Measuring collective attention in online content: Sampling, engagement, and network effects

Siqi Wu

PhD conclusion seminar
Aug 2020, ANU

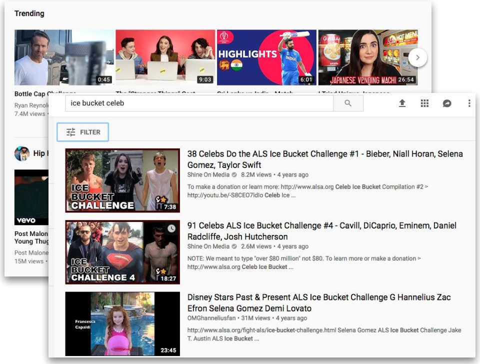
A pathway of consuming online content

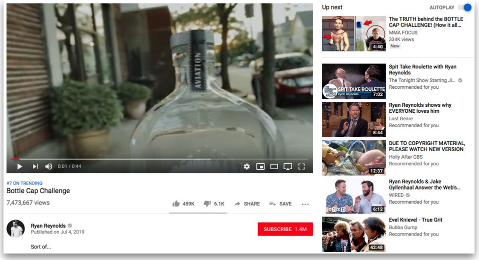
Social networking/community sites

twitter facebook reddit

Media sites [Krumme et al. '12]

YouTube Spotify goodreads





User-centric *Step 1: Which product should I **click**?* *Step 2: How should I **engage** with it?*

RQ1: Social data sampling
How does data sampling impact common measurements?

Item-centric

RQ2: Recsys network effects
How does the recommender systems drive user attention?

RQ3: Collective engagement
Why are some products more engaging?

Outline

1. Research scope
2. The effects of social data sampling
    - How are the tweets missing in the filtered stream?
    - What are the sampling effects on common measurements?
3. **The network effects induced by the recommender systems**
4. **The patterns of users engagement towards online content**
5. **Conclusion and looking ahead**

The effects of social data sampling

Paper:

[1] **Wu,** Rizoiu, and Xie. "Variation across Scales: Measurement Fidelity under Twitter Data Sampling." In *Proceedings of ICWSM, 2020.*

Dataset:

Complete/Sampled Retweet Cascades Datasets

Software:

Twitter-intact-stream: reconstructing the complete Twitter filtered stream

Variation across Scales: Measurement Fidelity under Twitter Data Sampling

Siqi Wu[1,3] and **Marian-Andrei Rizoiu**[2,3] and **Lexing Xie**[1,3]
[1]Australian National University, [2]University of Technology Sydney, [3]Data 61, CSIRO, Australia
{siqi.wu, lexing.xie}@anu.edu.au, marian-andrei.rizoiu@uts.edu.au

Abstract

A comprehensive understanding of data quality is the cornerstone of measurement studies in social media research. This paper presents in-depth measurements on the effects of Twitter data sampling across different timescales and different subjects (entities, networks, and cascades). By constructing complete tweet streams, we show that Twitter rate limit message is an accurate indicator for the volume of missing tweets. Sampling also differs significantly across timescales. While the hourly sampling rate is influenced by the diurnal rhythm in different time zones, the millisecond level sampling is heavily affected by the implementation choices. For Twitter entities such as users, we find the Bernoulli process with a uniform rate approximates the empirical distributions well. It also allows us to estimate the true ranking with the observed sample data. For networks on Twitter, their structures are altered significantly and some components are more likely to be preserved. For retweet cascades, we observe changes in distributions of tweet inter-arrival time and user influence, which will affect models that rely on these features. This work calls attention to noises and potential biases in social data, and provides a few tools to measure Twitter sampling effects.

1 Introduction

"Polls are just a collection of statistics that reflect what people are thinking in 'reality'. And reality has a well-known liberal bias." – Stephen Colbert[1]

Data quality is a timely topic that receives broad attention. The data noises and biases particularly affect data-driven studies in social media (Tufekci 2014; Olteanu et al. 2019). Overrepresented or underrepresented data may mislead researchers to spurious claims (Ruths and Pfeffer 2014). For example, opinion polls wrongly predicted the U.S. presidential election results in 1936 and 1948 because of unrepresentative samples (Mosteller 1949). In the era of machine learning, the data biases can be amplified by the subsequent models. For example, models overly classify agents doing cooking activity as female due to overrepresented correlations (Zhao et al. 2017), or lack the capacity to identify dark-skinned women due to underrepresented data (Buolamwini and Gebru 2018). Hence, researchers must be aware and take account of the hidden biases in their datasets for drawing rigorous scientific conclusions.

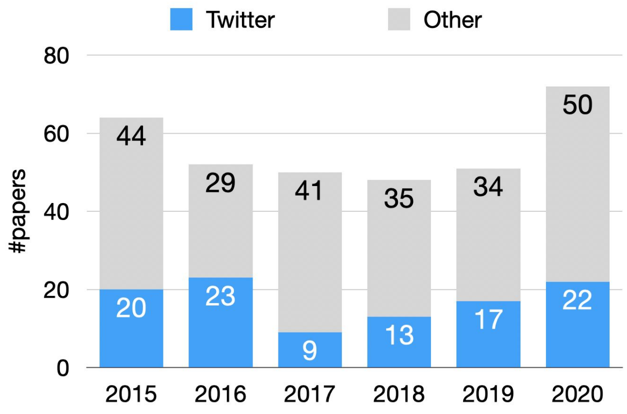
Twitter is the most prominent data source in ICWSM – 82 (31%) out of 265 full papers in the past 5 years (2015-2019) used Twitter data (listed in Section A of (Appendix 2020)), in part because Twitter has relatively open data policies, and in part because Twitter offers a range of public application programming interfaces (APIs). Researchers have used Twitter data as a lens to understand political elections (Bovet and Makse 2019), social movements (De Choudhury et al. 2016), information diffusion (Zhao et al. 2015), and many other social phenomena. Twitter offers two streaming APIs for free, namely *sampled* stream and *filtered* stream. The filtered stream tracks a set of keywords, users, languages, and locations. When the matched tweet volume is above a threshold, Twitter subsamples the stream, which compromises the completeness of the collected data. In this paper, we focus on empirically quantifying the data noises resulted from the sampling in the filtered stream and its impacts on common measurements.

This work addresses two open questions related to Twitter data sampling. Firstly, **how are the tweets missing in the filtered stream?** The sampling mechanism of the sampled stream has been extensively investigated (Kergl, Roedler, and Seeber 2014; Pfeffer, Mayer, and Morstatter 2018), but relatively little is said about the filtered stream. Since the two streaming APIs are designed to be used in different scenarios, it is pivotal for researchers who use the filtered stream to understand what, when, and how much data is missing. Secondly, **what are the sampling effects on common measurements?** Our work is inspired by Morstatter et al. (2013), who measured the discrepancies of topical, network, and geographic metrics. We extend the measurements to entity frequency, entity ranking, bipartite graph, retweet network, and retweet cascades. The answers to these questions not only help researchers shape appropriate questions, but also help platforms improve their data services.

We address the first question by curating two datasets that track suggested keywords in previous studies. Without leveraging the costly Twitter Firehose service, we construct the complete tweet streams by splitting the keywords and languages into multiple subcrawlers. We study the Twitter rate limit messages. Contradicting observations made by Sampson et al. (2015), our results show that the rate limit mes-

[1]At the 2006 White House Correspondents' Dinner.

Twitter data is prevailing, but it may get sampled



104 (31%) out of 337 ICWSM papers use Twitter data (2015-2020)

API	Search	Sampled streaming	Filtered streaming
Usage	Retrieving relevant tweets given a query	Streaming a sample of public tweets	Streaming matched tweets given a query
Rate limiting	180 or 450 calls / 15 minutes	Roughly 1% of all public tweets	50 tweets / 1 second
Affected studies	*Most*, since it only searches tweets of the past 7 days	*All*, by default roughly 1%	USC COVID-19: ~5% sampling rate [Chen et al. '20]

Q1. How are the tweets missing in the filtered stream?
Q2. What are the effects on common measurements?

Contribution: **a comprehensive measurement study of the Twitter sampling effects across different timescales and different subjects**

Twitter rate limit messages

- *Filtered streaming*: collecting tweets matching a set of prescribed predicates in realtime[1], e.g., "*COVID-19*"
- In each second, no more than 50 tweets will be returned[2].
- Rate limit messages indicate the cumulative number of missing tweets since the connection starts[3].

Blocks of streamed tweets

{"id_str":"1245501748485242881", ...}
{"limit":{"track":28469226,"timestamp_ms":"1585785737733"}}
{"id_str":"1245501752088150021", ...}

{"id_str":"1245501752968908802", ...}
{"limit":{"track":28469434,"timestamp_ms":"1585785738725"}}
{"id_str":"1245501756315860992", ...}

{"id_str":"1245501756987097089", ...}
{"limit":{"track":28469643,"timestamp_ms":"1585785739742"}}
{"id_str":"1245501760568995842", ...}

1 sec, 28469434 - 28469226 = 208 missing

1 sec, 28469643 - 28469434 = 209 missing

[1] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter
[2] https://developer.twitter.com/en/docs/labs/filtered-stream/faq
[3] https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/streaming-message-types

Constructing the complete filtered stream

- Strategy: splitting the filtering predicates into multiple subcrawlers.
- 2 datasets: Cyberbullying (sampling rate: 52.72%) and YouTube video sharing (91.53%).

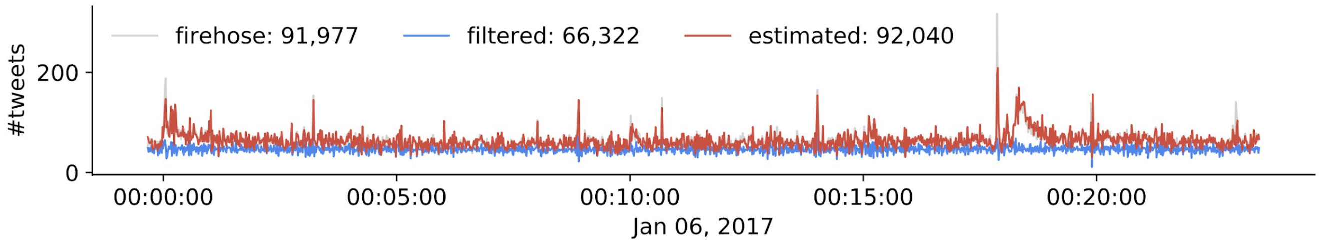
Id	Keywords	Languages	#collected tweets	#rate limit	#est. missing tweets	sampling rate
1	`should`	en	29,647,814	1,357	7,324	99.98%
2	`should`	all\en	801,904	0	0	100.00%
3	`live`	en	16,526,226	1,273	25,976	99.84%
4	`live`	all\en	7,926,325	233	7,306	99.91%
5	`kill`, `fight`, `poser`, `nerd`, `freak`, `pig`	all	15,449,973	16	108	100.00%
6	`dick`, `suck`, `gay`, `loser`, `whore`, `cunt`	all	13,164,053	15	125	100.00%
7	`pussy`, `fat`, `die`, `afraid`, `emo`, `slut`	all	21,333,866	89	1,118	99.99%
8	`bitch`, `wannabe`, `whale`, `slept`, `caught`	all	14,178,366	64	666	100.00%
complete	subcrawlers 1-8	all	114,488,537	3,047	42,623	99.96%
sample	all 25 keywords	all	60,400,257	1,201,315	54,175,503	52.72%

[1] Obtained via a Twitter data reseller https://discovertext.com/

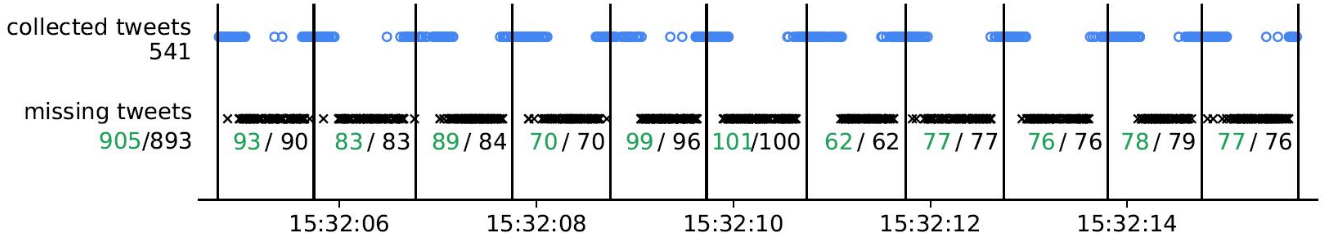
Constructing the complete filtered stream

- Strategy: splitting the filtering predicates into multiple subcrawlers.
- 2 datasets: Cyberbullying (sampling rate: 52.72%) and YouTube video sharing (91.53%).
- Validation: single crawler + rate limit messages *vs.* (1) Firehose stream[1] / (2) multiple subcrawlers.

vs. Firehose stream
MAPE: 0.007



vs. multiple subcrawlers
MAPE: 0.001

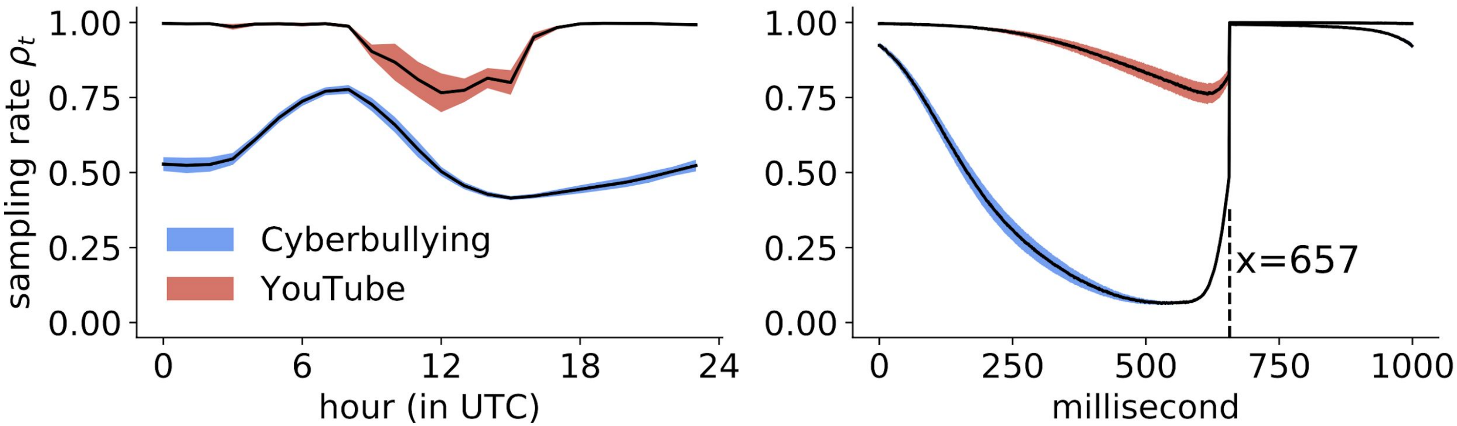


multiple subcrawlers ~= single crawler + rate limit messages ~= Firehose stream

[1] Obtained via a Twitter data reseller https://discovertext.com/

Temporal variation of sampling rates

- Sampling rates are uneven in different hours or in different milliseconds.



Outline

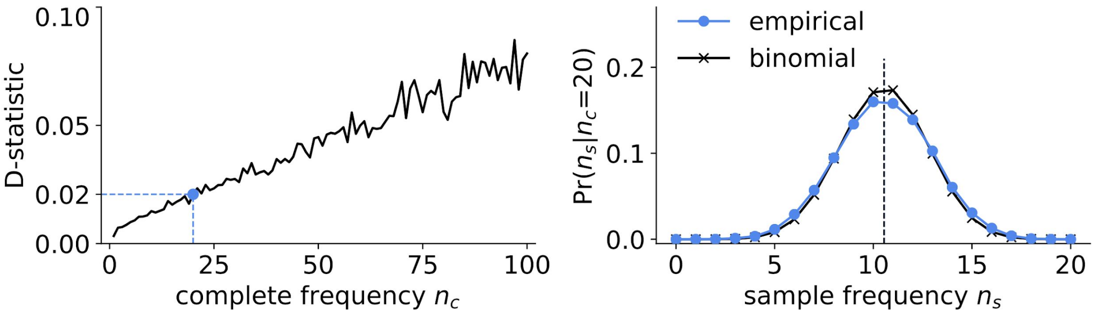
1. Research scope
2. **The effects of social data sampling**
   - How are the tweets missing in the filtered stream?
   - What are the sampling effects on common measurements?
3. **The network effects induced by the recommender systems**
4. **The patterns of users engagement towards online content**
5. **Conclusion and looking ahead**

Using Bernoulli process with a uniform rate to approximate the empirical data

- Assumption used in prior studies but not validated [Joseph et al. '13, Pfeffer et al. '18].
- *metric*: D-statistic [Leskovec and Faloutsos '06]. $D(G, G') = \max_x\{|G(x) - G'(x)|\}$
- Complete frequency → Sample frequency: binomial distribution $\Pr(n_s)$~*Binomial*(n_c, *p*).

$$\Pr(n_s|n_c,\bar{\rho}) = \binom{n_c}{n_s}\bar{\rho}^{n_s}(1-\bar{\rho})^{n_c-n_s}$$

$$\mathrm{E}(n_s) = n_c\bar{\rho}$$

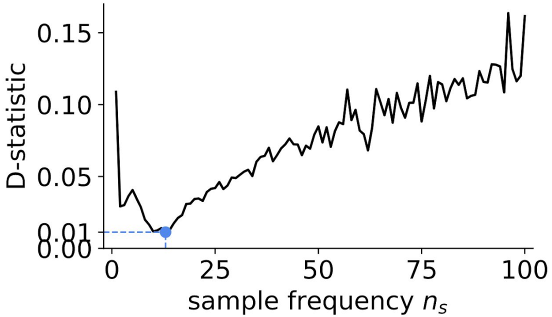


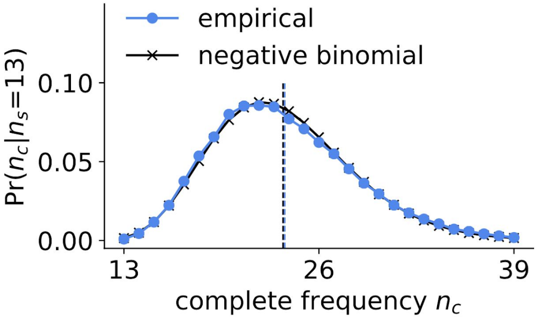
Using Bernoulli process with a uniform rate to approximate the empirical data

- Assumption used in prior studies but not validated [Joseph et al. '13, Pfeffer et al. '18].
- *metric*: D-statistic [Leskovec and Faloutsos '06]. $D(G, G') = \max_x\{|G(x) - G'(x)|\}$
- Complete frequency → Sample frequency: binomial distribution Pr(n_s)~*Binomial*(n_c, *p*).
- Sample frequency → Complete frequency: negative binomial distribution Pr(n_c)~*NegBinomial*(n_s, *p*).

$$\Pr(n_c | n_s, \bar{\rho}) = \binom{n_c - 1}{n_s - 1} \bar{\rho}^{n_s} (1 - \bar{\rho})^{n_c - n_s}$$

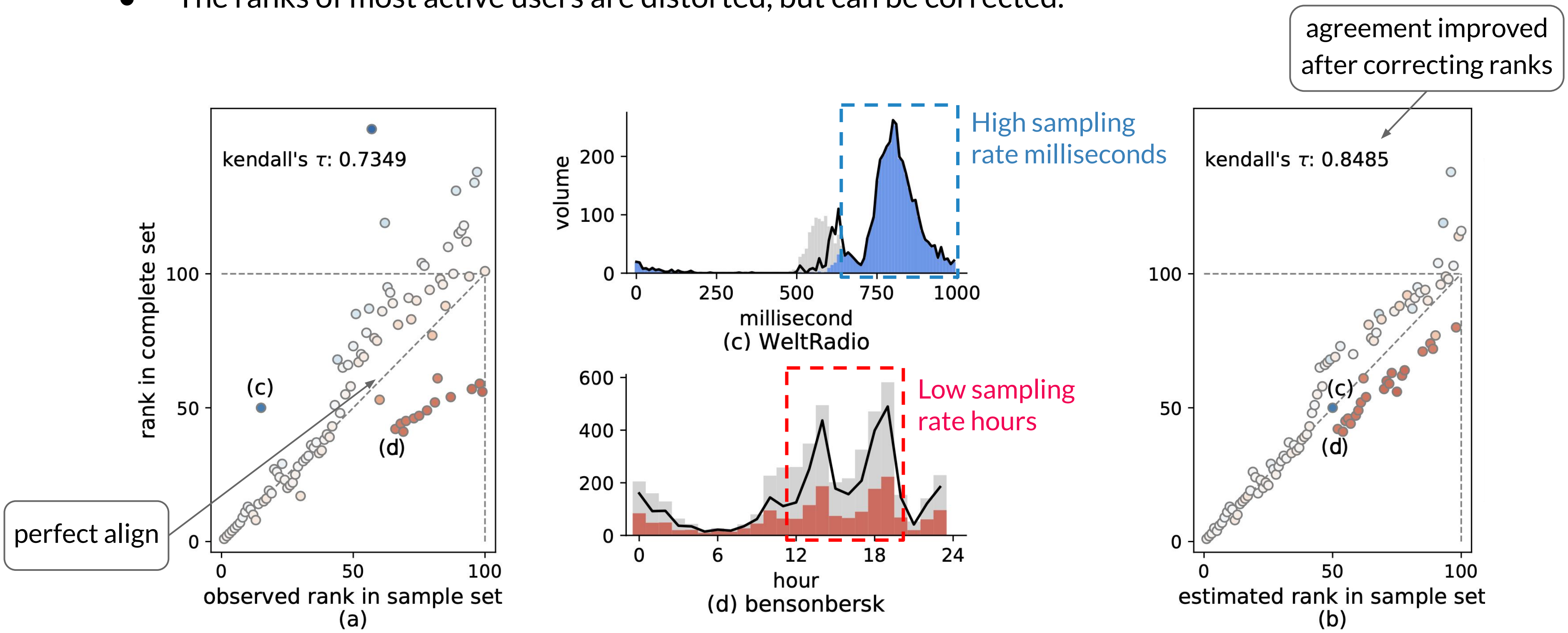
$$\mathrm{E}(n_c) = \frac{n_s}{\bar{\rho}}$$





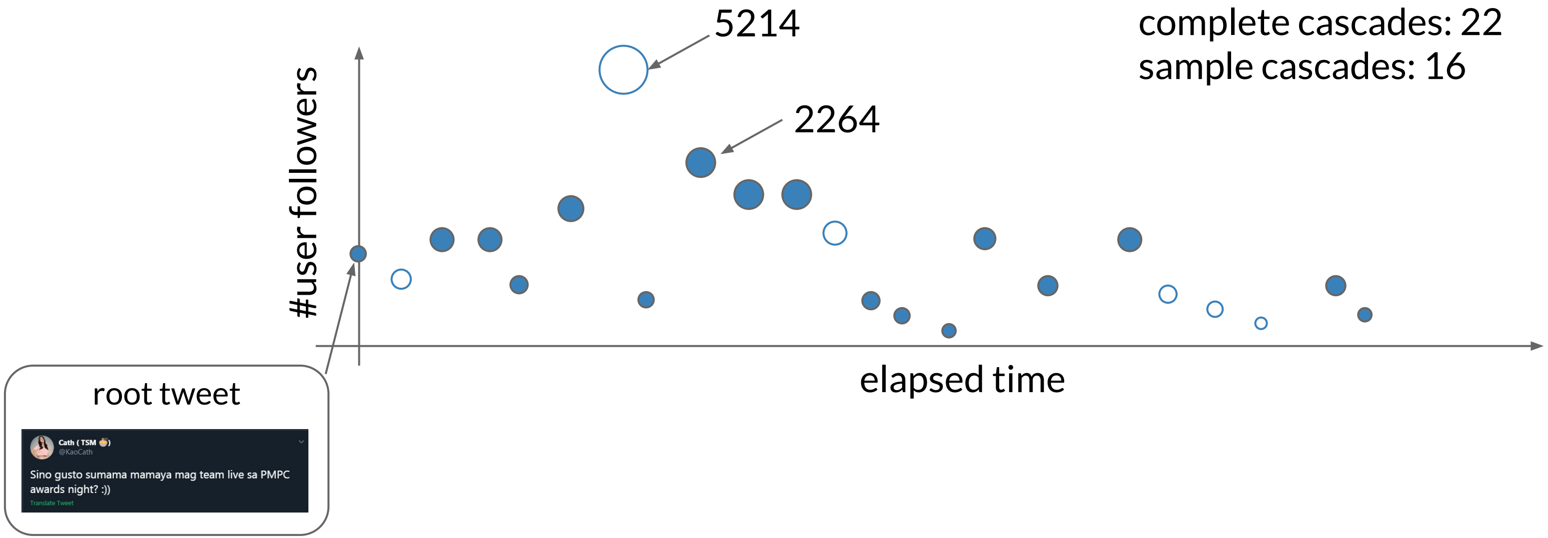
Estimating true ranking from the sample set

- The ranks of most active users are distorted, but can be corrected.



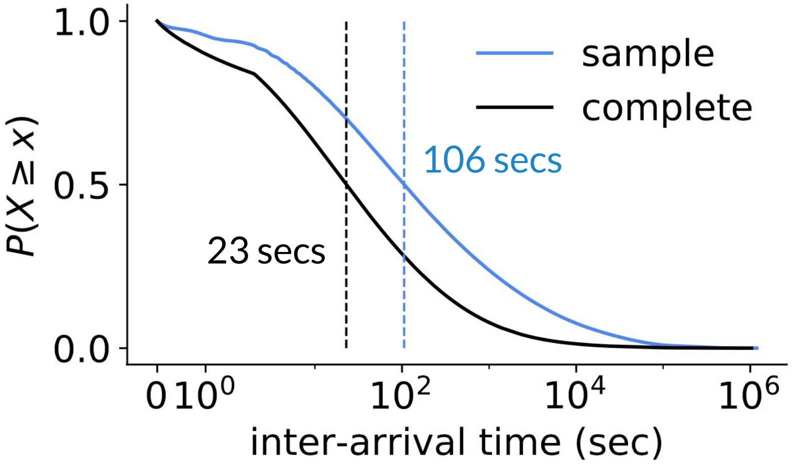
Impacts on retweet cascades

- 2 prominent features: *inter-arrival time*, *user influence* [Zhao et al. ’15, Mishra et al. ’16].
- Strong risks in research that concerns the activity history of each user [Gaffney and Matias ’18].

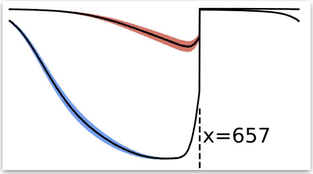


Impacts on retweet cascades

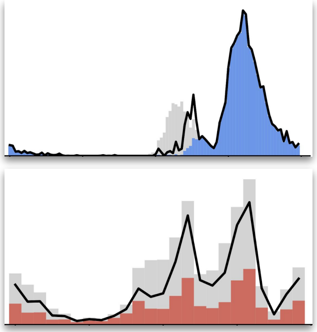
- 2 prominent features: *inter-arrival time, user influence* [Zhao et al. '15, Mishra et al. '16].
- Strong risks in research that concerns the activity history of each user [Gaffney and Matias '18].



Summary



1. **How are the tweets missing in the filtered stream?**
   - The volume of missing tweets can be estimated by Twitter rate limit messages.
   - Tweet sampling rates vary across different timescales.



2. **What are the sampling effects on common measurements?**
   - Bernoulli process with a uniform rate can approximate the empirical entity distribution.
   - True entity ranking can be inferred based on sampled observations.
   - Sampling compromises the quality of diffusion models, since inter-arrival time is significantly longer in the sampled stream, while user influence is lower.

Outline

1. Research scope
2. The effects of social data sampling
3. **The network effects induced by the recommender systems**
   - Characteristics of video recommendation network
   - How to model video popularity under recommender systems?
4. **The patterns of users engagement towards online content**
5. **Conclusion and looking ahead**

The network effects induced by the recommender systems

Papers:

[1] **Wu,** Rizoiu, and Xie. "Estimating Attention Flow in Online Video Networks." In *Proceedings of CSCW, 2019.*

[2] Shin*, Tran*, **Wu*,** Mathews, Wang, Lyall, and Xie. "AttentionFlow: Visualising Dynamic Influence in Ego Networks." Under review, *2020.*

Dataset:

Vevo Music Graph Dataset

Award:

CSCW'19 Best paper honorable mention award

Demo:

[AttentionFlow]

AttentionFlow: Visualising Dynamic Influence in Ego Networks

Category: Research

Paper Type: System

Estimating Attention Flow in Online Video Networks

SIQI WU, Australian National University & Data 61, CSIRO, Australia
MARIAN-ANDREI RIZOIU, University of Technology Sydney & Data 61, CSIRO, Australia
LEXING XIE, Australian National University & Data 61, CSIRO, Australia

Online videos have shown tremendous increase in Internet traffic. Most video hosting sites implement recommender systems, which connect the videos into a directed network and conceptually act as a source of pathways for users to navigate. At present, little is known about how human attention is allocated over such large-scale networks, and about the impacts of the recommender systems. In this paper, we first construct the Vevo network — a YouTube video network with 60,740 music videos interconnected by the recommendation links, and we collect their associated viewing dynamics. This results in a total of 310 million views every day over a period of 9 weeks. Next, we present large-scale measurements that connect the structure of the recommendation network and the video attention dynamics. We use the bow-tie structure to characterize the Vevo network and we find that its core component (23.1% of the videos), which occupies most of the attention (82.6% of the views), is made out of videos that are mainly recommended among themselves. This is indicative of the links between video recommendation and the inequality of attention allocation. Finally, we address the task of estimating the attention flow in the video recommendation network. We propose a model that accounts for the network effects for predicting video popularity, and we show it consistently outperforms the baselines. This model also identifies a group of artists gaining attention because of the recommendation network. Altogether, our observations and our models provide a new set of tools to better understand the impacts of recommender systems on collective social attention.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing.**

Additional Key Words and Phrases: YouTube; recommender system; empirical measurement; network effects; online attention; popularity prediction

ACM Reference Format:

Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2019. Estimating Attention Flow in Online Video Networks. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 183 (November 2019), 25 pages. https://doi.org/10.1145/3359285

1 INTRODUCTION

Many online platforms present algorithmic suggestions to help users explore the enormous content space. The recommender systems, which produce such suggestions, are central to modern online platforms. They have been employed in many applications, such as finding new friends on Twitter [41], discovering interesting communities on LinkedIn [40], and recommending similar goods on Amazon [18, 33]. In the domain of multimedia, service providers (e.g., YouTube, Netflix, and Spotify) use recommender systems to suggest related videos or songs [9, 16, 17, 20, 54]. Much

Authors' addresses: Siqi Wu, Australian National University & Data 61, CSIRO, Australia, siqi.wu@anu.edu.au; Marian-Andrei Rizoiu, University of Technology Sydney & Data 61, CSIRO, Australia, marian-andrei.rizoiu@uts.edu.au; Lexing Xie, Australian National University & Data 61, CSIRO, Australia, lexing.xie@anu.edu.au.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Association for Computing Machinery.
2573-0142/2019/11-ART183 $15.00
https://doi.org/10.1145/3359285

183

Recommender systems are ubiquitous in online platforms

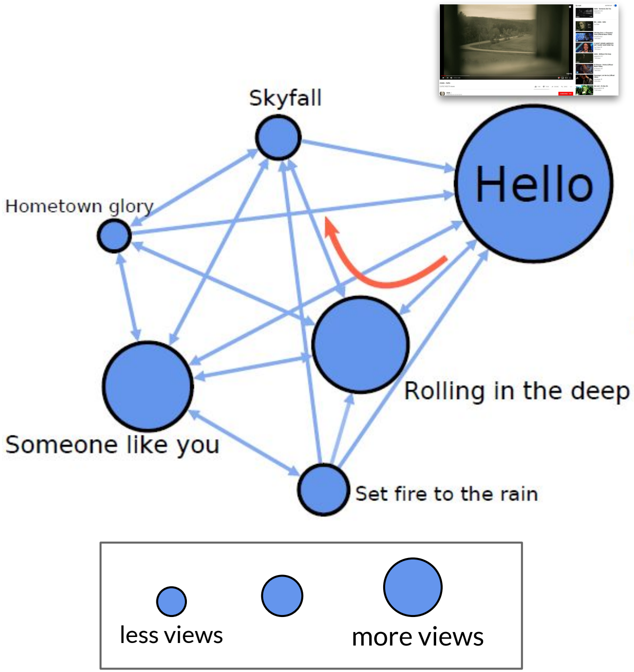
Method	Papers
Collaborative Filtering	[Davidson et al. *RecSys* ’10] [Bendersky et al. *KDD* ’14]
Deep Learning	[Covington et al. *RecSys* ’16] [Beutel et al. *WSDM* ’18]
Reinforcement Learning	[Chen et al. *WSDM* ’19] [Ie et al. *IJCAI* ’19]
Unbiased recommendation	[Zhao et al. *RecSys* ’19] [Yi et al. *RecSys* ’19]

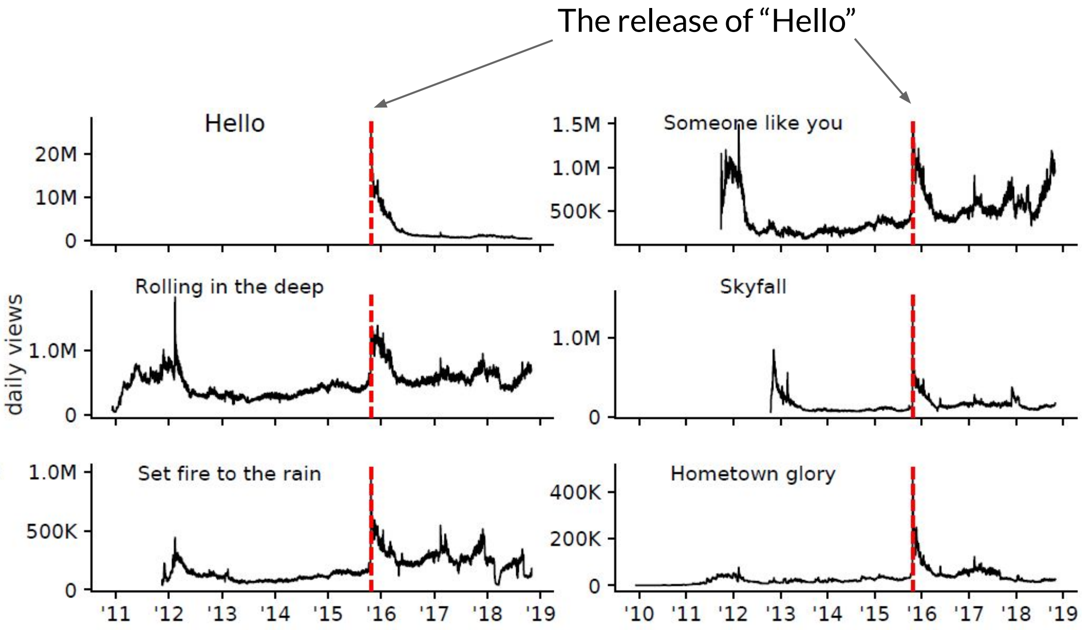
Q1. What does the recommendation network look like?
Q2. What are the network effects on video popularity?

Contribution: **the first large-scale study on the network effects induced by the YouTube recommender systems and a new model for estimating the attention flows over the network**

The “Hello” effect

- The release of “Hello” excited other videos from Adele[1].

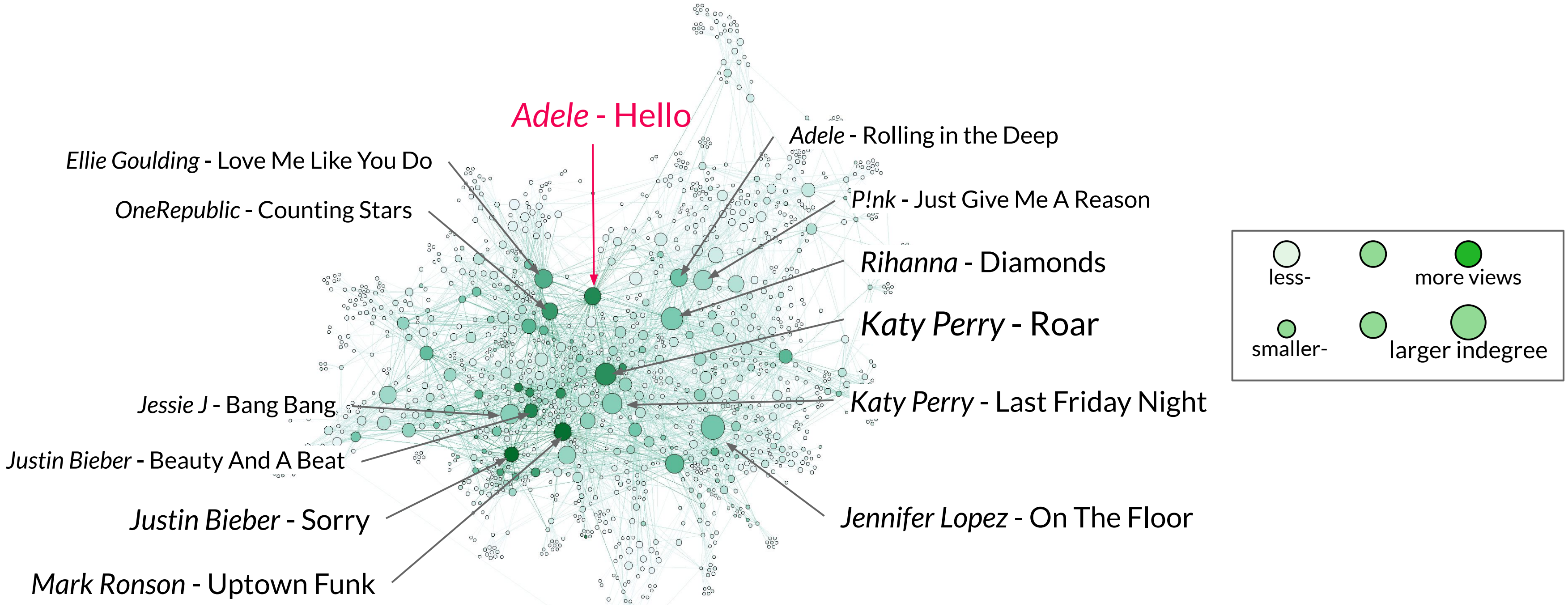




[1] Adele’s ’Hello’ Has Biggest YouTube Debut of Any Video This Year | Billboard https://bit.ly/2F6WLVQ

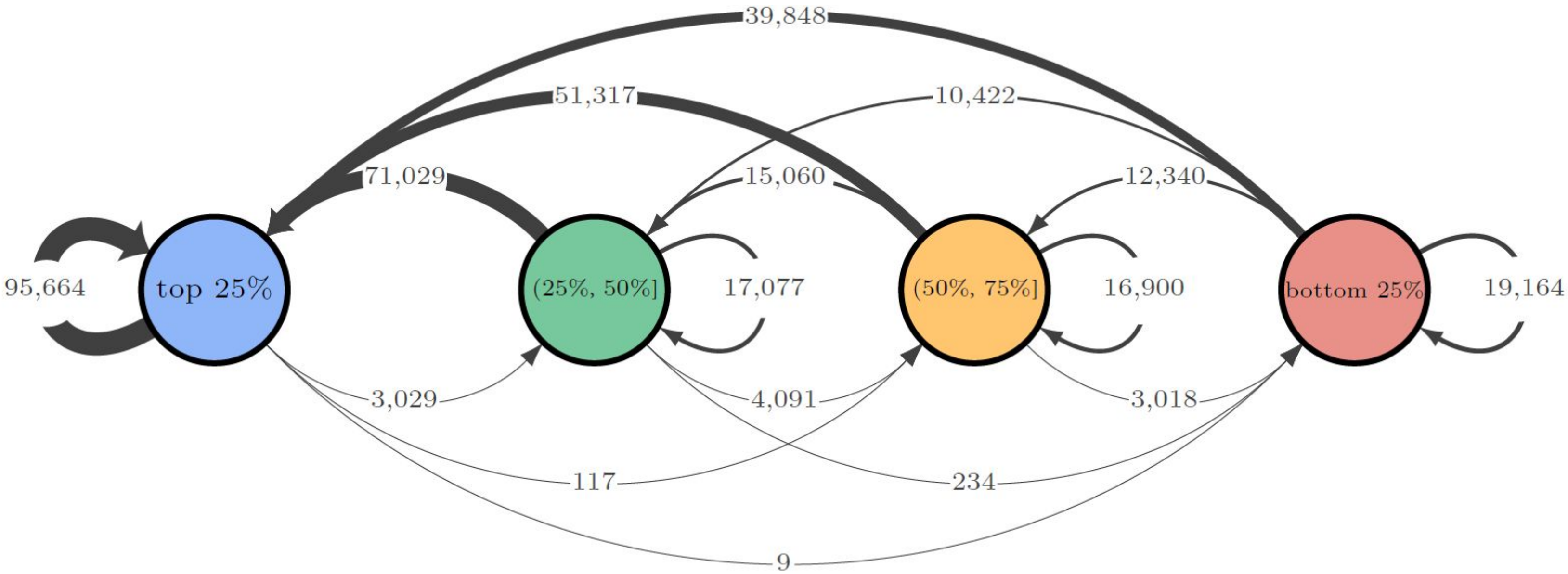
VEVO music graph dataset

- 60,740 music videos from 4,435 VEVO artists who are active in major English-speaking countries.
- 337K~394K directed links in 63 daily snapshots.
- Links consist of *non-personalized* feed from YouTube API.



Popularity bias in the recommendation network

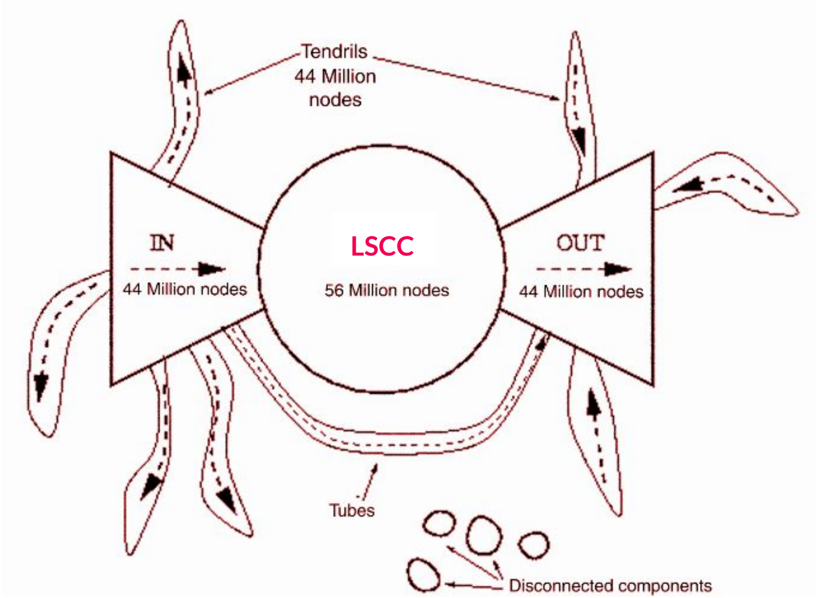
- Videos disproportionately point to more popular videos.



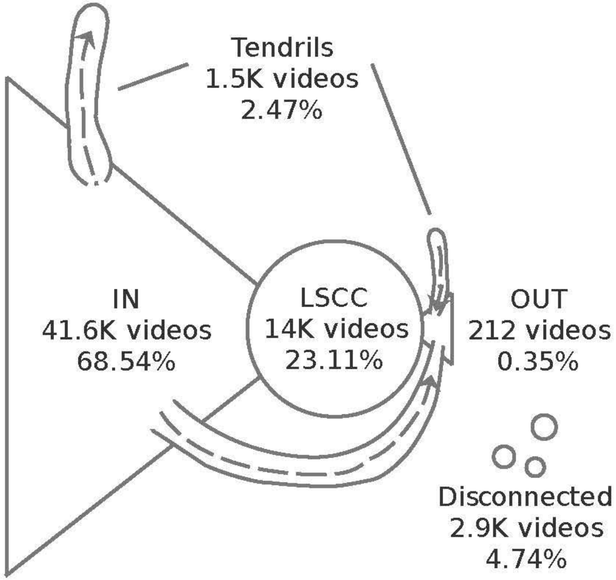
The bow-tie structure

- LSCC: largest strongly connected component.
- IN: nodes can reach LSCC, but not reachable from the nodes in LSCC.
- OUT: nodes that can be reached by LSCC but not pointing back to LSCC.

Web graph 1997 [Broder et al. '00]



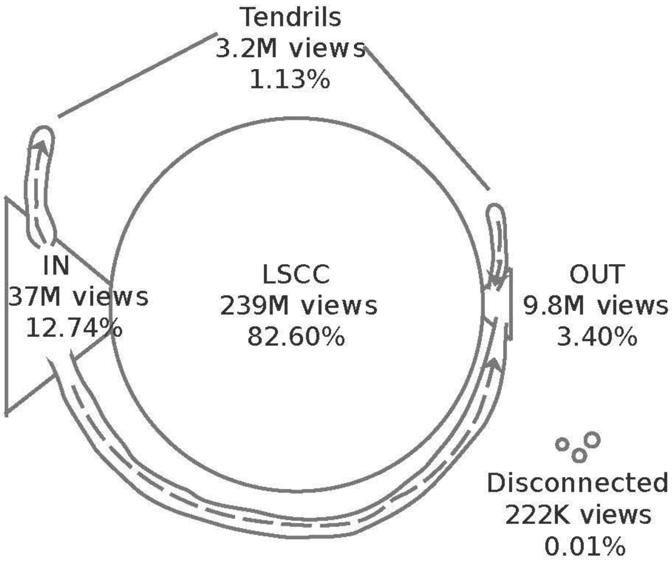
VEVO network



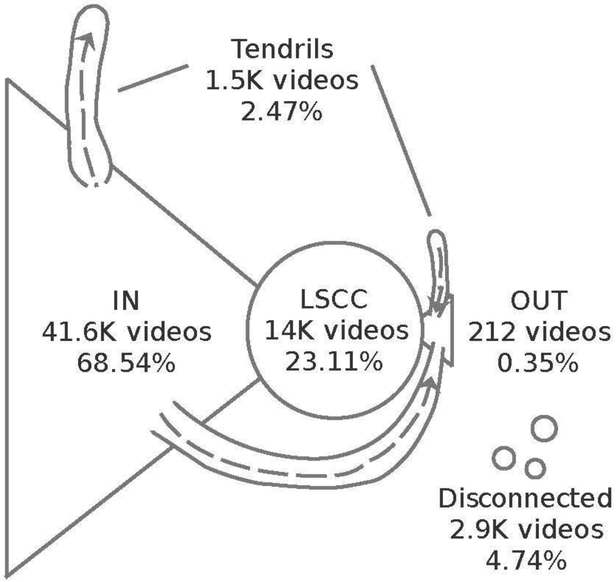
The attention bow-tie of Vevo network

- Attention flow in one direction: IN → LSCC → OUT.
- LSCC (23.1% of the videos) occupies most of the attention (82.6% of the views).
- IN component shrinks (68% → 12%).

attention bow-tie



structure bow-tie

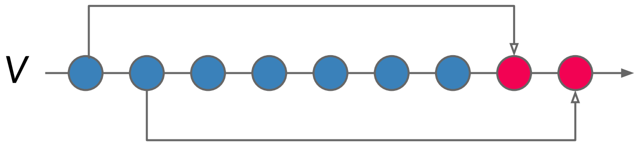


Outline

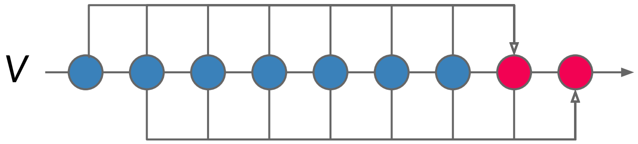
1. Research scope
2. The effects of social data sampling
3. **The network effects induced by the recommender systems**
   - Characteristics of video recommendation network
   - How to model video popularity under recommender systems?
4. **The patterns of users engagement towards online content**
5. **Conclusion and looking ahead**

Baseline methods

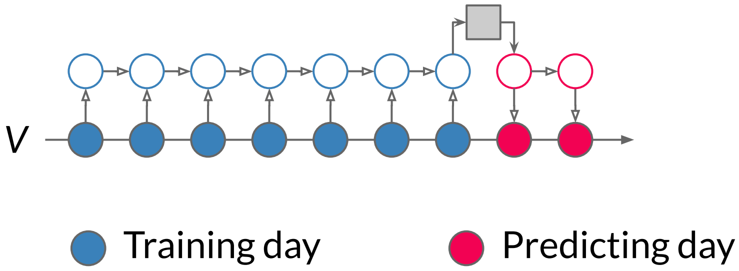
- Seasonality -> Seasonal Naive model (SN)

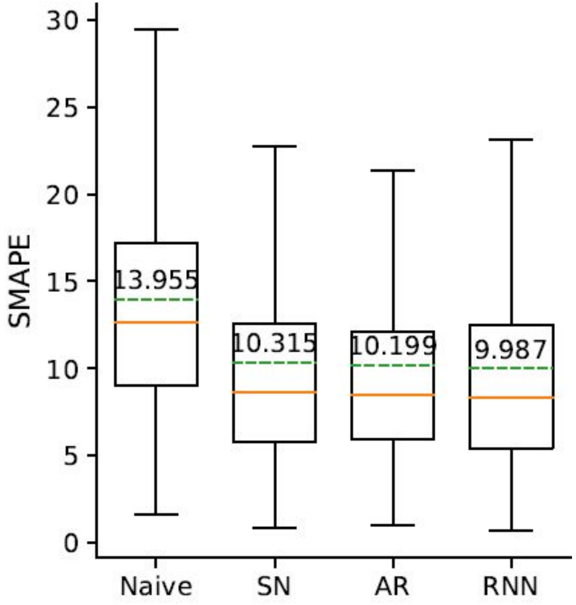


- Autocorrelation -> AutoRegressive model (AR)



- RNN with LSTM units





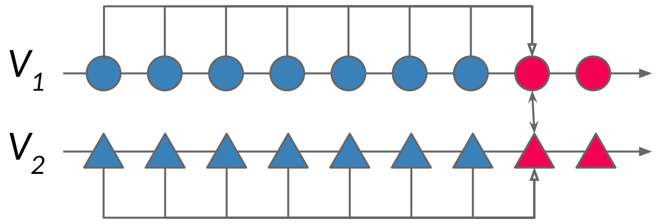
ARNet model and results

Baselines:

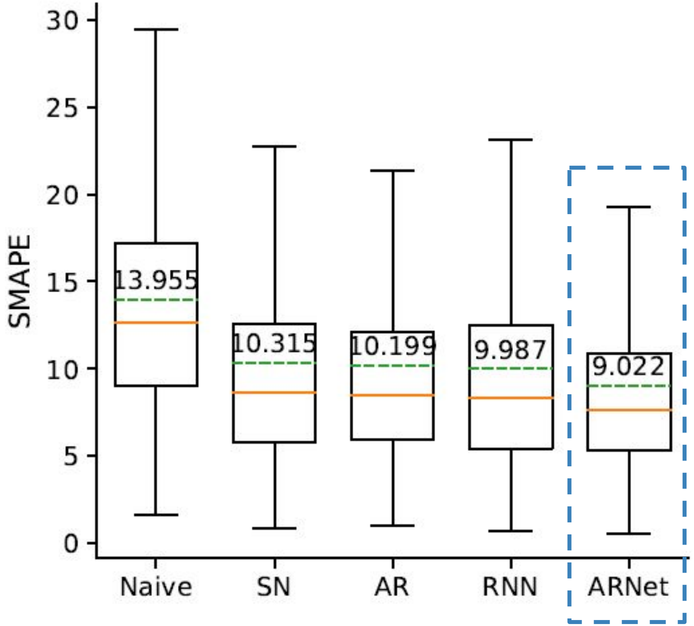
- Seasonality -> Seasonal Naive model (SN)
- Autocorrelation -> AutoRegressive model (AR)
- RNN with LSTM units

Proposed model:
AutoRegressive + Network (ARNet)

$$\hat{\mathbf{Y}}_v[t] = \underbrace{\sum_{\tau=1}^{w} \alpha_{v,\tau} \mathbf{Y}_v[t-\tau]}_{latent\ interest} + \underbrace{\sum_{(u,v)\in G} \beta_{u,v} \mathbf{Y}_u[t]}_{network\ effect}$$

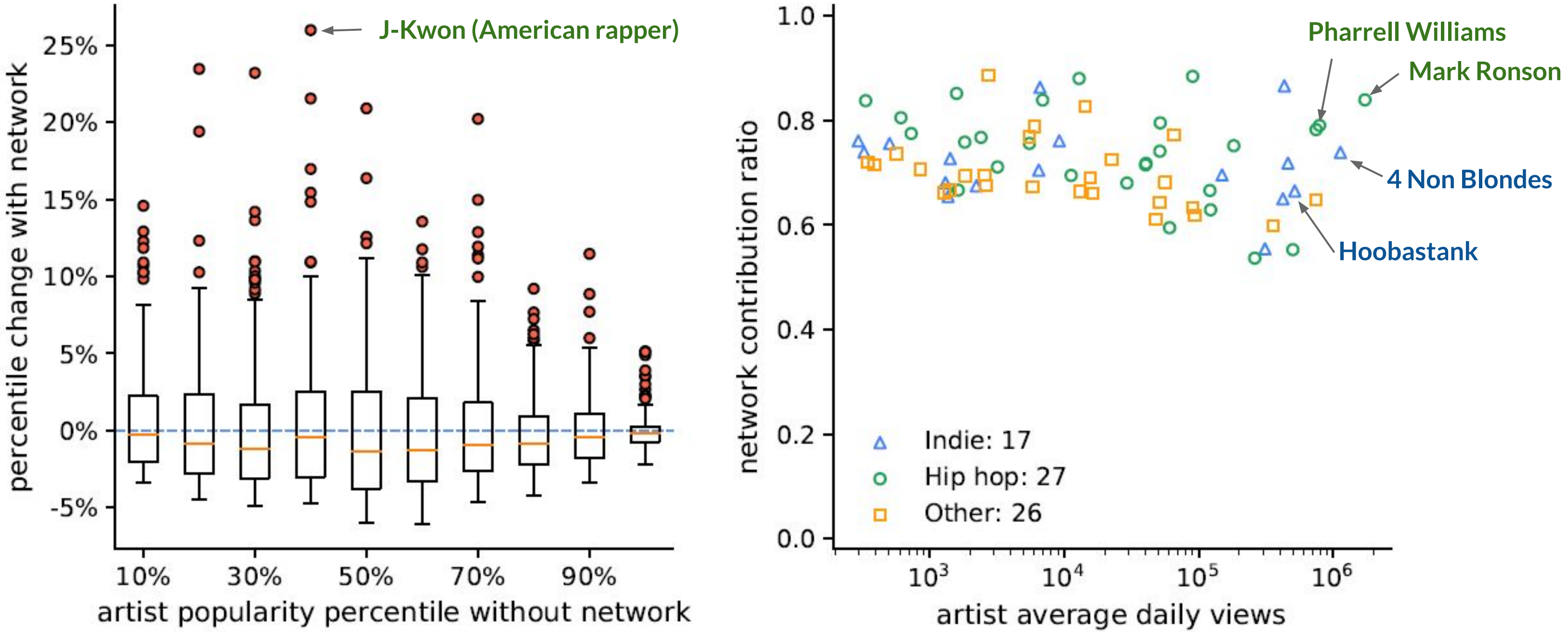


Training day   Predicting day



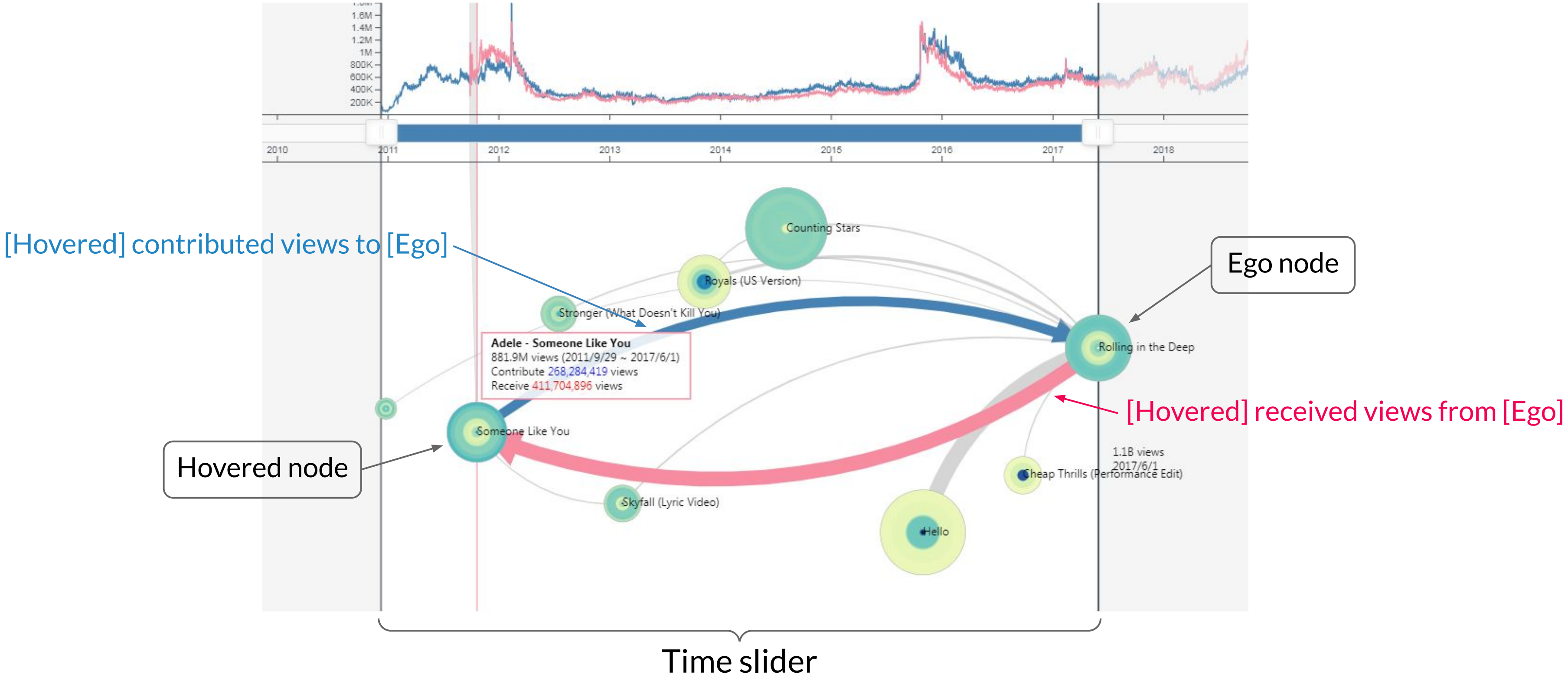
Which artists benefit the most from the recommendation network?

Estimated network contribution ratio: $\dfrac{\sum_{(u,v)\in G} \beta_{u,v} \mathbf{Y}_u}{\hat{\mathbf{Y}}_v}$

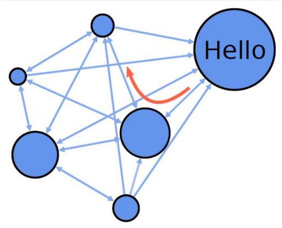


New demo: AttentionFlow

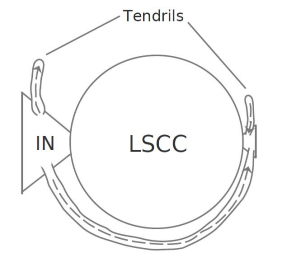
- *AttentionFlow*: a new system to visualise a collection of time series and the dynamic network influence.



Summary

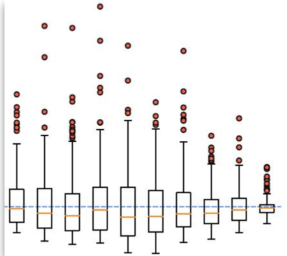


1. How to build the network of videos from recommender systems?
Building a non-personalized video network from YouTube API.



2. Characteristics of the recommendation network

- Popularity bias
- Unequal attention allocation



3. How to model video popularity under recommender systems?

- A model taking account of network information.
- Estimating link strength for each recommendation link.

Outline

1. Research scope
2. The effects of social data sampling
3. The network effects induced by the recommender systems
4. The patterns of users engagement towards online content
- How to measure aggregate engagement?
- Characteristics of aggregate engagement
- Can aggregate engagement be predicted?
5. Conclusion and looking ahead

The patterns of users engagement towards online content

Papers:
[1] **Wu,** Rizoiu, and Xie. “Beyond Views: Measuring and Predicting Engagement in Online Videos.” In *Proceedings of ICWSM, 2018.*
[2] Kong, Rizoiu, **Wu,** and Xie. “Will This Video Go Viral? Explaining and Predicting the Popularity of Youtube Videos.” In *Proceedings of WWW Companion, 2018.*

Dataset:
YouTube Engagement '16 Datasets

Software:
YouTube-insight: collecting metadata and insight data for YouTube videos

Demo:
[HIPie]

View count does NOT translate to watch time[1,2]

view count: 3,917,179 < view count: 7,833,595

watch time: 62 years > watch time: 32 years

[1] YouTube Now: Why We Focus on Watch Time https://bit.ly/2G9iuvc
[2] Facebook: Updating How We Account For Video Completion Rates https://bit.ly/2juca5b

Popularity and engagement for web content

Domains	Popularity metrics	Engagement metrics
Webpages	Visit number [Li and Moore *JMLR* '08]	Click-through-rate [Richardson et al. *WWW* '07]
Search ads	Display number [He et al. *ADKDD* '14]	Conversion rate [Barbieri et al. *WWW* '14]
Songs	Listening count [Bellogin et al. *ICWSM* '13]	Download number [Krumme et al. *PloS* '12]
Videos	View count [Pinto et al. *WSDM* '13] [Szabo and Huberman *Com.ACM* '10]	Watch time [Guo et al. *L@* [Park et al. *IC*

★ **No browser extension**
★ **Cold-start prediction**

Q1. How to measure aggregate engagement?
Q2. Can aggregate engagement be predicted?

Contribution: **the first large-scale measurement study of how users collectively engage with online content and a new metric to describe content quality**

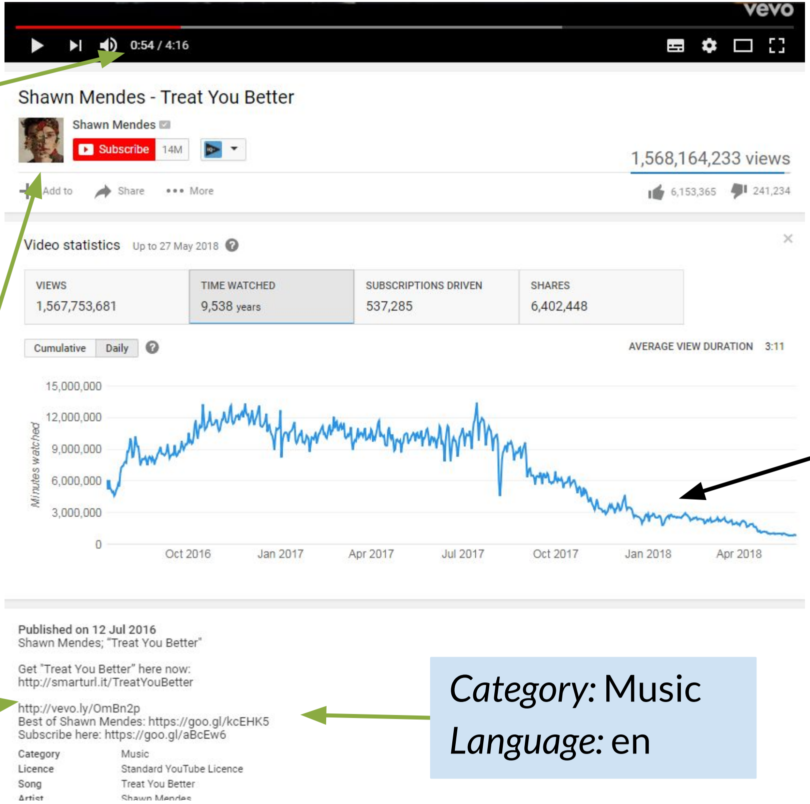
Tweeted Videos dataset

- 5 million YouTube videos published and tweeted in July and August 2016.

Video duration: 4M16S
Visual definition: HD or SD

Video Title:
Shawn Mendes - Treat You Better
Channel Id:
UC4-TgOSMJHn-LtY4zCzbQhw
Channel Title:
ShawnMendesVEVO

Freebase topics:
Shawn Mendes; Music; Music video; Pop music

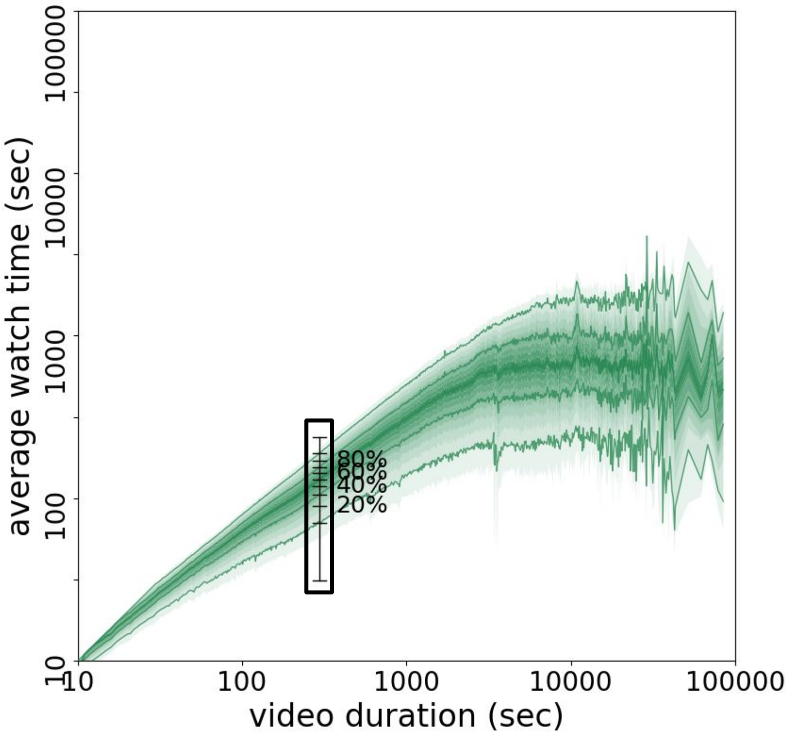


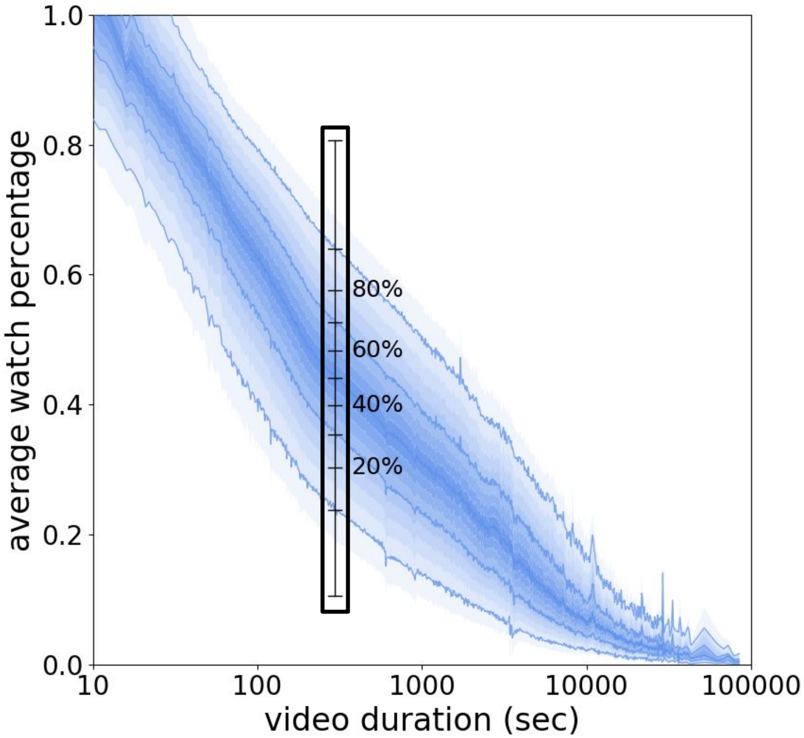
Insight time series:
(a) Daily watch time
(b) Daily view count
(c) Daily share count
(d) Avg. watch time

Category: Music
Language: en

Engagement maps

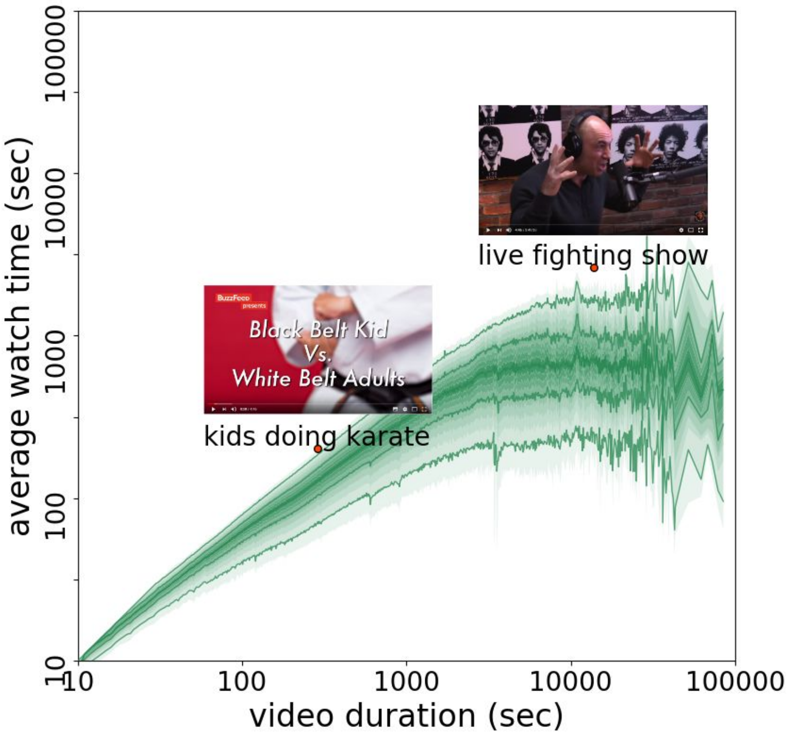
tch

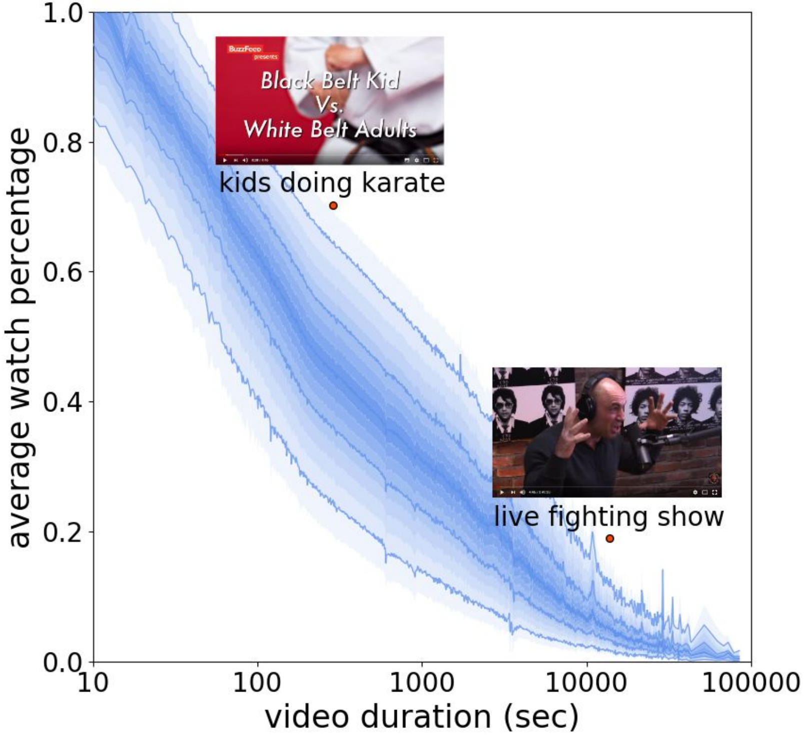




Engagement maps

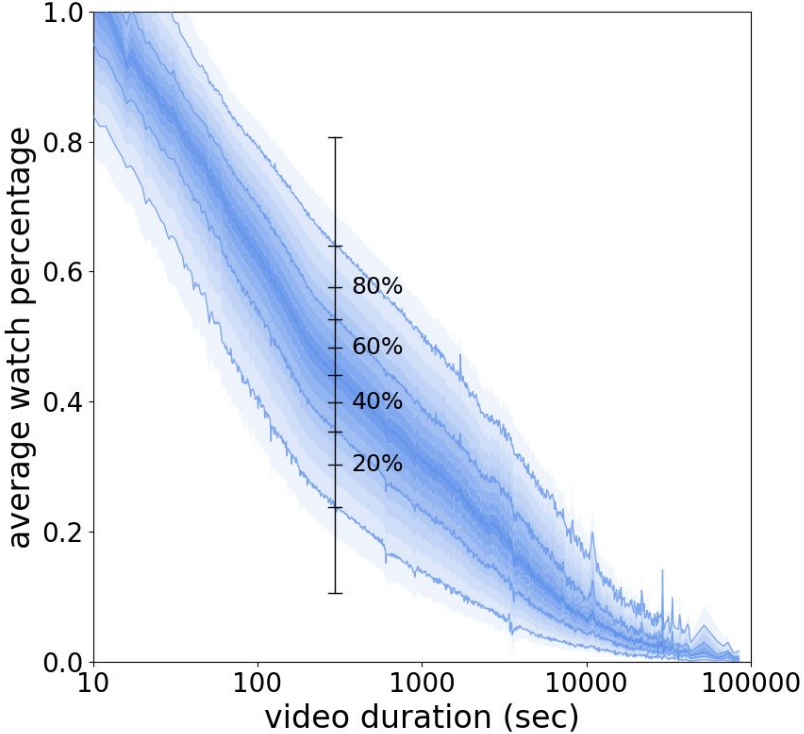
- x-axis: video length; y-axis: watch time or watch percentage.

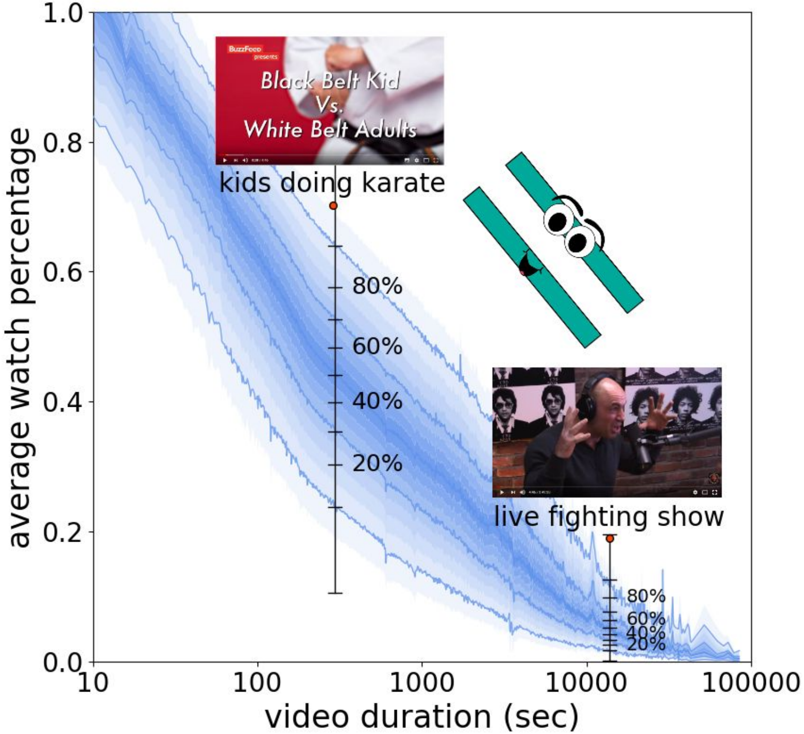




New metric: relative engagement

- Rank percentile of average watch percentage among videos with similar lengths





Outline

1. Research scope
2. The effects of social data sampling
3. The network effects induced by the recommender systems
4. **The patterns of users engagement towards online content**

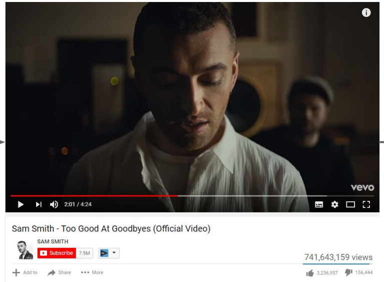
- How to measure aggregate engagement?
- Characteristics of aggregate engagement
- Can aggregate engagement be predicted?

5. **Conclusion and looking ahead**

Quality Videos datasets: Music and News

Music

Tweeted music clip
449,314 videos



Professional Vevo video
67,649 videos

Billboard top hit
63 videos

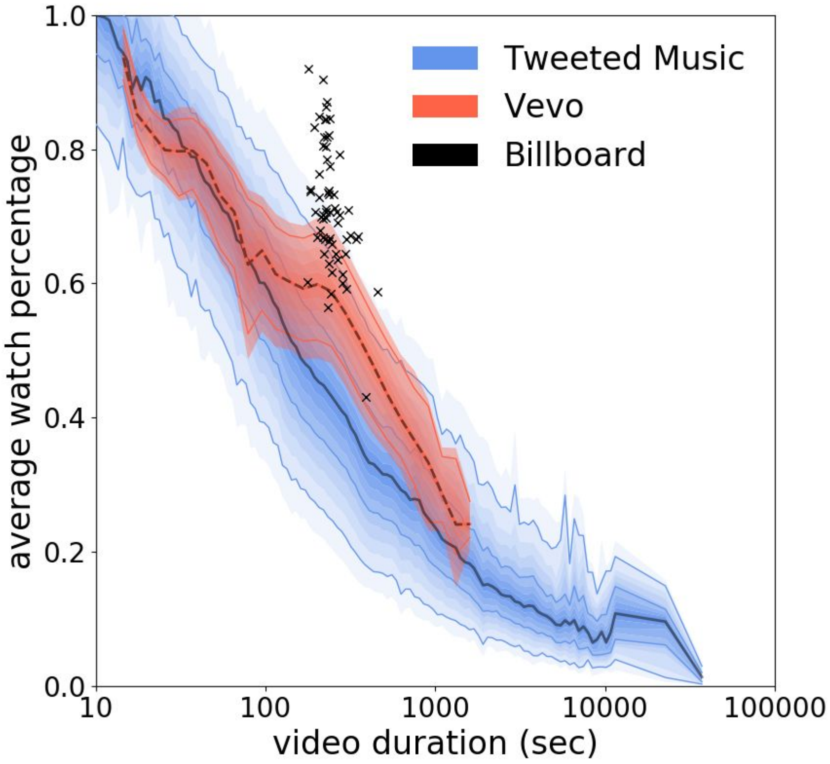
News

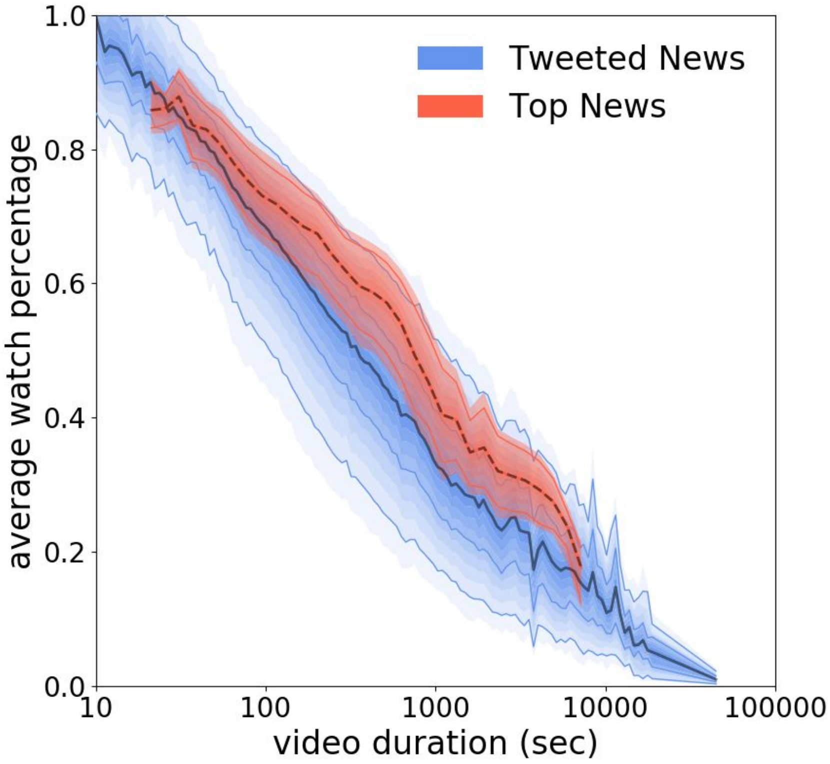
Tweeted news clip
459,728 videos

Quality is improved

Top news video
28,685 videos

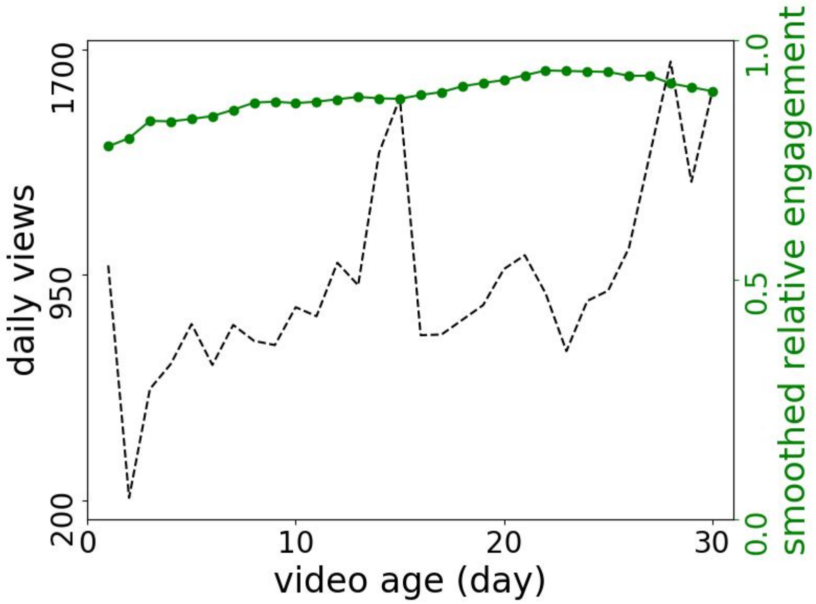
Relative engagement is correlated with video quality



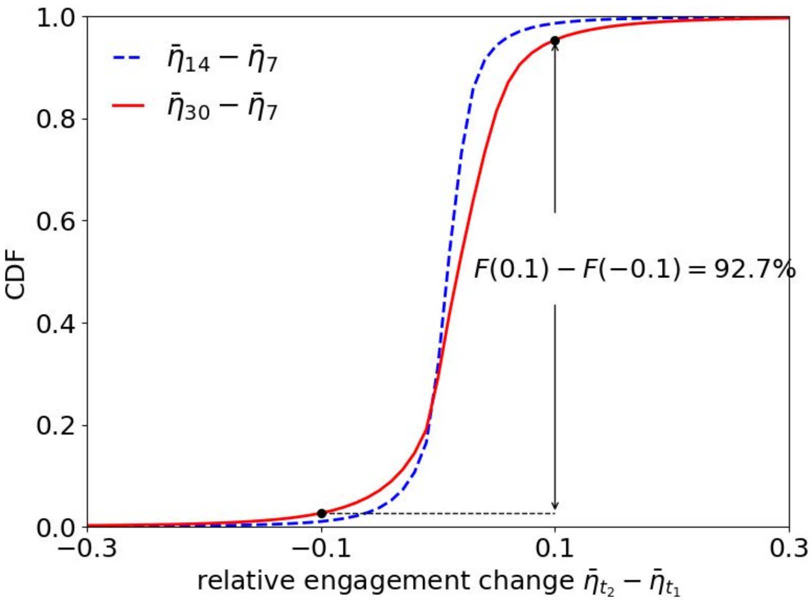


Post-clicking behavior is only relevant to product quality. [Krumme et al. *PloS* '12]

Relative engagement is stable over time



Video Id: XIB8Z_hASOs
Video Title: DC Super Hero Girls S02E10



92.7% of videos stay within 0.1 in relative engagement

Outline

1. Research scope
2. The effects of social data sampling
3. The network effects induced by the recommender systems
4. **The patterns of users engagement towards online content**

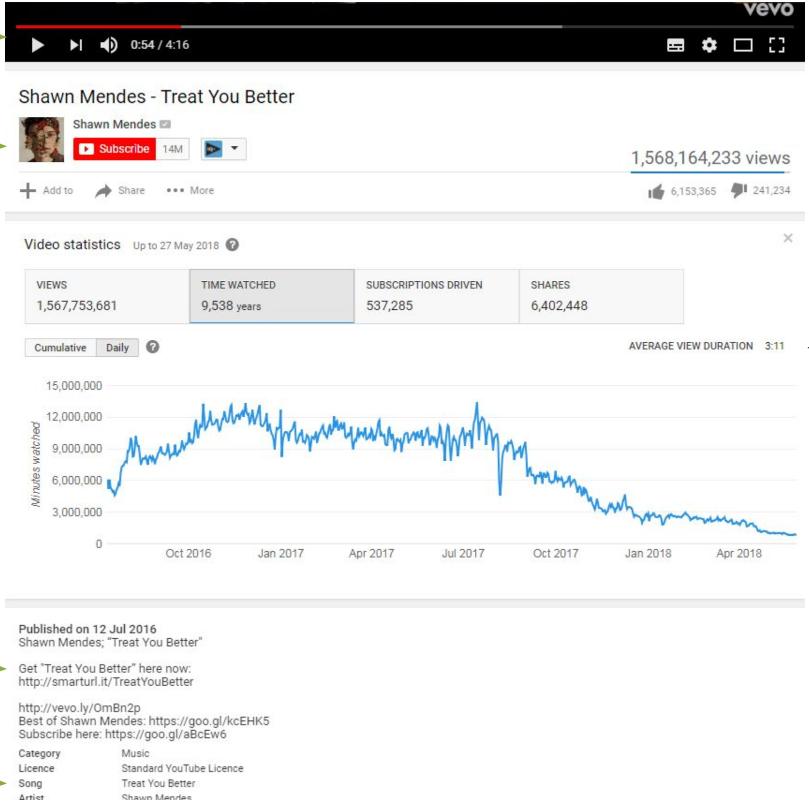
- How to measure aggregate engagement?
- Characteristics of aggregate engagement
- Can aggregate engagement be predicted in a *cold-start* setup?

5. **Conclusion and looking ahead**

Prediction task setup

Video duration: 4M16S

Channel activity level:
Daily upload number
Channel past engagement:
Summary of past performance



Prediction targets:
(a) Relative engagement
(b) Avg watch percentage

Prediction method:
Ridge regression

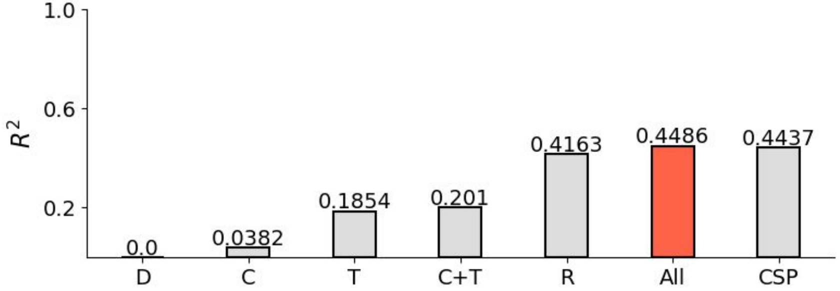
Evaluation metric:
R2

Visual definition: HD or SD
Category: Music
Language: en

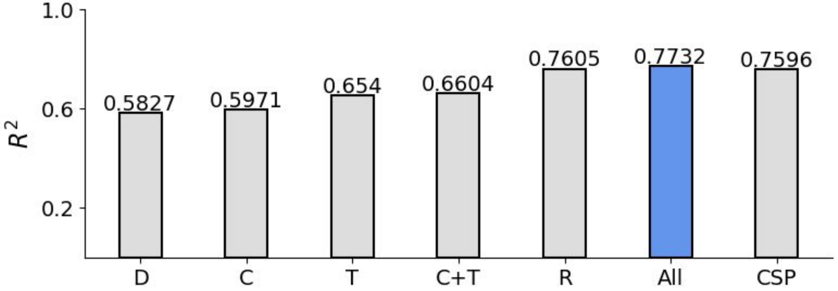
Freebase topics:
Shawn Mendes; Music; Music video; Pop music

Prediction results

Predict relative engagement



Predict average watch percentage

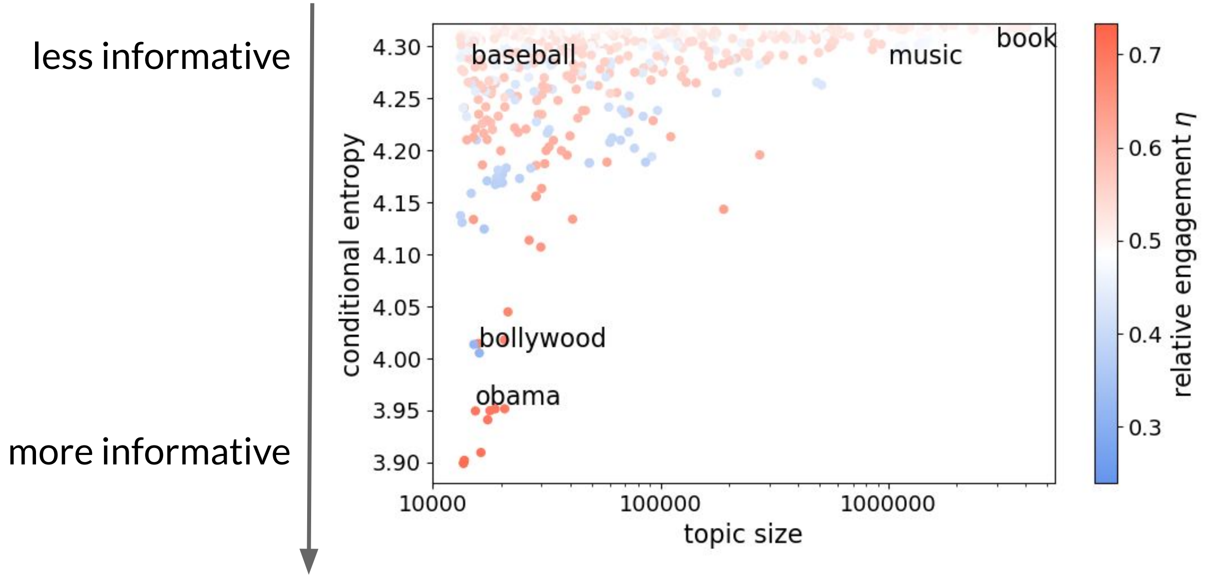


D: duration; **C:** context; **T:** topic; **C+T:** context+topic;
R: channel past reputation; **All:** all features; **CSP:** channel specific predictor

- R2 up to 0.45 for relative engagement and 0.77 for average watch percentage.
- Channel related features are the most predictive, consistent with [Cheng et al. *WWW* '14].
- Topic features are somewhat predictive, contrasting to [Martin et al. *WWW* '16].

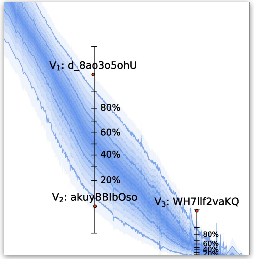
What are engaging topics?

- Conditional entropy: $H(Y|X_i = 1) = -\sum_{y \in Y} P(y|x_i = 1) \log_2 P(y|x_i = 1)$



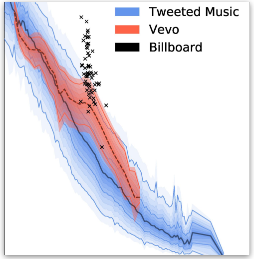
500 most frequent topics

Summary



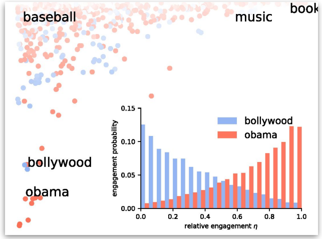
1. How to measure aggregate engagement?

Relative engagement - a new metric calibrated against video duration.



2. Characteristics of aggregate engagement

- Relative engagement is correlated with content quality.
- Relative engagement is stable over time.



3. Can aggregate engagement be predicted?

Engagement can be predicted before a video's upload, achieving R2=0.77.

Outline

1. Research scope
2. The effects of social data sampling
3. The patterns of users engagement towards online content
4. The network effects induced by the recommender systems
5. **Conclusion and looking ahead**

Conclusion

RQ1: Social data sampling

1. The volume of missing tweets can be estimated by rate limit messages.

2. Tweet sampling rates vary across different timescales.

3. True entity statistics (e.g., distribution and ranking) can be inferred based on sampled observations.

RQ2: Recsys network effects

1. Methods of building the non-personalized content recommendation network.

2. Characteristics of the recommendation network.

3. A model that estimates the network contribution between videos and artists.

RQ3: Collective engagement

1. A new metric "relative engagement" that calibrates against video length.

2. Correlated with video quality and stable over time.

3. Engagement metrics can be predicted in a cold-start setup.

Implications: methods and observations can be used by content producers, hosting sites, and online users to improve content production, advertising strategies, and recommender systems.

Long-term goal: developing principles for ***responsible platforms*** by using ***data visualization*** and ***machine learning*** to measure and model the ***collective user behavior***.

Publications

1. *Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model.*
   Cheng, Shu, **Wu,** Silva, Hall, and Liu. In proceeding of ***CIKM, 2020***. (Full paper)

2. *AttentionFlow: Visualising Dynamic Influence in Ego Networks.*
   Shin*, Tran*, **Wu*,** Mathews, Wang, Lyall, and Xie. Under review. (Short paper)

3. *Variation across Scales: Measurement Fidelity under Twitter Data Sampling.*
   Wu, Rizoiu, and Xie. In proceeding of ***ICWSM, 2020***. (Full paper)

4. *Estimating Attention Flow in Online Video Networks.*
   Wu, Rizoiu, and Xie. In proceeding of ***CSCW, 2019***. (Full paper, **best paper honorable mention**)

5. *How is Attention Allocated? Data-driven Studies of Popularity and Engagement in Online Videos.*
   Wu. In proceeding of ***WSDM, 2019***. (Doctoral Consortium)

6. *Beyond Views: Measuring and Predicting Engagement in Online Videos.*
   Wu, Rizoiu, and Xie. In proceeding of ***ICWSM, 2018***. (Full paper)

7. *Will This Video Go Viral? Explaining and Predicting the Popularity of Youtube Videos.*
   Kong, Rizoiu, **Wu**, and Xie. In proceeding of ***WWW Companion, 2018***. (Demo)

Software, Demo, Data, and Award

Software

- *Youtube-insight* [6]: collecting metadata and historical data for YouTube videos.
- *Twitter-intact-stream* [3]: reconstructing the complete Twitter filter stream.

Demos

- *AttentionFlow* [2]: a new system to visualise a collection of time series and the dynamic network influence.
- *HIPie* [7]: a new system to explain and predict the popularity of Youtube videos.

Datasets

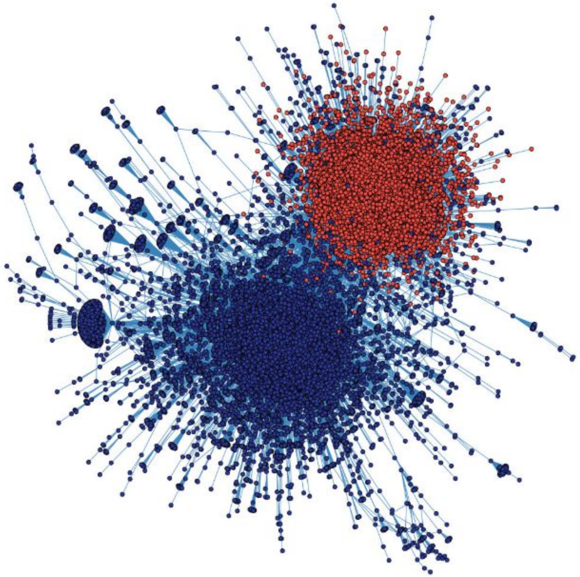
- *Complete/sampled retweet cascades* [3]: 2 sets of complete/sampled retweet cascades on cyberbullying and YouTube.
- *Vevo music graph* [4]: 60K videos with 63 daily snapshots of the video recommendation network.
- *YouTube engagement '16* [6]: 5M videos published and tweeted in Jul-Aug 2016.
- *YouTube active '17*: 330K videos published and tweeted 100+ times during 2017.

Awards

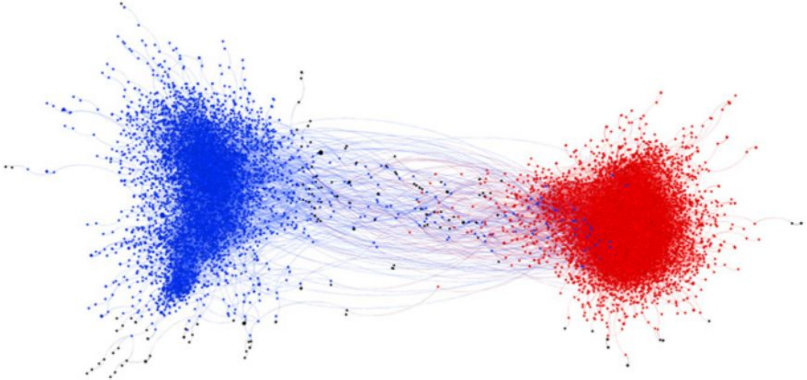
- CSCW 2019 Best paper honourable mention award [4]
- 2018 Google PhD fellowship

Future work: Political polarization on YouTube

- Political echo chamber on Twittersphere.



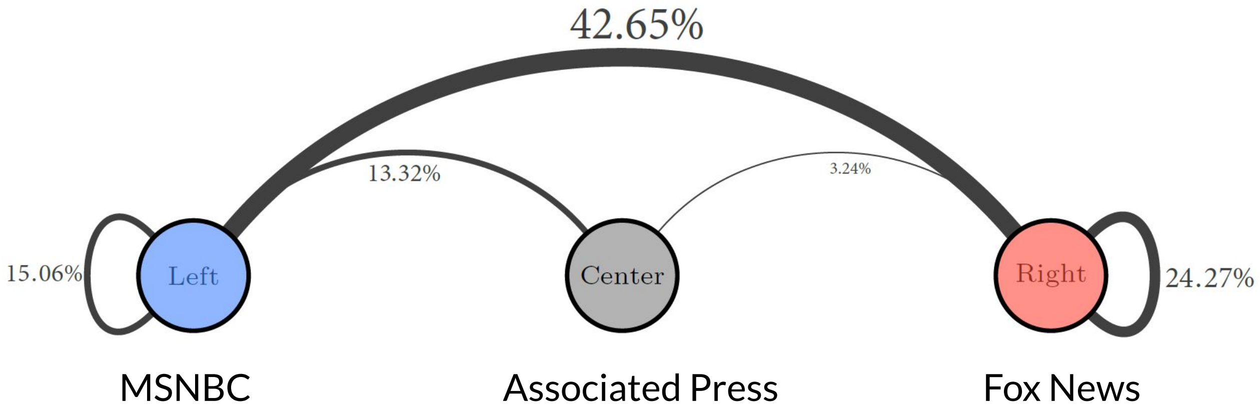
Twitter retweet network
[Conover et al. *ICWSM '11*]



Twitter retweet network
[Brady et al. *PNAS '17*]

Future work: Political polarization on YouTube

- Political echo chamber on Twittersphere.
- 42% users interact with both left- and right-leaning videos on YouTube.
- New ***YouTube media bias dataset***: 957 YouTube channels of media outlets, 414,923 videos published in 2020, and millions of comments and users who post the comments.
- Q1: Who are the users doing cross-partisan conversation?
- Q2: Are cross-partisan users getting more attention (e.g., votes, replies)?
- Q3: Do cross partisan users compliment one party and belittle the other?
- ...



Thank you!

Family Wife and parents, who always support me unconditionally!

Supervisory panel

Lexing Xie

Marian-Andrei Rizoiu

Cheng Soon Ong

CMLab folks Quyu Kong, Swapnil Misha, Alex Mathews, Dawei Chen, Minjeong Shin, Dongwoo Kim, Alasdair Tran, Rui Zhang, Umanga Bista, and many others!

Lifelong friends
Collaborators
Mentors

Lichen Liu (UPV-ITQ), Fangjie Xie (UCR), Lu Cheng (ASU),
Mert Ertugrul, Yu-Ru Lin (UPitt), Paul Resnick, James Park (UMich)