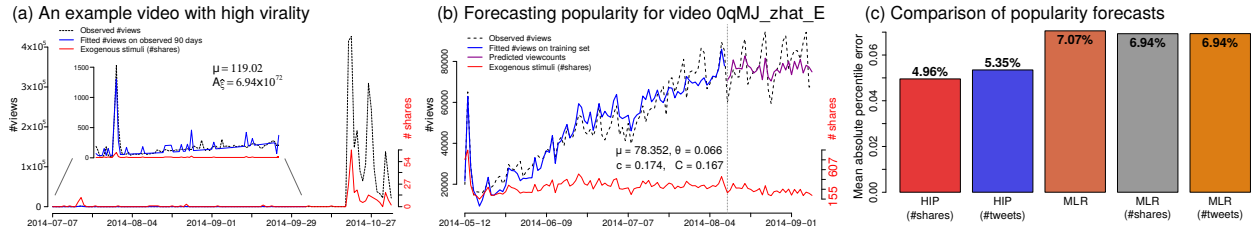


Figure 12: Popularity forecasting using the Hawkes intensity process. (a) Popularity series for video *1PuvXpv0yDM*, explaining a brain disorder. The video receives a total of 36 shares and 15,687 views in the first 90 days (see inset), it is estimated to have a high exogenous sensitivity and a high endogenous response ($\mu = 119.02$, $A_{\xi} = 6.94 \times 10^{72}$). Between day 91 and day 120, this video *jumped* from a popularity percentile of 5.85% to 94.9%, receiving 229 shares and gaining 2.42 million views. (b) Forecasting popularity for video *0qMJ_zhat_E*. Black dotted line: viewcounts series from day 1 to 120 after video upload. Red line: exogenous stimuli $s(t)$, also used in parameters estimation. Left of the gray dashed vertical line at $T \leq 90$ days: time period used for parameters estimation. Blue line: fitted viewcounts for $T \leq 90$ days, generated using Eq 11. Magenta line: viewcount forecast for day 91 to 120. (c) Comparison of average forecasting errors on the ACTIVETWEETS set. y-axis: Forecasting errors, calculated as the absolute difference between the popularity percentile at day 120 and that forecasted by each approach. x-axis, left to right: Hawkes intensity model, using either #shares or #tweets as $s(t)$; multivariate linear regression (MLR), using only popularity history, or #shares and #tweets, respectively.

the number of views that one unit of external promotion (via sharing or tweeting) will generate under the joint influence of endo- and exo- factors – a very small μA_{ξ} (e.g., $\mu A_{\xi} < 1e - 3$) is a hallmark of a video being *unpromotable*. Fig. 11(b) contains a zoomed-out view of the endo-exo map associated with the category People & Blogs. We found 63 videos ($\sim 3.2\%$) in this category to be unpromotable. Overall, 549 ($\sim 3.9\%$) videos in the ACTIVETWEETS set are deemed unpromotable. The thumbnail of one example video (a teenager video blog) is shown. It has $\mu = 2.88 \times 10^{-15}$ and $A_{\xi} = 1$, hence each online promotion is expected to generate 0 views. In contrast, for video v_1 in Fig. 11(a), each promotion is expected to generate 598 views.

4.3 Results: forecasting popularity

Via the endo-exo map, the Hawkes intensity process prescribes a video’s expected popularity dynamics under external promotions. This section explores the predictive power of such a model. We first illustrate the setting for popularity forecasts using video examples, and then present a quantitative evaluation. The full popularity forecasting results can be found in our paper [Rizoio et al., 2017].

4.3.1 A video that will go viral

We use HIP to identify videos that are not already popular but have a high potential to become so. This is similar to the phenomenon of delayed recognition in science [Ke et al., 2015]. Note that this approach is predictive in that we aim to find such potentially viral items before they become popular, rather than a measurement-driven approach that analyzes viral items in past history. Video

IPuvXpv0yDM in Fig. 12(a) is such an example, it received 15,687 views after being online for 90 days. The HIP model deems it to have high endogenous response ($A_{\xi} = 6.94 \times 10^{72}$) and high exogenous sensitivity ($\mu = 119.02$). Between days 91 and 120, the video received an additional 229 shares, more than 6 times the number of shares during its first 90 days. This video gained 2.42 million views, drastically improving its ranking on the popularity scale from 5.85% to 94.9%.

4.3.2 Evaluation of forecast

The HIP model takes as input the exogenous promotion $s[t]$ to produce estimates of the viewcount $\xi[t]$. To construct $\xi[t]$ in the future, $s[t]$ needs to be either estimable or known. We call this *forecasting* popularity, as opposed to predicting popularity where no information about future exogenous stimuli is assumed. Forecasting popularity has broad applications, such as estimating the effect of intended (promotional) interventions, and making decisions about when to promote.

Evaluating popularity forecast on temporal hold-out data. We design a protocol to quantitatively evaluate the predictive power of HIP. We use historical data held-out over time, thus avoiding the practical difficulty of generating realistic promotions and responses in a large-scale social network. Using (known) exogenous promotion $s[t]$, we forecast the popularity $\sum_{91}^{120} \xi[t]$ during the evaluation period (in purple) using Eq 12. Fig. 12(b) illustrates this setting with an example music video. A vertical line divides the observation period, day 1 to 90, and the evaluation period, day 91 to 120. The viewcount and sharing history in the observation period is used to fit model parameters and explain observed popularity (in blue). For this example, the forecast and the actual views are fairly similar.

Percentile-error metric. We obtain a predicted total viewcount at the end of evaluation period, i.e., $\sum_1^{90} \bar{\xi}[t] + \sum_{91}^{120} \xi[t]$, and compare it to the actual total viewcount $\sum_1^{120} \bar{\xi}$. Common error metrics such as root-mean-square-error (RMSE) or normalized RMSE are skewed by the large number of outliers in a long-tailed viewcount distribution. We map the number views to the popularity scale in Sec. 2, i.e. applying the mapping function $P_{120}(x)$ to normalize views into a metric between 0 and 1. We then compute the absolute error of the predicted percentile. Compared to error metrics based on the difference in views, this metric focuses on ranking videos correctly with respect to a large collection and is as useful as the broad class of learning to rank applications.

Baseline algorithms. The state-of-the-art approach for popularity prediction uses multivariate linear regression (MLR), based on the observation that historic viewcounts are predictive for future viewcounts [Pinto et al., 2013b, Szabo and Huberman, 2010b]. We train linear regressors to predict daily viewcounts for each day between 91 and 120, using a 90-dimensional feature corresponding to the number of views in days 1 to 90. To give the MLR forecast the same amount of information as the HIP model, we build two enhanced baselines, denoted by MLR (#shares) and MLR (#tweets), by introducing the exogenous influence as additional variables, both in the training and in the prediction. Note that the HIP models for each video are learned and evaluated independently, all baselines are trained on ACTIVETWEETS and we obtain predictions for each video using cross-validation.

4.3.3 Forecasting results

Fig. 12(c) summarizes forecasting performance for HIP and the MLR baselines. We can see that the forecasts made using HIP have lower average error compared to linear regression with exoge-

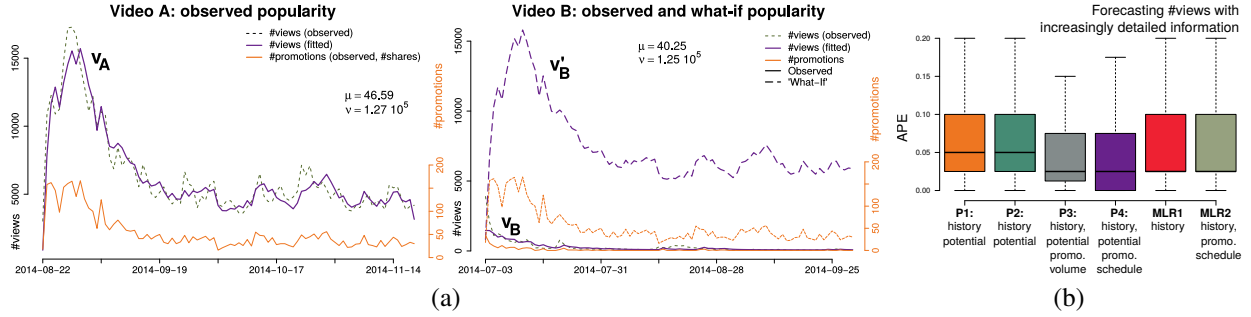


Figure 13: **(a)** The “What-If” analysis for two videos (v_A and v_B) with a similar viral potential, having achieved different popularity levels. (left) The observed and fitted popularity series, the observed promotion and fitted parameters for v_A . (right) Observed, fitted popularity series for v_B , and the hypothetical series (denoted as v'_B), if v_B had had the same promotion as v_A . **(b)** APE when predicting the number of views between 91 and 120 days, multiple predictors, using shares as promotions.

nous stimuli (#shares, #tweets) or without. The best forecast obtained an average percentile error of 4.96% (median 3%) for HIP (#shares) and 6.94% (median 3.75%) for the MLR (#shares), corresponding to a 28.6% relative reduction of error. These differences are statistically significant with paired t-test $p < 0.001$, medium effect size according to Cohen’s d [Cohen, 1988]. Within the HIP variants, we found that using the number of shares generates slightly better forecast than the number of tweets, but the differences are not statistically significant at $p = 0.001$.

We also observe that the performance gap doubles when forecasting popularity on more difficult videos – videos with a large exogenous shock in the forecasting period, defined as the mean number plus 100 times the standard deviation of the number of shares during the observed period. Fig. 12(a) shows an example of such a video. There are 4006 such videos in the ACTIVETWEETS dataset, for which HIP (#shares) achieves a mean percentile error of 5.11% (median 3.25%), whereas MLR (#shares) achieves a mean error of 9.24% (median 6.5%). A typical situation when HIP misses the forecast is when none or very little external influence is recorded during the observed period and during which the popularity is likely to have been driven by unseen exogenous sources.

4.4 Result: future popularity under promotion

In this section, we explore the relationship between viral potential, promotion and popularity. Sec. 4.4.1 illustrates the relationship between promotion and popularity using two example videos that have similar viral potential. HIP is the state of the art in popularity forecasting [Rizoiu et al., 2017], however the factors that influence the performance of the prediction were not well understood. We quantify the relative influence of four factors - the video’s viral potential, the amount of promotion, the timing of the promotion and the video’s prior position on the popularity scale - in two predictive exercises: predicting future views (Sec. 4.4.2) and predicting significant rises in popularity (Sec. 4.4.3).

Predictor	pred. / forecast	Historic Views	Promo.	Viral potential	Future promo. Volume	Time	Equation	mean APE $\pm$ std
B	prediction	✓					$\Psi_{120}^{-1}(\Psi_{90}(V_{1:90}))$	16.17% $\pm$ 14.07%
P₁	prediction	✓	✓	✓			$V_{1:90} + \Xi^{old}[91, 120]$	8.37% $\pm$ 11.17%
P₂	prediction	✓	✓	✓			$V_{1:90} + \Xi^{old}[91, 120] + \nu S_{61:90}$	8.12% $\pm$ 12.07%
P₃	forecast	✓	✓	✓	✓		$V_{1:90} + \Xi^{old}[91, 120] + \nu S_{91:120}$	4.98% $\pm$ 6.43%
P₄	forecast	✓	✓	✓	✓	✓	$V_{1:90} + \Xi^{old}[91, 120] + \Xi^{new}[91, 120]$	4.94% $\pm$ 6.38%
MLR₁	prediction	✓						7.07% $\pm$ 9.32%
MLR₂	forecast	✓	✓		✓	✓		6.94% $\pm$ 9.09%

Table 3: Summary of predictors used for predicting future views. The columns show the factors included in each predictor, the predicted value (equation) and the obtained prediction performance.

4.4.1 What-if scenarios for two videos

The relationships among viral potential, promotion, and popularity can be understood in a scenario where we can ask “what-if” certain promotions are applied to a given video, and compare the actual and the hypothetical outcomes. To this end, we select two videos from the ACTIVETWEETS dataset that have the same viral potential. Video v_A (id 8jPi r15Ms8c) is an Entertainment video, having achieved a high popularity percentile $\Psi_{120}(v_A) = 85\%$ and it is among the top 2.5% most shared videos, receiving 6192 shares in 120 days. Video v_B (id WtOC9Lw-gdc), is a Music video, with a low popularity percentile and few shares ($\Psi_{120}(v_B) = 15\%$, 162 shares). Although the two videos have similar viral potential scores ($\nu \simeq 126,000$), their popularity series present different dynamics (shown in Fig. 13a). The only factor that led to the popularity difference in video v_A and video v_B seems to be the amount of promotions they receive. One may ask “what if” v_B actually went through the same promotions as v_A . We simulate the popularity series of video v_B driven by the promotion series of video v_A , using Eq. (11). As shown in Fig. 13a (right), this “what-if” series is denoted as v'_B , and has much more views than v_B . The total popularity of v'_B is now almost the same as the popularity of v_A . This example shows that given sufficient promotion, an unpopular video with high potential could become popular. This insight is used to predict popularity in the rest of this section.

4.4.2 Predicting future views

In this section, we study the factors which impact the prediction of the number of views received by a video, as a result of promotions. Based on HIP, we construct four predictors, which embed increasingly detailed information about promotions, and we show that this leads to progressively better estimates. More precisely, we study the impact of the amount of promotion and the timing of the promotion. We also study the case when information about future promotions is not available, quantifying how much can be inferred about future popularity based on past popularity only. We denote this latter setup as *prediction*, in contrast with *forecasting* the popularity obtained through promotions.

Experimental setup. We construct a temporal holdout setup, each video is observed for the first 90 days, on which we fit the parameters of the HIP model, using either tweets or shares as external stimulation. The next 30 days (from 91 to 120) serve as the test period, and the tweets and the shares in this period are used as promotion. We opted for this setup due to the practical difficulty of performing interventions in large social environments, such as Youtube and Twitter. Using tweets and shares during test period as promotions is also supported by the lack of Granger

causality between the views, shares and tweets series, shown in [Rizoiu et al., 2017]. Commonly-used performance error metrics (such as the root-mean-square-error) are skewed by the long-tailed viewcount distribution shown in Fig. 6, and we chose not to use them. Instead, we evaluate each predictor using the Absolute Percentile Error (APE):

$$APE = |\Psi_{120}(\Xi) - \Psi_{120}(V_{1:120})|, \quad (43)$$

where Ξ and $V_{1:120}$ are the predicted and the observed number of views during the first 120 days, respectively.

Constructed predictors. Table 3 summarizes the four constructed predictors and the three used baselines, showing the information leveraged by each predictor, the predicted value formula (where applicable) and the prediction performance. The predictors $\mathbf{P}_1$ and $\mathbf{P}_2$ operate in a predictive setup, they do not make use of the promotion information. $\mathbf{P}_1$ leverages history (views and attention that the videos received during the training period) as well as the video viral potential. It quantifies the views in the period 91:120 obtained by the unfolding of the “old” promotions (promotion observed during the first 90 days) using $\Xi^{old}[1 : 90]$ (Eq. 37). The predictor $\mathbf{P}_2$ is similar to $\mathbf{P}_1$ and it makes a more realistic assumption about future promotions: it assumes that videos receive during the period 91-120 days the same amount of promotion as in the period 61-90 days (denoted as $S_{61:90}$). The third predictor, $\mathbf{P}_3$, leverages history, viral potential and future promotion volume; it works in a forecasting setting in which the number of promotions in test period are known ($S_{91:120}$). Predictor $\mathbf{P}_4$ leverages history, viral potential and promotion scheduling; in addition to $\mathbf{P}_3$, it leverages the promotion schedule in the test period $s(t), t \in [91, 120]$ and predicts the number of views using the formula in Eq. (37). We show in Table 3 the data baseline $\mathbf{B}$, which outputs, for each video, the number of views required for it to maintain, at 120 days, the same popularity percentile as at 90 days. We also implement two baseline predictors, using multivariate linear regression (MLR), based on the observation that past popularity is indicative of future popularity [Pinto et al., 2013a]. $\mathbf{MLR}_1$ leverages the past viewcounts to predict the viewcounts during the period 91-120 days. $\mathbf{MLR}_2$ is enhanced with information about promotions, both historic and future. Note that popularity prediction systems that require observing individual events [Zhao et al., 2015, Mishra et al., 2016a] cannot be used here, since the views information from Youtube is aggregated daily. $\mathbf{MLR}_1$ and $\mathbf{MLR}_2$ are evaluated with 5-fold cross validation. Shares are used to train the HIP model and $\mathbf{MLR}_2$ (and they serve as promotions in the forecasting setup).

Results. Fig. 13b and Table 3 summarize the performances of our predictors and the baselines, over the ACTIVETWEETS dataset. The data baseline $\mathbf{B}$ shows the difficulty of the problem, all other predictors and baselines outperform this data baseline. Visibly, introducing more detailed information consistently leads to a reduction in mean APE. Accounting for future promotions has the largest impact, reducing mean APE from 8.12% of $\mathbf{P}_3$ to 4.98% of $\mathbf{P}_2$. Further accounting for promotion schedule has a considerably smaller impact, with $\mathbf{P}_4$ outperforming $\mathbf{P}_3$ by only 0.04% in APE. Consistent with the results in [Rizoiu et al., 2017], both HIP-based forecasting $\mathbf{P}_3$ and $\mathbf{P}_4$ consistently outperform the baseline $\mathbf{MLR}_2$. The situation is different for the predictive exercise, where the baseline $\mathbf{MLR}_1$ outperforms both $\mathbf{P}_1$ and $\mathbf{P}_2$. Approximating the future volume of promotions (in $\mathbf{P}_2$) has only marginal impact.

Predictor	Oracle	Viral potential	Promo. volume	Pop. scale position	Equation	AUC
$\mathbf{R}_1$		✓			ν	0.6
$\mathbf{O}_1$	✓		✓		$\frac{V_{91:120}}{S_{91:120}}$	0.62
$\mathbf{R}_2$		✓	✓		$\nu S_{91:120}$	0.79
$\mathbf{O}_2$	✓				$V_{91:120}$	0.88
$\mathbf{R}_3$		✓	✓	✓	$\Psi_{120}(V_{1:90} + \nu S_{91:120})$	0.91
$\mathbf{O}_3$	✓			✓	$-\Psi_{90}(V_{1:90})$	1
					$\Psi_{120}(V_{1:120}) - \Psi_{90}(V_{1:90})$	

Table 4: Summary of predictors used for predicting popularity jumps. The columns show the factors included in each predictor, the predictors output value (equation) and the obtained prediction performance.

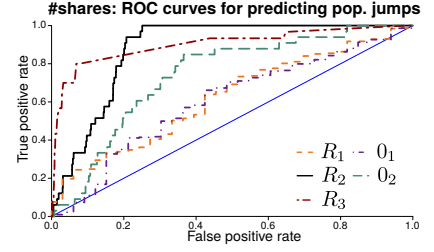


Figure 14: ROC curves of detecting popularity jumps – videos with a sudden increase in popularity of more than 20% between day 90 and 120, using shares as promotion.

4.4.3 Predicting popularity jumps

Recall in Fig. 6c there are 33 videos that had large *jumps* in the popularity scale. Predicting such sudden increase in popularity is relevant for content optimization or marketing. This can be done by building upon the methods in Sec. 4.4.2, and taking into account the non-linear effect of the popularity scale. We study the following factors which influence the prediction of popularity: the viral potential, the promotion volume and the position on the popularity scale.

Constructed predictors. Based on HIP, we construct three predictors which incorporate increasingly more information. $\mathbf{R}_1$ outputs the video viral potential. $\mathbf{R}_2$ leverages the viral potential and the amount of promotions: it outputs the forecasted number of views during the period 91-120 days. $\mathbf{R}_3$ leverages the viral potential, the amount of promotions and the position on the popularity scale at day 90: it outputs the forecasted popularity percentile gained during the period 91-120 days. Each predictor is doubled by an “oracle” predictor, based on the real data in the test period. The oracle predictors ($\mathbf{O}_1$, $\mathbf{O}_2$ and $\mathbf{O}_3$) serve as baselines to benchmark the maximum performance that can be achieved using a particular information. For example, $\mathbf{O}_1$ mirrors $\mathbf{R}_1$ in using the viral potential alone. Note that the viral potential is the number of views generated by a single promotion, and can be computed on the testing data as $\frac{V_{91:120}}{S_{91:120}}$ (where $V_{91:120}$ and $S_{91:120}$ are the series of view and shares during 91-120 days). Similarly, $\mathbf{O}_2$ outputs the *observed* number of view in the period 91:120 days, and $\mathbf{O}_3$ the popularity percentile gained in the same period.

Experimental setup. We employ the same temporal holdout setup as in Sec. 4.4.2: each video in the ACTIVETWEETS dataset is observed during the first 90 days, and the “jumping” behavior if predicted on the next 30 days. All predictors output real values, which we transform into a binary classifier using a threshold. We vary the threshold and we report the performance using ROC curves and AUC (Area Under Curve). Note that $auc \in [0.5, 1.0]$, higher is better, the random binary classifier obtains $auc = 0.5$ and perfect classification scores $auc = 1$. We varied the popularity jump threshold from 0.1 to 0.5 and the trends and the conclusions are similar for all values. Here, we show the results for a jump threshold of 0.2.

Results. Fig. 14 shows the ROC curves obtained for each predictor, using shares as promotions. The AUC obtained by each predictor is shown in Table 4. Both $\mathbf{R}_1$ and the oracle $\mathbf{O}_1$ achieve rather low prediction performances, with $auc(\mathbf{R}_1) = 0.60$ and $auc(\mathbf{O}_1) = 0.62$. This indicates that video potential is only weakly linked to jumping behavior, which happens due to a number of reasons:

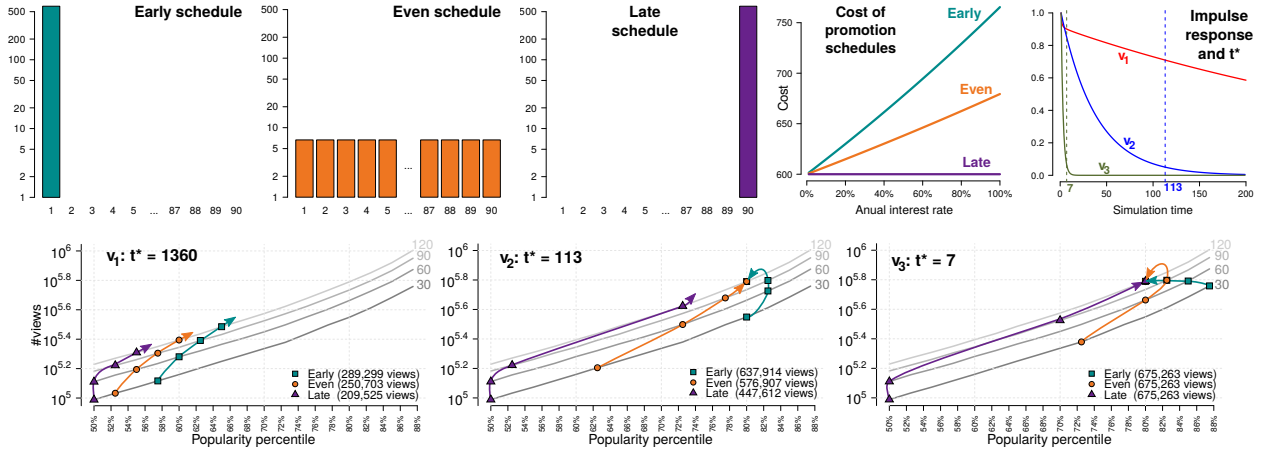


Figure 15: Analysis of three promotion schedules on the achieved popularity, after 120 days. (top) Distribution of 600 promotions over 90 days, for the **early**, **even** and **late** schedules (the y-axis is log-scale). The cost of schedules, function of interest rate. The impulse response and maturity time for the three subject videos: v_1 (id 9MwgExiIjnY) is a video of a Japanese youtuber, v_2 (id 5TQAnGjyN9A) is a short Entertainment video, v_3 (id d03B1Zp4GtQ) is a Brazilian boy gang video about playing games and drinking. (bottom) Reaction of each of the three videos to the promotion schedules.

videos of low potential that receive high promotion, or videos with high potential that receive low promotion, and the varying effects of views across the popularity scale. Accounting for the volume of promotion provides a boost in prediction performance, both for $\mathbf{R}_2$ ($auc(\mathbf{R}_2) = 0.79$) and for the oracle $\mathbf{O}_2$ ($auc(\mathbf{O}_2) = 0.88$). Note that the oracle $\mathbf{O}_2$ does not achieve perfect performance in predicting popularity jumps. This is due to the unequal impact of views on popularity, e.g. in order to jump 20% on the popularity scale, a video needs 83,245 views if it started from 25%, and 327,820 views if it starts from the median 50%. We obtain a performance boost when we account for the position on the popularity scale at day 90, with $auc(\mathbf{R}_3) = 0.91$. Note that the forecasts made using potential, promotion and scale information ($\mathbf{R}_3$) are more accurate than the oracle $\mathbf{O}_2$, achieving a 0.66 true positive rate with 0.05 false positive rate. This demonstrates the importance of taking into account the non-linear effect of the popularity scale. $\mathbf{O}_3$ obtains a perfect score ($auc(\mathbf{O}_3) = 1$).

4.5 Result: profiling promotion schedules

In this section, we focus on time as a key variable for deploying promotions and observing their outcome, and present two novel observations. The first one explores the trade-off between maturity time of a video and the time-sensitive value of promotions under the economic setting of compounding interest (Sec 4.5.1). In the second, we explain the common practice of advertising at regular intervals as an increase in perceived potential (Sec. 4.5.2).

4.5.1 The economics of time for promotions

Following from Corollary 3.1, the scheduling of S promotions does not matter in infinite time, as they eventually results in νS units of popularity gain. However, practical applications seek to

optimize the impact of promotions after a given time, e.g., how much was a video’s popularity boosted, a month after the promotion campaign has ended? The two counter-acting factors at play here are both about time – the time that the effects of a promotion have to unfold –, and the time-sensitive (monetary) value of carrying out promotions for a given piece of content. If the cost of promotion at any give time is constant, then the best strategy is to place all promotions as early as possible, this will lead to maximum unfolding of popularity in the target time window. A more realistic setting, however, needs to account for the varying value of promotions in carried out in different times. One simple way to model this is via compounding interest [Witt, 1613] with rate a ; say every dollar stored for one unit time (e.g. day) is worth $(1 + a)$ dollars. In this setting, putting all promotions at the beginning of k days actually needs $(1 + a)^k$ times the cost of putting all promotions at the last day.

The cost of three prototypical schedules. We first discuss three prototypical promotion schedules – late, even and early. Without loss of generality, we set the promotion budget at 600 promotions and the promotion time window at 90 days. **late** is spending the promotion budget at the last day of the promotion period. This is the least costly option, i.e. 600\$. **even** is spreading the promotion budget equally each day. Its cost is $\frac{600}{90} \sum_{k=1}^{90} (1 + a)^k$, visibly higher than the cost of late. **early** is spending the entire promotion budget in the first day, for a cost of $600(1 + a)^k$. These three schedules are general enough to provide an intuition of the trade-off between maturity and value of currency. An arbitrary schedule can be composed from them and its effects are also linear combinations of what is presented here, due to the LTI property (Sec. 1.2.3). The top row of Fig. 15 illustrates each of the three prototypical schedules and the cost of each schedule versus the annually interest rate.

Scheduling effects on three videos. We simulate the effects of each of the three promotion schedules on three videos and we plot the results for each video in Fig. 15 (bottom). The median values of the popularity scale at 30, 60, 90 and 120 days (as seen in Fig. 6) are shown as increasingly lighter gray lines. We plot the position (i.e. the popularity bin) obtained by the videos at each of these temporal checkpoints, when promoted with the three schedules. Here, we set each simulated video at median (50%) popularity and mandate that it remain at the median at 30, 60, 90, and 120 days if there were no promotions. The three videos have similar viral potential ($\nu \simeq 800$), but with very different maturity time t^* (shown in Fig. 15 (top right)): $t^* = 1360$ days for v_1 , $t^* = 113$ days for v_2 and $t^* = 7$ days for v_3 . For video v_1 , the slow unfolding of promotion effect makes the early schedule the most effective as the other schedules never catch-up. For video v_2 , the even schedule amounts to the same popularity percentile at 120 days as the early schedule. Not all promotions in the even schedule completely unfold, and the return of the two schedules is not exactly the same, however the 80% popularity bin covers volumes of view between 574,007 and 762,478 views, and both videos get assigned to it. Notice also that the early promotion achieves a popularity of 82% at 60 and 90 days, but is degraded to 80% at 120 days. This is because, after t^* days, promotions stop generating new views, and the video’s popularity slips. This phenomenon is more obvious for v_3 , for which the even schedule catches-up with early at 90 days, and all three schedules achieve the same popularity percentile at 120 days, but for different costs.

We conclude that promotion strategies should vary for videos of different maturity time: late schedule is the most economical for videos whose popularity unfolds fast enough; for videos with a long maturity time, early promotion is costly but will lead to the most attention.

4.5.2 Promotion and perceived potential

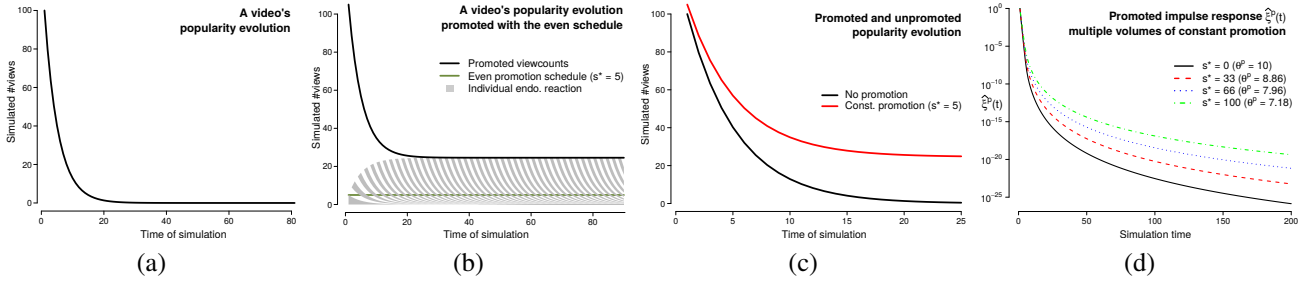


Figure 16: Effects of the even promotion schedule: apparent increase of video viral potential. (a) An example viewcount evolution; (b) the same evolution under an even promotion schedule with $s^* = 5$ promotions daily. (c) Side-by-side plotting of the two evolutions. (d) Equivalent *promoted* input response $\hat{\xi}^p(t)$, for increasing amounts of constant promotion.

In this section, we study in more detail the **even** promotion schedule. Despite not being the most advantageous in terms of cost or popularity unfolding, even promotion schedules are worth further examination, since they are very widely used in commercial settings – the car advertisement before the prime-time TV show every evening, for example. The process is shown in Fig. 16: we show in graph (a) the example of a common popularity pattern of Youtube videos: the daily viewcount has an initial peak, followed by a steady decay. Graph (b) shows the same system constantly promoted with a daily promotion of $s(t) = s^* = 5, \forall t \in 1, 2, \dots, 90$. With $t^* = 14$, the number of additional views gained through promotion stabilizes after about 20 days. The number of additional daily views is νs^* due to the LTI property (Sec. 1.2.3), as shown with the sliced graph [Rizoiu et al., 2017]. Fig. 16c contrasts the promoted and the unpromoted relaxation; we see that the daily views in the promoted system have a slower decrease and converge to a non-zero value – the evenly promoted system appears to possess a longer memory. We refit the HIP model parameters of the promoted system with varying s^* (without informing HIP about the promotion), as shown in Fig. 16d. We observe that the equivalent memory exponent θ^p is smaller and the *promoted* impulse response $\hat{\xi}^p(t)$ has a slower decay. Both are indicative of a longer system memory, and a higher viral potential. This provides a plausible explanation for the widely used even promotion schedule: by constantly injecting promotion into a social system, the perceived viral potential appears increased and the video stays longer in the public attention, thus generating more views.

4.6 Results: analyzing *relative engagement*

In this section we study the behavior of the *relative engagement* over a large collections of videos and we show that it demonstrates correlation to external notions of video quality (Sec. 4.6), and is stable over time (Sec. 4.6.2).

4.6.1 *Relative engagement* and video quality

In this section, we show the relation between *relative engagement* and video quality. We plot comparatively, on the *engagement map*, the three QUALITY VIDEOS datasets defined in Sec. 2.1

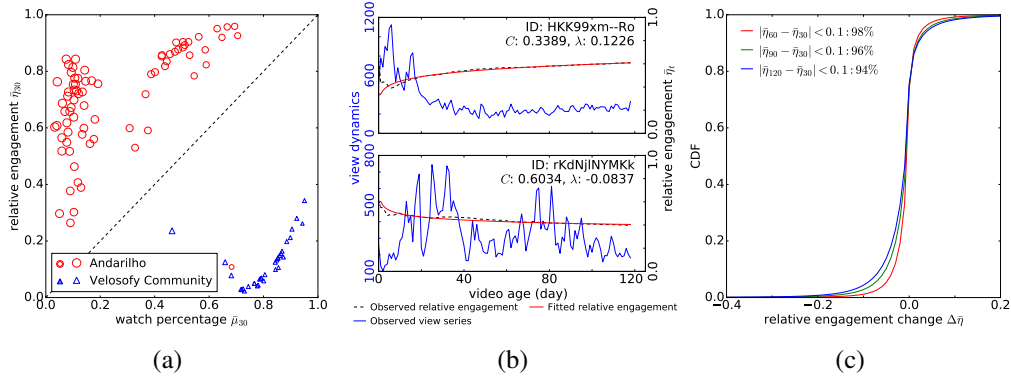


Figure 17: Stability of *relative engagement*. **(a)** Relative engagement and video quality for individual channels. **(b)** The views series (in blue) and the *relative engagement* over time, observed in dashed black line and fitted by a power-law of parameters C and λ in red; **(c)** The Cumulative Distribution Function of the change in *relative engagement* over a period of 30, 60 and 90 days (in red, green and blue respectively).

against their corresponding video categories in the TWEETED VIDEOS dataset. Fig. 9c plots the *engagement map* of all the Music videos in the TWEETED VIDEOS dataset (in blue shades), the *engagement map* of the videos in the VEVO VIDEOS dataset (in red shades) and projects the videos in the BILLBOARD VIDEOS as a scatter plot. Similarly, Fig. 9d plots the *engagements maps* of all the News videos in TWEETED VIDEOS (blue) and of the videos in TOP NEWS VIDEOS (in red), respectively. Visibly, the QUALITY VIDEOS are skewed towards higher *relative engagement* values in both figures. Most notably, 32 videos in the BILLBOARD VIDEOS dataset (51% of the dataset) possess a remarkably high *relative engagement* of over 0.95. This indicates a direct link between *relative engagement* and video quality, as quantified by external domain experts.

Relative engagement across channels The *relative engagement* appears to be consistent across videos of the same channel. Fig. 17a shows the example of two channels. Videos from the *Andarilho* channel (red circles) have an average video *watch percentage* of 0.25. Those from the *Velosofy Community* channel (blue triangles) has a higher average *watch percentage* of 0.79. However, merely accounting for *watch percentage* is biased by video duration (shown by the size of the circles and triangles). The videos of *Velosofy Community* are in average much shorter than those of *Andarilho*. The two channels have average *relative engagement* of 0.71 and 0.14, respectively.

4.6.2 Temporal dynamics of *relative engagement*

As argued, good engagement measures should be stable throughout the videos' lifetime. In this section, we study the stability of *relative engagement* and we predict its dynamics over time. Note that for a given video, the relation between $\bar{\mu}$ and $\bar{\eta}$ is constant throughout its lifetime. Consequently, all the conclusions of this section apply to both *average watch time* and *relative engagement*.

Relative engagement is stable over time. Fig. 17 (left) shows the dynamics of views and of *relative engagement* over a period of 120 days, for two example videos. While the views series show complex evolutions, with multiple highs and lows, the *relative engagement* is remarkably stable with only a slight positive trend for video *HKK99xm-Ro* and a slight negative trend for *rKdNjI NYMKk*. The same stability can be observed for all the videos in the TWEETED VIDEOS

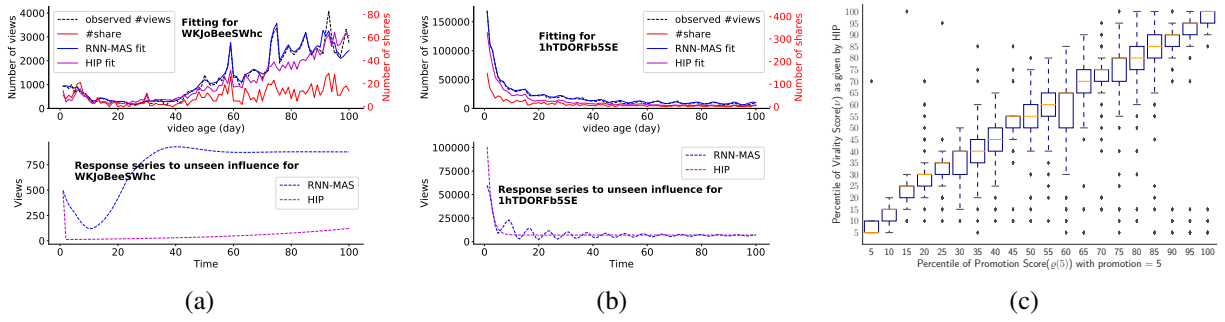


Figure 18: Figure (a) and (b) sample fitting for RNN-MAS and HIP for video *WKJoBeeSWhc* and *1hTDORFb5SE* in top row and response series to unseen influence in bottom row for RNN-MAS and HIP. (c) virality ν of HIP against response to a unit promotion $\rho(p)$ for RNN-MAS with $p = 5$. ν and $\rho(5)$ are binned into percentile popularity of 5% each.

dataset. Fig. 17 (right) plots on the x-axis the absolute change in *relative engagement* between day 30 and days 60, 90 and 120 respectively; and the Cumulative Distribution Function on the y-axis. The figure reveals two observations: first, more than 94% of the videos observe a change of *relative engagement* of less than 0.1 between day 30 and day 120; second, there are more videos for which the *relative engagement* decreases over time than those for which it increases.

Fitting *relative engagement* dynamics. We set out to model and predict *relative engagement* over time. The power-law function has shown good results in modeling the dynamics of human-related measures [Crane and Sornette, 2008b, Yu et al., 2015, Mishra et al., 2016b]. We also consider modeling the evolution of *relative engagement* using a simple linear model or approximating it with a constant value. For every video in the TWEETED VIDEOS dataset, we fits the *relative engagement* dynamics during the first 120 days using each of the three models, and we report the Mean Absolute Error (MAE). The power-law decay models best the dynamics of *relative engagement*, with a mean MAE of 0.0119.

4.7 Result: grassroot users vs. celebrities

In this section, we exemplify how the response series to unseen influence and metrics *promotion score* and *loudness level* described in Section 3.3 can be used to analyze the popularity of online videos.

4.7.1 Understanding unseen influence

We present case studies on real and simulated data to illustrate various complex dynamics captured by response series to unseen influence for our models when compared to HIP.

Real data fittings. We show sample fittings for two videos *WKJoBeeSWhc* and *1hTDORFb5SE* in the top row of Figure 18a and Figure 18b respectively. Bottom row in both figures show the response series of unseen influence for RNN-MAS and HIP. We observe response series for RNN-MAS captures more dynamic trends, a delayed effect for *WKJoBeeSWhc* and a seasonal trend for *1hTDORFb5SE*, which HIP can't. The gain in fitting performance for RNN-MAS can be attributed to better modelling of response to unseen influence.

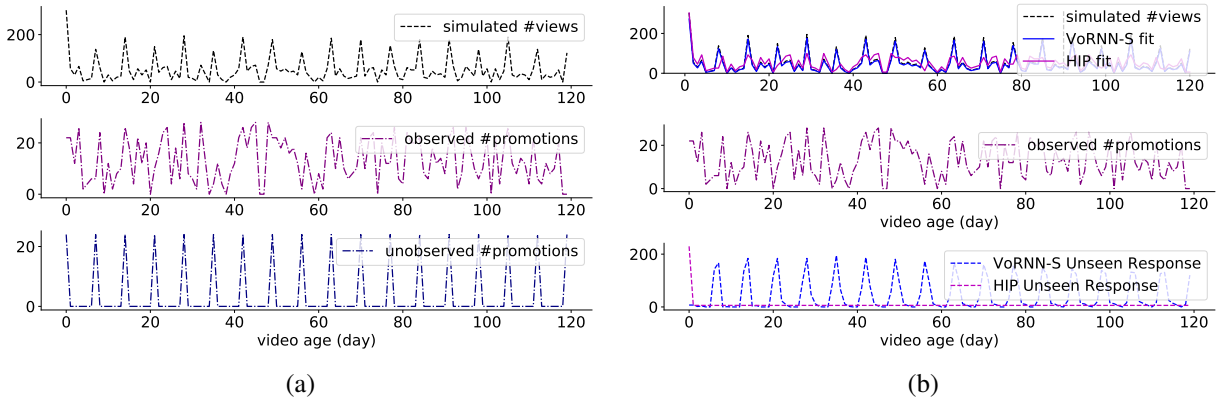


Figure 19: Figure (a) Top row is a simulated view series from HIP with two external sources: simulated random promotion shown in middle row and simulated cyclic promotion with weekly spikes shown in bottom row. (b) Top row shows fitting of VoRNN-S and HIP on simulated view series (random promotion is observed by both models whereas cyclic series is hidden) . Bottom row shows the response series to unseen influence for both models. Response series of VoRNN-S recovers the weekly cycles in unobserved promotion.

Synthetic data. We use synthetic time series to understand the ability of volume RNN models in recovering important seasonal trends. We simulate two sources of promotion for 120 days: a series with weekly cyclic component of 24 units of promotion, shown in bottom row of Figure 19a; a series with random promotions, shown in middle row of Figure 19a. We now simulate a view series from HIP using both sources as external stimulus, shown in top row of Figure 19a. We adopt the fitting procedure in Section 4.7.2 to train HIP and VoRNN-S on simulated view series using only random promotion as the external stimulus. Exact details for simulation can be seen in supplementary information of our paper [Mishra et al., 2018].

Fitting results for both the models are shown in the top row of Figure 19b. We can see the fitting of VoRNN-S (blue line) follows the simulated views (dashed black lines) more closely than the HIP’s fit (magenta line). The response series for VoRNN-S and HIP is shown in bottom row of Figure 19b with colors blue and magenta respectively. The estimated response from VoRNN-S shows a clear seasonal component. HIP failed to capture this due to its simplistic parametric assumptions about external influence.

4.7.2 Average unit promotion score

HIP was used by Rizoio and Xie [2017] to compute virality score (ν), for capturing total attention generated by a unit promotion. We compare the promotion score $\varrho(p)$ for our models against ν . The change in percentile rank for promotion score $\varrho(p)$ of RNN-MAS with $p = 5$ is plotted against ν in Figure 18c. We observe strong correlation between relative gain in promotion for non-viral videos (lower percentiles like 5%, 10%) and highly viral videos (90% and above), indicating strong performance in modeling effects of promotions on popularity. We also observe decrease in $\varrho(p)$ as p increases (e.g. median for $\varrho(1) \approx 263$, $\varrho(5) \approx 260$, $\varrho(10) \approx 259$), pointing towards modeling of diminishing returns Jones [1990] by RNN-MAS. HIP can’t model diminishing returns because of its LTI property.

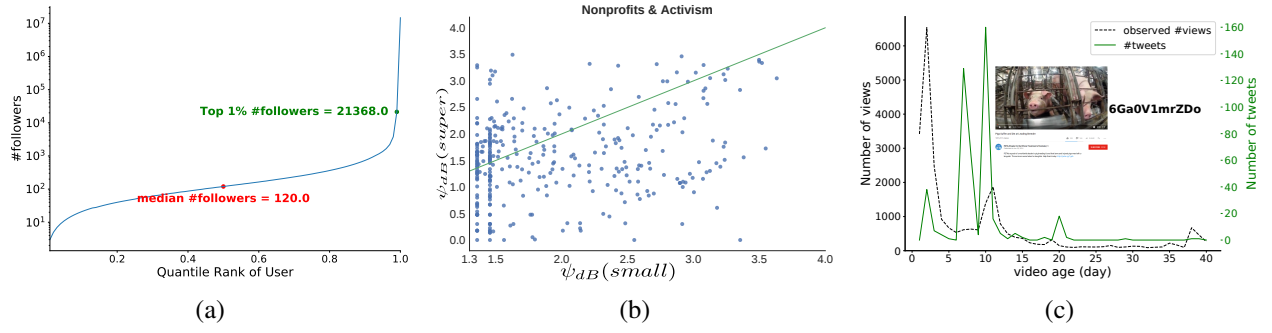


Figure 20: (a) Quantile rank versus number of followers of an user on ACTIVETWEETS. The median user has 120 followers whereas a user with 21368 followers is in top 1%. (b) $\psi_{dB}(super)$ versus $\psi_{dB}(small)$ where super is a series of events generated by a super user(followers=21368) and small is series of events generated by cohort (178) of small users(followers=120) for the category *Nonprofits & Activism*. For 37% of super users have greater influence than a cohort of small users. (c) view and tweet series for video 6Ga0V1mrZDo, where $\psi_{dB}(super)$ is 3.3 whereas as $\psi_{dB}(small)$ is 1.3. Only first and third peak of tweet series correlates with peaks in view series, as we have a super user with 494851 and 23561 followers around those peaks respectively.

4.7.3 User Influence

Setup. Bakshy et al. [2011] showed for promoting content on Twitter the most reliable and cost-effective method is to utilize a group of individuals who have average or even less number of followers. We evaluate the influence of a super user against a cohort of small users. As per Figure 20a, a super user is estimated to be one with #followers=21,368 which puts him in top 1% of all users, whereas a small user is the median user with #followers=120. The cohort has 178 users as in expectation they generate same user influence (#followers) as the super user. Now we generate a series of tweets from a trained RPP model for a video as promotions for our user groups. The effect for promotion is captured by metric ψ_{dB} for each group as per the procedure in Section 3.3. To contrast the effects of both user groups we plot ψ_{dB} for each of them against each other for every video, for example Figure 20b. For more details on setup refer supplementary information of our paper [Mishra et al., 2018].

Results. Figure 20b shows the value of ψ_{dB} for super user against a cohort of small users on all videos in *Nonprofits & Activism* category. We plot similar plots for all categories in the supplementary information of our paper [Mishra et al., 2018]. We observe type of content greatly affects the promoting capabilities of the super user. For example in Figure 20b 37% of videos (above the green line) of category *Nonprofits & Activism* have ψ_{dB} of super user is greater than ψ_{dB} of cohort of small users. Whereas for categories *HowTo & Style* and *Gaming* more than 50% of videos have ψ_{dB} greater for super user when compared to a cohort of small users, refer to supplementary information of our paper [Mishra et al., 2018].

We present a real video (6Ga0V1mrZDo) from *Nonprofits & Activism* category in Figure 20c, where $\psi_{dB}(super)$ is 3.3 whereas as $\psi_{dB}(small)$ is 1.3. The video here is uploaded by PETA, talking about deaths of pigs in slaughter houses. In Figure 20c not all peaks in tweets corresponds towards the increase in views of the video. On examining the real data we find that peaks where views and tweets correlates are the points where a super user is part of the tweets. In first peak

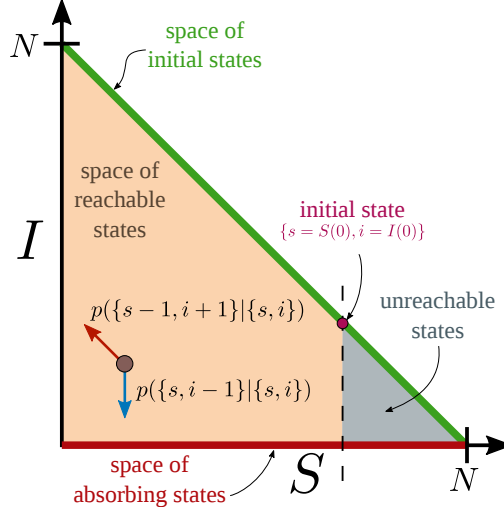


Figure 21: Visualization of the space of states of the SIR bivariate Markov Chain $\{s, i\}$. $s, i \in \mathbb{N}$ and $s + i \leq N$, in other words the space of valid states sits under the green line $s + i = N$. The initial state in an SIR epidemic is always on the **green line** ($S(0) + I(0) = N$). Given an initial state $\{s = S(0), i = I(0)\}$ (shown by the **magenta circle**), the **orange area** shows the space of reachable states and the gray area depicts the *unreachable states*. From a state $\{s, i\}$ the system can to $\{s - 1, i + 1\}$ (new infection); and to $\{s, i - 1\}$ (new recovery). The absorbing states ($\{s, 0\}$) is shown with a **red line**.

around 3rd day we have a user with 494851 followers and in third peak around 12th day we have a user with 23561 followers, whereas at two other peaks there are cohort of small users tweeting about video with none with more than 4000 followers. Presence of super user around peaks in view series supports our estimate of super user being more effective than a cohort of small users for promoting the given video.

4.8 Result: explaining popularity unpredictability

In this section, we detail a technique for computing the final size distribution of a partially observed cascade, and we provide a new result concerning the unpredictability of online popularity. In Sec. 4.8.1 we review known results on the final size distribution of an SIR epidemic. In Sec. 4.8.2 we employ the equivalence shown in Theorem 1.2 to compute the final size distribution of a cascade modeled with HawkesN. Finally in Sec. 4.8.3, we profile cascade size distributions and we provide a new explanation for the perceived popularity unpredictability.

4.8.1 Epidemic size distributions in SIR

The final size of an infection is defined as the total number of individuals that have been infected (and recovered) during the epidemic. Estimating the final size while the epidemic is in its early stages is a well-studied problem in epidemiology. In this section, we review a solution to this problem for the stochastic SIR model.

SIR as a Markov chain. The stochastic SIR introduced in Sec. 1.3.1 can be formulated as a bivariate continuous time-homogeneous Markov chain [Allen, 2008]. Each state $\sigma \in \Sigma$ is uniquely

defined by the ordered pair of random variables $\{S_t, I_t\}$, denoting the sizes of the susceptible and the infected populations. Σ is the finite space of all possible states, visually represented in Fig. 21 as a triangle on the two-dimensional surface with the number of susceptible on the x-axis and the number of infected on the y-axis. From a given state $\{S_t = s, I_t = i\}$ (denoted from here on as $\{s, i\}$), there are only two other states in which the system can transition depending on the type of event that occurs: a new infection arrives, and the system transitions $\{s, i\} \rightarrow \{s - 1, i + 1\}$ with the probability $p(\{s - 1, i + 1\}|\{s, i\})$; or a new recovery arrives, and the system transitions $\{s, i\} \rightarrow \{s, i - 1\}$ with the probability $p(\{s, i - 1\}|\{s, i\})$ (shown in the figure by the red and blue arrows, respectively).

Transition matrix and probabilities. The epidemic ends when $I_t = 0$, i.e. no more infectious individuals exist to propagate the epidemic. Consequently, the states $\{s, 0\}$ are *absorbing states* – once the system arrives in one of these states, it does not transition to any other state. From a non-absorbing state $\{s, i\}$, new infections are observed at the rate $\frac{\beta}{N}si$ and new recoveries at the rate of γi (see Eq. (19)). We obtain the transition probabilities:

$$\begin{aligned} p(\{s - 1, i + 1\}|\{s, i\}) &= \frac{\frac{\beta}{N}si}{\frac{\beta}{N}si + \gamma i} = \frac{\beta s}{\beta s + N\gamma}, \text{ for } i > 0 \\ p(\{s, i - 1\}|\{s, i\}) &= \frac{\gamma i}{\frac{\beta}{N}si + \gamma i} = \frac{N\gamma}{\beta s + N\gamma}, \text{ for } i > 0. \end{aligned} \quad (44)$$

Suppose that the states in Σ are ordered from 1 to $|\Sigma|$. We define the *transition matrix* $T = [t_{kl}]$ of size $|\Sigma| \times |\Sigma|$, where $t_{kl} = p(\sigma_k = \{s_k, i_k\}|\sigma_l = \{s_l, i_l\})$ is the probability of transitioning from state σ_l to state σ_k . From Eq. (44) we obtain:

$$t_{kl} = \begin{cases} \frac{\frac{\beta s_l}{\beta s_l + N\gamma}}{\frac{\beta s_l}{\beta s_l + N\gamma} + \frac{N\gamma}{\beta s_l + N\gamma}} & , s_k = s_l - 1, i_k = i_l + 1, i_l > 0 \\ \frac{\frac{N\gamma}{\beta s_l + N\gamma}}{\frac{\beta s_l}{\beta s_l + N\gamma} + \frac{N\gamma}{\beta s_l + N\gamma}} & , s_k = s_l, i_k = i_l - 1, i_l > 0 \\ 1 & , s_k = s_l, i_k = i_l = 0 \\ 0 & , \text{otherwise} \end{cases}. \quad (45)$$

Note that the sum of each column j in the transition matrix M is equal to 1, as it contains the probabilities of transitioning from σ_l to another state.

Probability state vector and size distribution. Let π be the probability state vector $\pi \in \mathbb{R}^{|\Sigma| \times 1}$, with the l^{th} position of π giving the probability that the system is currently found in state σ_l . Let $\pi_0 = [0, \dots, 0, 1, 0, \dots, 0]$ be the initial probability vector, with the value of 1 corresponding to $\tilde{\sigma}$ the *initial state*, and zero everywhere else. Starting from π_0 , we compute the probability state vector after one transition as $\pi^{(1)} = M \times \pi_0$. $\pi^{(2)} = M^2 \times \pi_0$ gives the probabilities after two transitions, $\pi^{(3)} = M^3 \times \pi_0$ after three etc. Given that there are at most $N - 1$ infection events and N recovery events in an SIR realization, the system is guaranteed to converge after $2N - 1$ steps [Allen, 2008]. At convergence, all states except the absorbing states have a probability of zero in $\pi^{(2N-1)}$. We denote as $P(s)$ the value in $\pi^{(2N-1)}$ for the state $\{s, 0\}$, $s = 0, 1, \dots, N - I_0$.

From an initial state $\tilde{\sigma}$ the system can finish in any of the absorbing states $\{s, 0\}$ with the probability $P(s)$. The distribution of final size of the diffusion is that of the random variable $N' = N - s$.

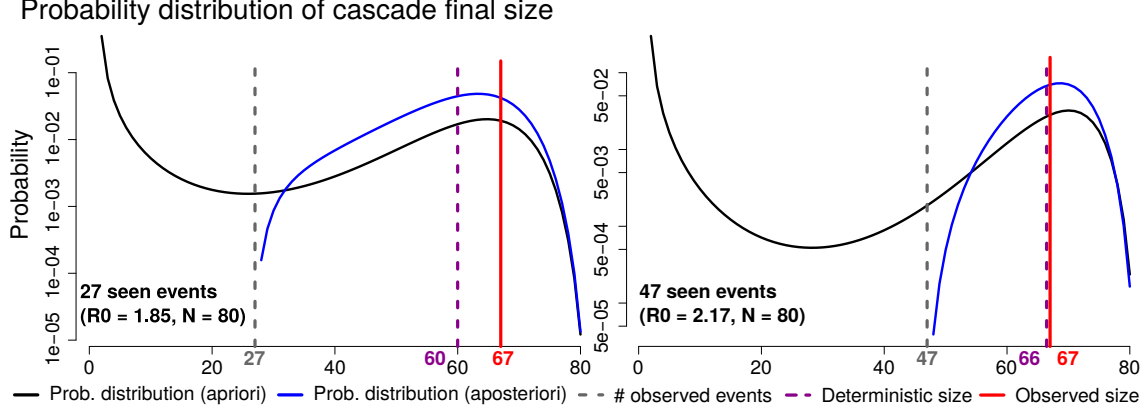


Figure 22: Final size probability distribution for a NEWS cascade. HawkesN was fitted on the first 27 events (*left*) and 47 events (*right*). The black line shows the *apriori* distribution, after seeing a single event. the blue line shows the *aposteriori* distribution, after seeing the observed events. Shown with vertical lines: the number of observed events in gray dashed, the deterministic prediction in magenta dashed and the actual observed cascade size in red.

4.8.2 Cascade size distribution in HawkesN

Given Theorem 1.2, computing the probability distribution over the final size of the cascade after observing an arbitrary number of events, conceptually amounts to changing the initial state $\tilde{\sigma} = \{\tilde{s}, \tilde{i}\}$ and using the method described in Sec. 4.8.1. Suppose we observed l events in HawkesN, we have $\tilde{s} = N - l$. The recovery events are not observed and the exact size of the infected population $\tilde{i}$ is not known. We compute its expectation over times to recovery $\mathcal{T}$ as:

$$\mathbb{E}_{\mathcal{T}}[\tilde{i}] = \mathbb{E}_{\mathcal{T}} \left[\sum_{j=1}^l \mathbb{1}(t_j^I + \tau_j > t_l) \right] \stackrel{\text{Eq. (23), (24)}}{=} \sum_{j=1}^l e^{-\gamma(t_l - t_j^I)} \quad (46)$$

and we run the method in Sec. 4.8.1 starting from the initial state $\tilde{\sigma} = \{N - l, \mathbb{E}_{\mathcal{T}}[\tilde{i}]\}$. $\mathbb{E}_{\mathcal{T}}[\tilde{i}]$ in Eq. (46) is a real number, that we round to the closest integer.

In our discussion in Sec. 4.8.3, we study two probability distributions: the *apriori* distribution is computed starting from the initial state $\tilde{\sigma} = \{N - 1, 1\}$ and it is dependent on model parameters only. The *aposteriori* distribution is the size distribution after observing l events, it is computed starting from $\tilde{\sigma} = \{N - l, \mathbb{E}_{\mathcal{T}}[\tilde{i}]\}$ and it is dependent on model parameters and the observed event times $t_1, t_2, \dots, t_l$ (cf. Eq. (46)).

4.8.3 Explaining popularity unpredictability

In this section, we study the probability distribution of population size for real-life cascades. Both left and right plots in Fig. 22 show the same cascade from the NEWS dataset, with the HawkesN parameters fit on 27 and 47 events respectively (here N is a meta-parameter fixed at $N = 80$). The *apriori* probability size distribution – the distribution after observing only the first event – shows two maxima: one around very small values of cascade size, and one around ~ 65 . This provides the following explanation for the general perceived unpredictability of online popularity. For cascades showing a bi-modal *apriori* size distribution, there are two likely outcomes: either it

dies out early or it reaches a large size compared to the maximum population N . At time $t = 0$ is it impossible to differentiate between the two outcomes. The situation is different after observing a number of events. The *a posteriori* probability distribution (shown in Fig. 22 with a blue line) reflects the information gained from the observed events and it shows a single maximum towards the higher size values. The more events we observe, the higher the likelihood of the true value of cascade size. We also observe that the size distribution gets narrower as we observe more events, i.e. there is less uncertainty in cascade size prediction. This provides another explanation to why autoregressive popularity prediction approaches [Chang et al., 2014, Pinto et al., 2013a, Szabo and Huberman, 2010a] achieve higher results. Online popularity has been previously claimed to be unpredictable [Martin et al., 2016], however as far as we know this is the first explanation for it, based on analytical results on size distributions.

References

- Vevo in Wikipedia, 2017. URL <https://en.wikipedia.org/wiki/Vevo>.
- Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Spatio-temporal models for estimating click-through rate. In *WWW '09*, pages 21–30, 2009.
- Linda J. S. Allen. An Introduction to Stochastic Epidemic Models. In *Mathematical Epidemiology*, chapter 3, pages 81–130. Springer, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-78911-6_3. URL http://link.springer.com/10.1007/978-3-540-78911-6_3.
- Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone’s an influencer. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, WSDM '11, page 65, New York, New York, USA, February 2011. ACM Press. ISBN 9781450304931. doi: 10.1145/1935826.1935845.
- Peng Bao, Hua-Wei Shen, Xiaolong Jin, and Xue-Qi Cheng. Modeling and Predicting Popularity Dynamics of Microblogs using Self-Excited Hawkes Processes. In *WWW '15*, pages 9–10, 2015. ISBN 9781450334730. doi: 10.1145/2740908.2742744.
- Georgiy V. Bobashev, D. Michael Goedecke, Feng Yu, and Joshua M. Epstein. A hybrid epidemic model: Combining the advantages of agent-based and equation-based approaches. In *Proceedings - Winter Simulation Conference*, pages 1532–1537. IEEE, 2007. ISBN 1424413060. doi: 10.1109/WSC.2007.4419767. URL <http://ieeexplore.ieee.org/document/4419767/>.
- Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system. In *IMC '07*, pages 1–14. ACM, 2007.
- Biao Chang, Hengshu Zhu, Yong Ge, Enhong Chen, Hui Xiong, and Chang Tan. Predicting the Popularity of Online Serials with Autoregressive Models. In *CIKM '14*, pages 1339–1348, nov 2014. ISBN 9781450325981. doi: 10.1145/2661829.2662055.
- J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv e-prints*, December 2014.

- Aaron Clauset, Cosma Rohilla Shalizi, and M E J Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, November 2009. ISSN 0036-1445. doi: 10.1137/070710111.
- Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ, 2nd edition, 1988. ISBN 978-0805802832. URL <http://www.amazon.com/Statistical-Analysis-Behavioral-Sciences-Edition/dp/0805802835>.
- Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS* '08, 105(41):15649–15653, oct 2008a. ISSN 0027-8424. doi: 10.1073/pnas.0803685105.
- Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS*, 105(41):15649–15653, 2008b. doi: 10.1073/pnas.0803685105.
- D J Daley and D Vere-Jones. *An introduction to the theory of point processes. Vol. I*, volume I. 2008. ISBN 978-0-387-21337-8. doi: 10.1007/b97277. URL <http://www.springer.com/us/book/9780387955414>.
- Wanying Ding, Yue Shang, Lifan Guo, Xiaohua Hu, Rui Yan, and Tingting He. Video Popularity Prediction by Sentiment Propagation via Implicit Network. In *CIKM '15*, pages 1621–1630, oct 2015. ISBN 978-1-4503-3794-6. doi: 10.1145/2806416.2806505. URL <http://dl.acm.org/citation.cfm?id=2806416.2806505>.
- Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. Understanding the impact of video quality on user engagement. In *ACM SIGCOMM Computer Communication Review*, volume 41, pages 362–373. ACM, 2011.
- Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD '16*, pages 1555–1564, 2016.
- Jeffrey L. Elman. Finding structure in time. *COGNITIVE SCIENCE*, 14(2):179–211, 1990.
- Flavio Figueiredo, Jussara M Almeida, Fabrício Benevenuto, and Krishna P Gummadi. Does content determine information popularity in social media?: A case study of youtube videos' content and their popularity. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 979–982. ACM, 2014.
- Flavio Figueiredo, Jussara M Almeida, Marcos A Gonçalves, and Fabricio Benevenuto. Trendlearner: Early prediction of popularity trends of user generated content. *Information Sciences*, 349:172–187, 2016.
- Vladimir Filimonov and Didier Sornette. Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *arXiv preprint*, aug 2013. URL <http://arxiv.org/abs/1308.6756>.
- Shuai Gao, Jun Ma, and Zhumin Chen. Modeling and Predicting Retweeting Dynamics on Microblogging Platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, pages 107–116, New York, New York, USA, feb 2015. ACM Press. ISBN 9781450333177. doi: 10.1145/2684822.2685303. URL <http://dl.acm.org/citation.cfm?id=2684822.2685303>.
- A. Graves. Generating Sequences With Recurrent Neural Networks. *ArXiv e-prints*, August 2013.
- Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1): 83–90, 1971. ISSN 0006-3444. doi: 10.1093/biomet/58.1.83.

- Alan G. Hawkes and David Oakes. A Cluster Process Representation of a Self-Exciting Process. *JSTOR: Journal of Applied Probability*, 11:493–503, 1974.
- Agnès Helmstetter and Didier Sornette. Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research: Solid Earth*, 107(B10):ESE 10–1–ESE 10–21, 2002. ISSN 2156-2202. doi: 10.1029/2001jb001580.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL <http://dx.doi.org/10.1162/neco.1997.9.8.1735>.
- JP Jones. The double jeopardy of sales promotions. *Harvard business review*, 68(5):145—152, 1990. ISSN 0017-8012. URL <http://europaemc.org/abstract/MED/10113338>.
- Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *PNAS*, 112(24):7426–7431, 2015.
- Ryota Kobayashi and Renaud Lambiotte. TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. In *ICWSM 2016*, number ICWSM, 2016. ISBN 9781577357582.
- Quyu Kong, Marian-Andrei Rizoïu, Siqi Wu, and Lexing Xie. Will this video go viral? explaining and predicting the popularity of youtube videos. In *WWW 2018 Companion: The 2018 Web Conference Demos*, Lyon, France, 2018. doi: 10.1145/3184558.3186972. URL <https://arxiv.org/pdf/1801.04117.pdf>.
- Coco Krumme, Manuel Cebrian, Galen Pickard, and Sandy Pentland. Quantifying social influence in an online cultural market. *PloS one*, 7(5):e33785, 2012.
- Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
- Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, aug 1989. ISSN 0025-5610. doi: 10.1007/BF01589116.
- Travis Martin, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. Exploring Limits to Prediction in Complex Social Systems. In *Proceedings of the 25th International Conference on World Wide Web*, pages 683–694, feb 2016. doi: 10.1145/2872427.2883001. URL <http://arxiv.org/abs/1602.01013><http://dx.doi.org/10.1145/2872427.2883001>.
- Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: Model and implications. *KDD '12*, 2012.
- Swapnil Mishra, Marian-Andrei Rizoïu, and Lexing Xie. Feature Driven and Point Process Approaches for Popularity Prediction. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM '16*, page 10, 2016a. doi: 10.1145/2983323.2983812.
- Swapnil Mishra, Marian-Andrei Rizoïu, and Lexing Xie. Feature Driven and Point Process Approaches for Popularity Prediction. In *Proceedings of the 25th ACM International Conference on Conference on Information and Knowledge Management, CIKM '16*, Indianapolis, IN, USA, 2016b. doi: 10.1145/2983323.2983812.

- Swapnil Mishra, Marian-Andrei Rizoïu, and Lexing Xie. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, pages 1–10, Stanford, CA, USA, 2018. URL <https://arxiv.org/abs/1804.02101>.
- AV Oppenheim, AS Willsky, and SH Nawab. *Signals and Systems*. Prentice Hall, 1997.
- Minsu Park, Mor Naaman, and Jonah Berger. A data-driven study of view duration on youtube. In *ICWSM*, pages 651–654, 2016.
- Henrique Pinto, Jussara M. Almeida, and Marcos A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *WSDM '13*, page 365. ACM Press, feb 2013a. ISBN 9781450318693. doi: 10.1145/2433396.2433443. URL <http://dl.acm.org/citation.cfm?id=2433396.2433443>.
- Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *WSDM '13*, pages 365–374. ACM, 2013b.
- Marian-Andrei Rizoïu and Lexing Xie. Online popularity under promotion: Viral potential, forecasting, and the economics of time. In *The International AAAI Conference on Web and Social Media (ICWSM)*, pages 1–10, Montreal, Canada, 2017. URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15553>.
- Marian-Andrei Rizoïu, Lexing Xie, Tiberio Caetano, and Manuel Cebrian. Evolution of Privacy Loss on Wikipedia. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining, WSDM '16*, San Francisco, CA, USA, 2016. doi: 10.1145/2835776.2835798.
- Marian-Andrei Rizoïu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be HIP: Hawkes intensity processes for social media popularity. In *World Wide Web 2017, International Conference on, WWW '17*, pages 735–744, Perth, Australia, 2017. doi: 10.1145/3038912.3052650. URL <https://arxiv.org/pdf/1602.06033>.
- Marian-Andrei Rizoïu, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie. #DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, pages 1–10, Stanford, CA, USA, 2018a. URL <https://arxiv.org/abs/1802.09808>.
- Marian-Andrei Rizoïu, Young Lee, Swapnil Mishra, and Lexing Xie. Hawkes processes for events in social media. In Shih-Fu Chang, editor, *Frontiers of Multimedia Research*, pages 191–218. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA, 2018b. doi: 10.1145/3122865.3122874. URL <https://arxiv.org/pdf/1708.06401.pdf>.
- Marian-Andrei Rizoïu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference, WWW '18*, Lyon, France, 2018c. doi: 10.1145/3178876.3186108. URL <https://arxiv.org/pdf/1711.01679.pdf>.
- Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks. In *ICML'11*, pages 561–568, 2011. ISBN 978-1-4503-0619-5.
- Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

- Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. In *AAAI '14*, pages 291–297, 2014. ISBN 978-1-57735-661-5. URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8370>.
- Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communicat. of the ACM*, 53(8):80–88, 2010a.
- Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010b.
- 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*, 106(1):25–57, 2006.
- Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* '04, 160(6):509–16, sep 2004. ISSN 0002-9262. doi: 10.1093/aje/kwh255. URL <http://aje.oxfordjournals.org/content/160/6/509.full>.
- Yichen Wang, Xiaojing Ye, Haomin Zhou, Hongyuan Zha, and Le Song. Linking Micro Event History to Macro Prediction in Point Process Models. In *AISTATS*, pages 1375–1384, 2017.
- Richard Witt. *Arithmetical Questions*. 1613.
- Fang Wu and Bernardo A Huberman. Novelty and collective attention. *PNAS* '07, 104(45):17599–601, nov 2007. ISSN 0027-8424. doi: 10.1073/pnas.0704916104.
- Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond Views: Measuring and Predicting Engagement in Online Videos. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, Stanford, CA, USA, 2018. URL <https://arxiv.org/abs/1709.02541>.
- Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M Chu. Modeling the Intensity Function of Point Process Via Recurrent Neural Networks. *AAAI'17*, pages 1597–1603, 2017. URL <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14391>.
- Ping Yan. Distribution Theory, Stochastic Processes and Infectious Disease Modelling. In Wu J. Brauer F., van den Driessche P., editor, *Mathematical Epidemiology*, chapter 10, pages 229–293. Springer, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-78911-6_10. URL http://link.springer.com/10.1007/978-3-540-78911-6_10.
- Honglin Yu, Lexing Xie, and Scott Sanner. Twitter-driven youtube views: Beyond individual influencers. In *Proceedings of the ACM International Conference on Multimedia*, pages 869–872. ACM, 2014.
- Honglin Yu, Lexing Xie, and Scott Sanner. The Lifecycle of a Youtube Video: Phases, Content and Popularity. In *Ninth International AAAI Conference on Web and Social Media*, page 10, 2015. URL <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewPaper/10537>.
- Ali Zarezade, Utkarsh Upadhyay, Hamid Rabiee, and Manuel Gomez Rodriguez. RedQueen: An Online Algorithm for Smart Broadcasting in Social Networks. In *10th ACM International Conference on Web Search and Data Mining*, feb 2017. URL <http://arxiv.org/abs/1610.05773>.
- 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 '15*, 2015. ISBN 9781450336642.