Measuring Collective Attention in Online Content: Sampling, Engagement, and Network Effects

Siqi Wu

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

March 2021

Except where otherwise indicated, this thesis is my own original work.

Siqi Wu
9 March 2021

Acknowledgments

I would like to express my sincere gratitude to the people who have supported me in this Ph.D grind:

- Prof. Lexing Xie. I am extremely honored to have Lexing to be my advisor. Her extensive knowledge in various fields and strong dedication to research motivate me to become a good researcher.

- Dr. Marian-Andrei Rizoiu. Andrei is one of the smartest people that I know. He provided many insightful ideas when I was stuck. His research also sparked my interests in modeling online popularity.

- Dr. Cheng Soon Ong. Cheng is my "think tank". He never turned me down when I felt discouraged or desperately looked for advices, both in research and in life.

- Members of the ANU Computational Media Lab – Swapnil Mishra, Quyu Kong, Alexander Mathews, Dawei Chen, Minjeong Shin, Dongwoo Kim, Jooyoung Lee, Rui Zhang, Alasdair Tran, Umanga Bista, Yuli Liu, Qiongkai Xu, and many others. I am grateful that I can work with a group of supportive talents. Much of my work is benefited from the discussions with them.

- External collaborators – Yu-Ru Lin, Ali Mert Ertugrul, and Xian Teng from University of Pittsburgh, Paul Resnick and James Park from Univeristy of Michigan, and Lu Cheng from Arizona State University. It has been a great pleasure to work with them. Our collaborations also lead to fruitful research outcomes.

- Data61, CSIRO. I would like to thank Data61 for providing my scholarship. It has also provided me a strong network in both academia and industry.

- ANU Research School of Computer Science. I would like to thank the admin and HDR team in CECS for creating a friendly working environment for us.

- National eResearch Collaboration Tools and Resources (Nectar). I would like to thank Nectar for providing computational resources that facilitate my research.

This research is supported in part by Asian Office of Aerospace Research and Development Grant 19IOA078, Air Force Research Laboratory Grant FA2386-15-1-4018, Australian Research Council Project DP180101985, and a Google PhD Fellowship.

Finally, and most importantly, I would like to express deep thanks to my wife Rongqin Tan, my parents Suying Deng and Sanji Wu, and my extended family, who have all given their unreserved supports to me in so many years. May my grandparents rest in peace. I love you and always will.

Abstract

The production and consumption of online content have been increasing rapidly, whereas human attention is a scarce resource. Understanding how the content captures collective attention has become a challenge of growing importance. In this thesis, we tackle this challenge from three fronts – quantifying sampling effects of social media data; measuring engagement behaviors towards online content; and estimating network effects induced by the recommender systems.

Data sampling is a fundamental problem. To obtain a list of items, one common method is sampling based on the item prevalence in social media streams. However, social data is often noisy and incomplete, which may affect the subsequent observations. For each item, user behaviors can be conceptualized as two steps – the first step is relevant to the content appeal, measured by the number of clicks; the second step is relevant to the content quality, measured by the post-clicking metrics, e.g., dwell time, likes, or comments. We categorize online attention (behaviors) into two classes: popularity (clicking) and engagement (watching, liking, or commenting). Moreover, modern platforms use recommender systems to present the users with a tailoring content display for maximizing satisfaction. The recommendation alters the appeal of an item by changing its ranking, and consequently impacts its popularity.

Our research is enabled by the data available from the largest video hosting site YouTube. We use YouTube URLs shared on Twitter as a sampling protocol to obtain a collection of videos, and we track their prevalence from 2015 to 2019. This method creates a longitudinal dataset consisting of more than 5 billion tweets. Albeit the volume is substantial, we find Twitter still subsamples the data. Our dataset covers about 80% of all tweets with YouTube URLs. We present a comprehensive measurement study of the Twitter sampling effects across different timescales and different subjects. We find that the volume of missing tweets can be estimated by Twitter rate limit messages, true entity ranking can be inferred based on sampled observations, and sampling compromises the quality of network and diffusion models.

Next, we present the first large-scale measurement study of how users collectively engage with YouTube videos. We study the time and percentage of each video being watched. We propose a duration-calibrated metric, called relative engagement, which is correlated with recognized notion of content quality, stable over time, and predictable even before a video's upload.

Lastly, we examine the network effects induced by the YouTube recommender system. We construct the recommendation network for 60,740 music videos from 4,435 professional artists. An edge indicates that the target video is recommended on the webpage of source video. We discover the popularity bias – videos are disproportionately recommended towards more popular videos. We use the bow-tie structure to characterize the network and find that the largest strongly connected component consists of 23.1% of videos while occupying 82.6% of attention. We also build models to estimate the latent influence between videos and artists. By taking into account the network structure, we can predict video popularity 9.7% better than other baselines.

Altogether, we explore the collective consuming patterns of human attention towards online content. Methods and findings from this thesis can be used by content producers, hosting sites, and online users alike to improve content production, advertising strategies, and recommender systems. We expect our new metrics, methods, and observations can generalize to other multimedia platforms such as the music streaming service Spotify.

Publications, Software, and Data

The majority of the thesis has been published in peer-reviewed conference proceedings. Software and data developed as a part of this thesis are provided for reproducing experiment results and facilitating future work.

Publications

- Minjeong Shin*, Alasdair Tran*, **Siqi Wu***, Alexander Mathews, Rong Wang, Georgiana Lyall, and Lexing Xie. "AttentionFlow: Visualising Dynamic Influence in Ego Networks." *ACM International Conference on Web Search and Data Mining (WSDM)*, 2021. (Demo | Chapter 5)
- Lu Cheng, Kai Shu, **Siqi Wu**, Yasin N. Silva, Deborah L. Hall, and Huan Liu. "Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model." *ACM International Conference on Information and Knowledge Management (CIKM)*, 2020. (Full paper, acceptance rate: 21%)
- **Siqi Wu**, Marian-Andrei Rizoiu, and Lexing Xie. "Variation across Scales: Measurement Fidelity under Twitter Data Sampling." *AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2020. (Full paper, acceptance rate: 17% | Chapter 3)
- **Siqi Wu**, Marian-Andrei Rizoiu, and Lexing Xie. "Estimating Attention Flow in Online Video Networks." *ACM International Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2019. (Best paper honorable mention award, acceptance rate: 31% | Chapter 5)
- **Siqi Wu**. "How is Attention Allocated? Data-driven Studies of Popularity and Engagement in Online Videos." *ACM International Conference on Web Search and Data Mining (WSDM)*, 2019. (Doctoral consortium)
- **Siqi Wu**, Marian-Andrei Rizoiu, and Lexing Xie. "Beyond Views: Measuring and Predicting Engagement in Online Videos." *AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2018. (Full paper, acceptance rate: 16% | Chapter 4)

- Quyu Kong, Marian-Andrei Rizoiu, **Siqi Wu**, and Lexing Xie. "Will This Video Go Viral? Explaining and Predicting the Popularity of YouTube Videos." *International Conference on World Wide Web Companion (WWW)*, 2018. (Demo | Chapter 4)

Preprints

- Ali Mert Ertugrul*, **Siqi Wu***, Jooyoung Lee*, Lexing Xie, and Yu-Ru Lin. "Linking Collective Attention Across Platforms: Do More Tweets Beget More Video Views?" *Under revision*.
- **Siqi Wu** and Paul Resnick. "Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives." *Under review*.

Software

- Twitter-intact-stream (Chapter 3): a Python package to reconstruct the complete Twitter filtered stream.

  https://github.com/avalanchesiqi/twitter-intact-stream
- YouTube-insight (Chapter 4): a Python package to collect metadata and historical data for YouTube videos.

  https://github.com/avalanchesiqi/youtube-insight
- HIPie (Chapter 4): a web interface to explain and predict the popularity of YouTube videos.

  http://www.hipie.ml/
- AttentionFlow (Chapter 5): a web interface to visualize a collection of time series and the dynamic network influence among them.

  http://www.attentionflow.ml/

Data

- Complete/Sampled Retweet Cascades Datasets (Chapter 3) include 2 sets of complete/sampled retweet cascades on the topics of cyberbullying (sampling rate: 52.72%, 3M complete cascades, 1.17M sampled cascades) and YouTube video sharing (sampling rate: 91.53%, 2.02M complete cascades, 1.8M sampled cascades).

  https://tinyurl.com/rguqamr

- YouTube Engagement '16 Datasets (Chapter 4) include (1) a tweeted videos dataset, which contains 5M YouTube videos that are uploaded and tweeted from 2016-07-01 to 2016-08-31, and are watched at least 100 times within 30 days of their onsets; (2) three quality videos datasets, which contain 96K videos deemed of high quality by domain experts. To our knowledge, they are the only publicly available datasets containing information of video watch time.

  https://tinyurl.com/wswvtbj

- Vevo Music Graph Dataset (Chapter 5) contains the metadata and historical data of 60,740 YouTube videos from 4,435 Vevo artists who are active in English-speaking countries, and 63 daily snapshots of the video recommendation network.

  https://tinyurl.com/tqzeaps

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Online content has shown a tremendous increase in both content production and attention consumption. In the era of information overload, while users have an unprecedented volume of products to choose from, the products in turn compete for their limited attention [Weng et al., 2012; Zarezade et al., 2017]. It has become increasingly difficult for the users to differentiate a set of valuable information sources, which also hinders them from allocating attention efficiently.

The inefficient attention allocation can be attributed to two reasons. Firstly, online content is overwhelmingly presented. For example, more than 500 million tweets are sent on Twitter every day [twitter.com, 2013] and more than 500 hours of videos are uploaded on YouTube every minute [tubefilter.com, 2019]. Secondly, the negative effects of platforms' proprietary algorithms remain an active matter of debate. For example, the recommender systems have been criticized for exposing users to a narrower spectrum of content over time, creating a filter bubble phenomenon [Resnick et al., 2013; Nguyen et al., 2014]. Therefore, understanding how the content captures human attention is a fundamental step towards building responsible platforms.

Instead of focusing on each individual user, we take an item-centric approach. We conceptualize the behaviors of consuming a digital product as two steps – the first step is relevant to the content appeal, measured by the number of clicks or views; the second step is relevant to the content quality, measured by the post-clicking metrics, e.g., dwell time, likes, or comments. The two steps characterization was also adopted in the MIT MusicLab experiment [Salganik et al., 2006; Krumme et al., 2012]. Based on it, online attention (behaviors) can be categorized into two classes: popularity (clicking) and engagement (watching, liking, or commenting). Intuitively, the notions of popularity and engagement respectively describe the decision to click on an item and the decision to interact after clicking.

In this thesis, we use the largest video hosting platform YouTube as a lens to study the collective user behaviors. YouTube currently ranks the second in the most-visited websites [alexa.com, 2020] and attracts over a billion hours watch time every day [youbube.com, 2017]. Broadly, online videos account for 73% of internet traffic in 2016, and they are projected to have 82% of all traffic by 2021 [cisco.com, 2017]. We tackle three research questions related to attention consumption on YouTube videos from the perspectives of sampling, engagement, and network effects.

Our first question quantifies **the effects of social data sampling on widely-used attention measures**. Data sampling is a common yet fundamental problem in social media studies [Morstatter et al., 2013; Olteanu et al., 2019]. Most platforms deploy a request quotation system to avoid malicious attacks. For example, the prevailing data source Twitter allows 15 to 900 requests every 15 minutes [twitter.com, 2020f], which is often inadequate to construct a complete, unsampled dataset. The data sampling introduces noises and biases [Boyd and Crawford, 2012; Tufekci, 2014]. Hence, researchers must be aware and take account of hidden noises in their curated datasets for drawing rigorous scientific conclusions.

Our second question measures and predicts **the collective engagement patterns towards online content**. While there has been a rich body of literature studying online content, current research extensively focuses on measuring and modeling the popularity metrics [Pinto et al., 2013; Rizoiu et al., 2017b]. On the other hand, engagement measures, sometimes referred to as “active participation” [Khan, 2017] or ”post-clicking behaviors” [Yi et al., 2014], remain understudied in academia despite becoming core metrics in practice [youtube.com, 2012; facebook.com, 2017]. Considering video watch time or webpage dwell time, the audience may immerse in the content once they click on the item, or quickly abandon it.

Our third question measures and models **the network effects induced by the recommender systems**. Many modern platforms provide algorithmic suggestions to help users explore the enormous content space. Users may first react to exogenous stimuli such as breaking news events and social media promotions [Lehmann et al., 2012; De Choudhury et al., 2016]. Their attention is then amplified and steered through the platforms’ recommender systems, creating an endogenous effect [Rizoiu and Xie, 2017]. On YouTube, despite that the recommender system accounts for 70% watch time [cnet.com, 2018], little is known about how the system drives user attention and how the induced network affects video popularity.

We show one example for illustrating this discrepancy between popularity and engagement metrics. Figure 1.1 (a) is the Trump[1] inauguration speech video and Figure 1.1 (b) is one of the most viewed Ice Bucket Challenge[2] (IBC) videos. In terms of the popularity metric, the Trump video has 4.1M views while the IBC video is far more popular with 31M views. However, the Trump video attracts more engaging

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

[2] https://en.wikipedia.org/wiki/Ice_Bucket_Challenge

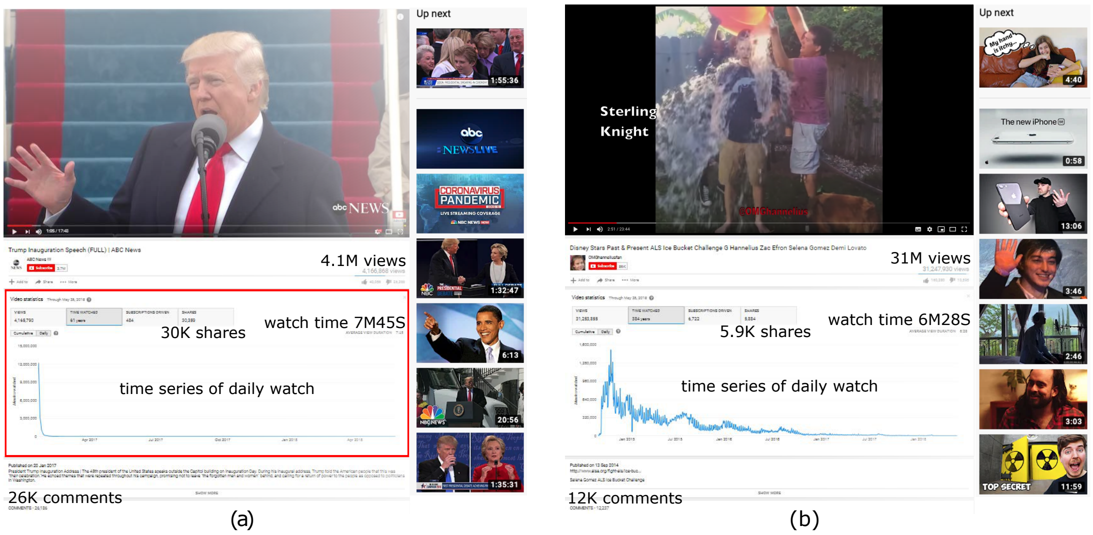

Up next
Trump Inauguration Speech (FULL) | ABC News
4.1M views
30K shares
watch time 7M45S
time series of daily watch
26K comments
(a)
Sterling Knight
Disney Stars Past & Present ALS Ice Bucket Challenge ft Hannelius Zac Efron Selena Gomez Demi Lovato
31M views
5.9K shares
watch time 6M28S
time series of daily watch
12K comments
(b)

Figure 1.1: Screenshots of two YouTube videos, their metadata, insight data, and recommended videos. **(a)** "Trump Inauguration Speech (FULL) | ABC News" (source: https://www.youtube.com/watch?v=sRBsJNdK1t0); **(b)** "Disney Stars Past & Present ALS Ice Bucket Challenge" (source: https://www.youtube.com/watch?v=TCY0_tP_cbk). The screenshots were taken in May, 2018. The insight dashboard (highlighted in red box) is deprecated in current YouTube interface.

behaviors with 26K comments and 30K shares. Additionally, even the Trump video is 6 minutes shorter than the IBC video, the audience still spend on average 1 more minute watching it. This example depicts that videos with a higher number of views do not necessarily lead to more interacting behaviors. It also highlights a significant challenge that many researchers face – that of choosing an appropriate measure and understanding the interplay of measures from different dimensions.

On the right-hand panel of each video page, a list of recommended videos is generated by YouTube recommender systems. From this Trump video, users may continue to watch the same content from another cable news (1st position), or the presidential debate between Trump and Clinton (4th position), or Obama's speech (5th position). The recommender system effectively provides mechanical pathways for user attention to flow, and consequently changes the video popularity.

1.1 Thesis overview

This thesis examines the three research questions in greater detail. In Chapter 2, we present a broad introduction of online attention. We start with the attention theories derived in economics and cognitive science, and then use a supply and demand framework to demonstrate the competition of attention. Next, we review recent advances in measuring collective behaviors in complex social systems from the fronts of social data sampling, popularity and engagement measures, and recommender systems. We are interested in understanding how users reach the content, what they consume, and how they interact in the information-overloaded world.

In Chapter 3, we address the first research question by presenting a comprehensive study of the Twitter sampling effects on common measurements [Wu et al., 2020]. Because we use Twitter prevalence as a proxy to obtain YouTube videos, it is crucial to first understand the potential sampling effects on attention measures. By constructing two sets of complete and sampled tweet datasets on cyberbullying and YouTube sharing, we show that Twitter rate limit message is an accurate indicator for the volume of missing tweets, and sampling rates vary across different timescales. We find the Bernoulli process with a uniform rate can approximate the empirical entity distribution well. More importantly, the true entity distribution and ranking can be inferred based on sampled observations. In network measures, we observe that the structures are altered with denser components more likely to be preserved. Lastly, sampling compromises the quality of diffusion models since tweet inter-arrival time is significantly longer in the sampled stream, while user influence is lower.

We curate a longitudinal dataset that tracks tweets containing YouTube URLs from 2015 to 2019. On average, more than 3M tweets are collected every day. We extract the associated URLs to acquire YouTube video ids. We develop a new Python package to crawl YouTube data, and use it to construct two large video datasets, which are our basis to study the collective user behaviors in online videos.

In Chapter 4, we address the second research question by presenting the first large-scale measurement study of video engagement on YouTube [Wu et al., 2018]. In contrast to prior work that requires auxiliary toolkits to record user actions [Buscher et al., 2009; Arantes et al., 2016], our collection (5.3M videos) raises the data volume by several orders of magnitude. We study a set of metrics including time and percentage of videos being watched. We observe that video duration is an important covariate on watching patterns. Longer videos generally make the users stay for a longer time but are less likely to keep them watching till the end. To calibrate watching metrics against video duration, we construct a 2-dimensional tool, called engagement map. Based on it, we propose a new metric, called relative engagement, as the watch percentage rank percentile among videos of similar lengths. This metric is closely correlated with recognized notions of content quality. Moreover, we find that engagement measures are stable over time and predictable even before a video's upload. They have most of the variance explained by video context, topics and channel information at coefficient of determination R^2 of 0.77. Using a Hawkes process model to forecast the video attention dynamics, we find the time series of the engagement metric (daily watch time) is more predictable than that of the popularity metric (daily view count). The result is significant as it separates the concerns for modeling engagement and popularity – the latter is known to be unpredictable aprior and driven by external promotions [Kong et al., 2018].

In Chapter 5, we address the third research question by presenting the first large-scale measurement study on the network effects induced by YouTube recommender systems [Wu et al., 2019]. We construct the content recommendation network for 60,740 music videos from 4,435 professional artists. An edge indicates that the target video is recommended on the webpage of source video. Our work is motivated by a few key observations that the recommender systems drive user attention, especially when a blockbuster video is uploaded or when a breaking news event arises. By systematically measuring the entire network, we find that videos are disproportionately recommended towards more popular videos. This means the recommender system is likely to take a random viewer to more popular videos and keep them there, thus reinforcing the "rich get richer" phenomenon. Furthermore, we use the bow-tie structure [Broder et al., 2000] to characterize the recommendation network. We find that its core component (23.1% of the videos) occupies most of the attention (82.6% of the views). This is indicative of the connection between video recommendation and the inequality of user attention allocation. Finally, we estimate the attention flow in the video recommendation network. We propose a new model, called ARNet, which accounts for the network structure and can predict video popularity 9.7% better than other baselines. The ARNet model also allows us to identify a group of artists who gain significant attention from the recommender systems. Furthermore, we develop a new demo called AttentionFlow [Shin et al., 2021] to visualize the effects of recommendation for videos and artists in the YouTube network.

Finally, we summarize our work and present a number of interesting future directions in Chapter 6.

1.2 Key contributions and impact

The key contributions of this thesis include new observations, methods, metrics, datasets, software, and web demonstrations.

- **New observations and methods on 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) The Bernoulli process approximates the empirical entity distribution well. (4) Sampling compromises the quality of network and diffusion models. (5) A new method to infer true entity statistics (e.g., missing volume, entity distribution, and ranking) based on sampled observations.

- **New metrics and observations on collective engagement patterns.** (1) A new tool called engagement map to capture the nonlinear relationship between video length and watch patterns. (2) A new metric – relative engagement – that calibrates against video length, correlates with video quality, and appears stable over time. (3) Engagement metrics can be predicted in a cold-start setup, achieving R^2=0.77.

- **New observations and methods on recommendation network effects.** (1) The first large-scale characterization of the video recommendation network intersecting with video attention consumption. (2) Popularity bias – videos are disproportionately recommended towards more popular videos. (3) A new model called ARNet that accounts for the network structure to predict video popularity and to estimate the network contribution between videos and artists.

- **Large-scale datasets.** We curate and release two YouTube datasets. (1) YouTube Engagement '16 dataset contains 5.3M videos published and tweeted between July and August, 2016. (2) Vevo Music Graph dataset contains 60K music videos with 63 daily snapshots of the video recommendation network. We also release two sets of Complete/Sampled Retweet Cascades datasets on the topics of cyberbullying and YouTube sharing.

- **Open software.** We release two new data collection tools. (1) Twitter-intact-stream, for reconstructing the complete filtered stream on Twitter; (2) YouTube-insight, for collecting metadata and historical data for videos on YouTube.

- **Web demonstrations.** We build two new web demonstrations. (1) HIPie, for explaining and predicting the popularity of YouTube videos; (2) AttentionFlow, for visualizing a collection of time series and the dynamic network influence.

Overall, a better understanding of how online content attracts human attention provides us with a set of useful tools to improve user experience in online platforms. The observations of the sampling study can help researchers be aware of and mitigate hidden noises in social media datasets. The observations of the engagement study can help content producers choose engaging topics to create better products, and help hosting sites prioritize quality products in recommender systems. The observations of the network effects study can help content owners understand how traffic is driven for better promotion strategies, help hosting sites combat social optimization, and help online users be conscious of the relevance, novelty and diversity trade-offs in the content they are recommended to.

Looking forward, the popularity bias in the recommendation network sheds light on building a responsible online platforms. But we still lack knowledge of how the recommender systems change a user's cognition, stance and behavior. Future research should include a user-centric qualitative study that surveys user cognition on the effects of recommender systems, and a quantitative study that uses auditing methods to examine the biases in the recommendation network.

Chapter 2

Related work

The collective user behaviors that we investigate in this thesis have been studied in multiple disciplines including computer science, sociology, and economics. Firstly, we provide a brief introduction to online attention in Section 2.1. Next, Section 2.2 details the usage of social media APIs and their sampling effects. Guiding by the two steps framework, we review work studying the engagement and popularity patterns in Section 2.3 and Section 2.4, respectively. Lastly, we discuss recommender systems and their effects on driving user attention in Section 2.5.

2.1 The anatomy of online attention

In psychology, attention is a cognitive process of selectively concentrating on a specific piece of information [James, 2007]. American economist, Nobel Prize and Turing Award winner Herbert Simon articulated the concept of *attention economy*, in which he pointed out that attention is the limiting factor for information consumption since human beings cannot digest all the information [Simon, 1971]. In modern society, attention becomes a scarce commodity. Users need to allocate their attention efficiently to avoid getting lost in a wealth of information.

Competition for finite attention. The attention competition is exacerbated with the explosive growths of online products and social media content. Whenever a novel item occurs, it often captures immediate attention. This effectively reduces the attention to other items, leading to inattentional blindness [Chabris and Simons, 2010]. Wu and Huberman [2007] studied the dynamics of collective attention in Digg stories. They observed a natural time scale over which the attention fades. The dynamics can be described by an exponentially decaying model characterized by a single novelty factor. Weng et al. [2012] investigated how the competition shapes the spread of information, especially on content popularity, diversity, and lifetime. Valera and Gomez-Rodriguez [2015] modeled and illustrated the impact of social influence on the adoption of competing products.

The attention competition also exists in social networks. First proposed by British anthropologist Robin Dunbar, the *Dunbar's number* suggests the cognitive limit of the number of people with whom one can maintain stable relationships [Dunbar, 1992]. He found a correlation between primate brain size and average social group size. The Dunbar's number for human beings is often perceived around 150. However, the prevailing usage of social media really boosts the number of relationships one can possible make. Although initially obtained in an offline experimental setting, can the Dunbar's number generalize to online relationships?

Gonçalves et al. [2011] validated the Dunbar's number on a large Twitter networks. They found that the empirical data is in agreement with Dunbar's result: users can only maintain a close relationship circle of 100-200 people. On another social media platform Facebook, Backstrom et al. [2011] analyzed how the users balance their attention among social contacts. They found that communication-based activities (e.g, messages and comments) are much more focused with a higher fraction of attention going towards top contacts, while viewing-based activities (e.g, profile views and photo views) are significantly more dispersed across contacts.

The above research altogether gives an example of saturated attention economy: the ability of information consumption is greatly limited by the finite human attention. This can partly explain many social and economic phenomena, such as "rich get richer" [Piketty, 2015] and "winner takes all" [Giridharadas, 2019].

Strategies to gain attention. Nowadays, most user-generated content (UGC) sites have three stakeholders: content consumers, content producers, and hosting platforms. We can use a supply and demand framework to explain their goals. Consumers have the demand of consuming information and the resources (e.g., time, money) to spend. On the supply end, producers aim at maximizing exposure and monetization by attracting as many consumers as possible; hosting platforms want to keep the consumers satisfied by aligning them with the needed products [Konstan and Riedl, 2012]. Therefore, a surrogate metric of attention is useful and desirable, since it can serve as a quantifiable signal for the producers to improve their productions, and for the platforms to improve their services.

For the above three stakeholders, their strategies for gaining attention vary even though it is intuitive that higher quality products will probably beget more attention. In this thesis, we focus on the perspectives of content producers and hosting sites. Shen et al. [2015] investigated product rating on Amazon and found strategic behaviors among online reviewers. Their results suggest that reviewers are more likely to post reviews for popular but less crowded products, and tend to be more conservative when posting controversial opinions. In terms of the platforms, recommender systems are widely used to attract users' long term attention [Chen et al., 2019].

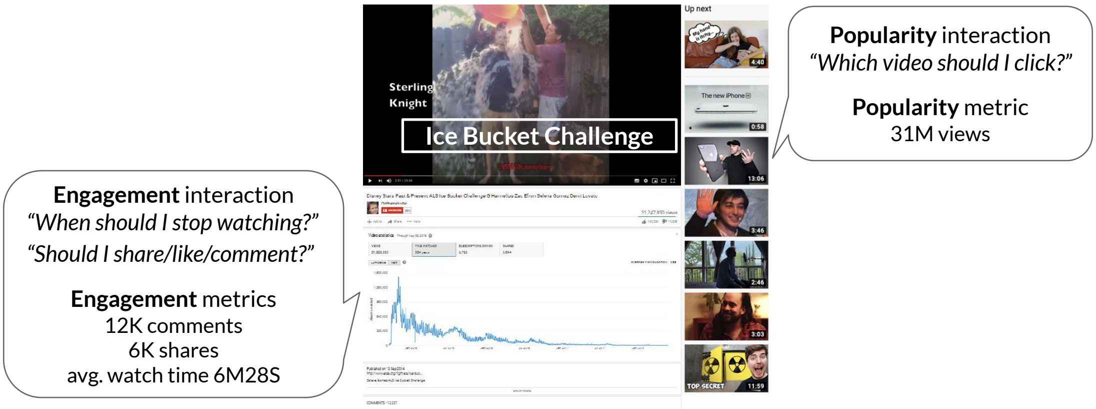


Figure 2.1: A YouTube video webpage and user interactions on it. The notions of popularity and engagement respectively describe the decision to click on a video (right box) and the decision to interact after clicking (left box). (Screenshot source: https://www.youtube.com/watch?v=TCY0_tP_cbk)

Characterizing online attention. Salganik et al. [2006] explored how social influence and inherent quality jointly affect a product's market share in the "MusicLab" experiment. Content quality only partially determines the product success, while an increasing strength of social influence increases the unpredictability of success. Krumme et al. [2012] proposed a two-step framework to characterize how users consume digital items. The first step is based on the product appeal, measured by the number of clicks; the second step is based on the product quality, measured by post-clicking metrics, e.g., dwell time, comments, or shares. The appearance of product is an implementation choice, thus empowering the platforms to manipulate attention allocations by changing the presentation order of items. [Lerman and Hogg, 2014].

In the web search community, a similar framework has also been adopted to differentiate page views from dwell time [Yue et al., 2010]. Stoddard [2015] measured this framework on two social news aggregators, Reddit and Hacker News. Furthermore, Van Hentenryck et al. [2016] showed popularity alone is a poor proxy to represent quality in online market. For example, clickbait may attract users' attention and receive lots of clicks while failing to provide a positive user experience, suggesting that popularity and engagement metrics indeed capture different product properties.

Following this idea, we categorize online attention into popularity and engagement. Figure 2.1 illustrates the user interactions and metrics on a YouTube video. Popularity metric refers to the number of views that a video receives, while engagement metrics refer to the time spent on watching the video, the number of the comments and external sharing.

2.2 Social data sampling

We rely on social media APIs to collect data used in this thesis, specifically, Twitter and YouTube APIs. YouTube API does not subsample the data, yet it requires an input field of video id. Because we use collected tweets as a proxy to obtain YouTube video ids, it is crucial to first understand Twitter data sampling.

Twitter APIs. Twitter has different levels of access (Firehose, Gardenhose, Spritzer) and different ways to access (search API, sampled stream, filtered stream). As the complete data service (Firehose) incurs excessive costs and requires severe storage loads, here we only discuss the free APIs.

- Twitter search API returns relevant tweets for a given query, but it only fetches results published in the past 7 days [twitter.com, 2020d]. The search API also bears the issue of data attrition. Research using this API to construct a "complete" dataset would inevitably miss parts of desired tweets [Wang et al., 2015] since tweet creation and deletion are highly dynamic [Almuhimedi et al., 2013]. To overcome this limitation, researchers can pivot to the streaming APIs, which return public tweets at the time of their creation.

- Twitter sampled streaming API returns roughly 1% of all public tweets in realtime [twitter.com, 2020c]. Pfeffer et al. [2018] detailed its sampling mechanism and identified potential tampering behaviors. Ghosh et al. [2013] compared the 1% sampling with expert sampling. They observed that elite users share more trustworthy content and react faster to breaking events. González-Bailón et al. [2014] examined the biases in the retweet network from the 1% sample and the search API. Some researchers found that the 1% sample can be treated as a representative sample of all Twitter activities since the hashtag frequencies from 1% sample largely overlay (within 3 standard deviations) with the bootstrapped random samples from the complete Firehose stream [Morstatter et al., 2014]. However, it is worth noting that data filtering can only be conducted after the data is collected. Therefore, the sampled streaming API is not suitable to create ad hoc datasets, e.g., tracking *all* tweets that contain the hashtag #coronavirus.

- Twitter filtered streaming API collects tweets matching a set of prescribed predicates in realtime [twitter.com, 2020a]. Suppose that the streaming rate is below Twitter limit, the pre-filtering makes the filtered stream possible to construct the complete datasets without using the costly Firehose stream, e.g., on social movements [De Choudhury et al., 2016], on news outlets [Mishra et al., 2016], and on controversial topics [Bista et al., 2019]. We focus on the scenes where the data streams are sampled. The most relevant work is done by Morstatter et al. [2013], in which they compared the filtered stream with the Firehose, and measured the discrepancies in various metrics.

Another important observation is that Twitter sampling is deterministic. Joseph et al. [2014] found no practical differences in tweets seen in different streaming clients, as long as the filtered configurations are the same. Therefore, simply stacking crawlers with the same predicates will not yield more data. However, users can improve the sample coverage by splitting the keyword set into multiple disjoint predicate sets, and monitoring each set with a distinct subcrawler. Sampson et al. [2015] successfully inflated the volume of collected tweets by 10 times through 20 subcrawlers.

Effects of missing social data. The data quality problem has received growing attention in academic studies. Social data, which records ubiquitous human activities in digital form, plays a fundamental role in social media research. Boyd and Crawford [2012] pointed out the necessity to interrogate the assumptions and biases in data. Ruths and Pfeffer [2014] discussed the biases and flaws in social media data. Tufekci [2014] outlined five issues on data representativeness and validity. The hidden data biases may alter some research conclusions and even impact human decision making [Olteanu et al., 2019].

Sampling from graphs and cascades. Leskovec and Faloutsos [2006] studied different graph sampling strategies for drawing representative samples. Wagner et al. [2017] considered how sampling impacts the relative ranking of groups in the attributed graphs. The effects of graph sampling have been extensively discussed by Kossinets [2006]. In a retweet graph, the missing tweets can cause edge weights to decrease, and some edges to even disappear. On sampling a cascade, De Choudhury et al. [2010] found that combining network topology and contextual attributes distorts less the observed metrics. Sadikov et al. [2011] proposed a k-tree model to uncover some properties from the sampled data. They both sampled the cascades via different techniques (e.g., random, activity-based, forest fire) and varying ratios.

In this thesis, we do not design new sampling mechanism, instead, we study the sampling effects of Twitter's proprietary algorithm. Based on a widely-used Reddit corpus, Gaffney and Matias [2018] identified and suggested strong risks in research that concerns user history or network information, and moderate risks in research that uses aggregate counts. We use these qualitative observations as starting points and conduct a set of in-depth quantitative measurements in Chapter 3. We corroborate the risks in user history study and network analysis and further extend the scope of measured subjects. Moreover, we use the sampled observations to estimate the unobserved, complete entity statistics. Our observations are of great importance to researchers who use Twitter data for empirical measurement and user modeling.

2.3 User engagement

Many researchers have analyzed user engagement behaviors towards web content. For example, the line of work that measures webpage reading patterns often exploits auxiliary toolkits such as mouse-tracking [Arapakis et al., 2014; Lagun and Lalmas, 2016] or eye-tracking [Buscher et al., 2009] instrumented browser. Dwell time, which is conceptually close to video watch time, has been widely used in the domains of web search and recommendation to improve retrieval performance [Dupret and Lalmas, 2013; Lalmas et al., 2015].

In recommender systems, Yi et al. [2014] compared two systems that optimize for clicks and dwell time, and found the one using dwell time achieves better performance on ranking relevant products. On YouTube, both explicit (e.g., rating) and implicit (e.g., watch time) feedback signals are vital for recommending satisfied items to users [Davidson et al., 2010; Covington et al., 2016]. More recently, YouTube switched to promote videos that can keep the audience watching for longer time rather than these optimizing for clicks [youtube.com, 2012]. The same adjustment was also seen in Facebook videos [facebook.com, 2017].

Individual versus collective measurement. Most of the above research is user-centric as they study engagement for each individual user. User activity trajectories (e.g., search history, watch history) are often assumed available in this approach. This information is obtained either from internal logs or from crowdsourcing platforms such as Amazon Mechanical Turk (MTurk).

Figueiredo et al. [2014] asked 72 MTurk users to rate pairs of YouTube videos, and found that in most evaluations users could not reach consensus on which video is of better quality. Arantes et al. [2016] monitored a campus network and analyzed how users interact with video-ad on YouTube. Sun et al. [2017] interviewed a small group of recruited participants on how they watch a video together. Swart et al. [2020] surveyed 300 MTurk users to evaluate if advertising banner on YouTube is noticeable. Survey, interview, field deployment, and diary study are important means of user-centric studies in human-computer interaction (HCI). However, they are unlikely to provide quantitative conclusions due to limited data size, typically ranging from a dozen to several hundred.

Unfortunately, user-level data is generally inaccessible on YouTube, even the content owners only observe an aggregate analytic report. To this end, researchers can take an item-centric approach as the aggregate statistics for each item is public and can be crawled by automated scripts [Yu et al., 2015].

Dobrian et al. [2011] correlated user engagement on video watching with network quality, e.g., join time, buffering ratio, rendering quality. Guo et al. [2014] presented a case study of 4 pre-selected edX courses, where mixed methods were applied – mining edX server logs and interviewing course designer. This approach is infeasible at the scale of system level and its implications are limited within the area of educational videos. The most relevant work to this thesis is from Park et al. [2016], who measured watch time of a small set of YouTube videos, and showed the predictive power of collective reactions.

Modeling user engagement. Guo and Agichtein [2012] monitored cursor moving and scrolling to estimate document relevance on webpage. Drutsa et al. [2015] used gradient boosting decision tree model to predict user browsing behaviors in search engine. Barbieri et al. [2016] used survival analysis techniques to estimate the distribution of time that users will spend on online advertisements. Dupret and Lalmas [2013] argued that solely relying dwell time is not enough, the time between two consecutive user visits, or the absence time, should also be taken into consideration for modeling engagement.

On online videos, Chen et al. [2013] correlated watch time to video length, type, and popularity measures. Their model is a simple concatenation of multiple linear components. Park et al. [2016] showed that watch percentage is positively associated with the view count, the number of likes per view, and perhaps most surprisingly, the negative sentiment in the comments. One drawback of these features is that they require observing videos for some period of time. Yet, a large fraction of videos do not have comments [Cheng et al., 2008], making this prediction setup inapplicable to any random YouTube video. Tong et al. [2020] studied video engagement from a neuroscience viewpoint. They found that brain activities are informative for predicting when users stop watching.

Complementary to the studies on engagement behaviors for individual user, our work in Chapter 4 focuses on measuring and modeling content engagement at the aggregate level and at large scale. We propose a new metric, called relative engagement, and we find it closely correlates with recognized notion of quality. We also show that the aggregate engagement is stable throughout a video's lifetime, and it can be predicted before the video gathers any view or comment.

2.4 Item popularity

Individual user actions on social media platforms give rise to complex phenomena at the aggregate level (e.g., spiky, irregular, seasonal popularity). Cha et al. [2007] are among the first to observe the long-tail distribution of popularity on YouTube, in which they explain by the preferential attachment effect. Gill et al. [2007] analyzed the YouTube network traffic inside the campus of the University of Calgary. They found that the viewing patterns for videos vary significantly by the time of day and day of week. Zhou et al. [2010] revealed that internal search and video recommendation are the two most important traffic sources for video popularity. Figueiredo et al. [2011] characterized the growth patterns of video popularity. In particular, top videos often get most of the views much earlier in their lifetimes.

Modeling item popularity. Popularity dynamic is the most studied attribute of online products. One line of work aims at predicting the volume of final popularity [Szabo and Huberman, 2010]. On the other hand, many researchers focus on predicting the shape (e.g., time series) of future popularity [Figueiredo et al., 2016].

A number of models have been proposed to describe it, such as a series of endogenous relaxations [Crane and Sornette, 2008] or multiple power-law phases [Yu et al., 2015]. Other studies link popularity dynamics to epidemic contagion [Bauckhage et al., 2015; Kong et al., 2020b], external stimulation [Yu et al., 2014] or geographic locality [Brodersen et al., 2012]. We broadly categorize the approaches into four directions: time series analysis, feature-driven model, point process, and deep learning.

- Time series analysis is based on the autocorrelation between past observations and future trend. Szabo and Huberman [2010] showed that early viewing pattern is a strong predictor for future views and the relation appears to be log-linear. Later, Pinto et al. [2013] extended the log-linear model to multivariate linear regression (MLR) to account for the different weights of past observations and a Radial Basis Functions (RBF) kernel MLR to account for the similarity of popularity growth patterns with a set of pre-selected video clusters. Instead of predicting numerical value, Ahmed et al. [2013] discretized popularity into several states and used transition graph to infer hidden popularity states. Autoregressive Integrated Moving Average (ARIMA) is one of the most used methods for modeling time series data. It decouples the trend part and the seasonality part. Gürsun et al. [2011] applied seasonal ARIMA to predict future view counts on YouTube videos.

- Feature-driven approach falls into the regime of traditional machine learning that relies on feature engineering. It achieves the trade-off between predictability and interpretability [Hofman et al., 2017] and often provides instrumental insights. Studying the sharing behaviors of Facebook posts, Cheng et al. [2014] observed that temporal and structural features are key predictors of final popularity size. In contrast, Martin et al. [2016] found that even with unlimited data, the predictive ability in complex social systems would be bounded below a theoretic, deterministic threshold. Nevertheless, both works point out that past success is the most predictive feature. On YouTube, Ma et al. [2017] utilized metadata features from video and its uploader in a regression model to predict lifetime popularity. Abisheva et al. [2014] conducted a cross-platform prediction task that uses the sharing signals on Twitter to predict the video popularity on YouTube.

- Point processes contain a family of generative models that account for the impacts from past events, for example, poisson point process, determinantal point process, Hawkes process, to name a few. They are often associated with a decaying kernel (e.g., power-law or exponential) since the interest towards a new event naturally fades away [Wu and Huberman, 2007]. Zhao et al. [2015] proposed a doubly stochastic poisson process called SEISMIC to model the information diffusion on Twitter. Mishra et al. [2016] integrated content features into a marked Hawkes self-exciting point process to estimate content virality, memory decay, and user influence. Kong et al. [2020a] extended self-exciting processes to dual-mixture processes for characterizing the resharing cascades of online items. Rizoiu et al. [2017b] used Hawkes process to explain the complex popularity dynamics of YouTube videos as two components: exogenous stimuli and endogenous response. For a detailed introduction of Hawkes process on social media, we refer to the tutorials by Rizoiu et al. [2017a].

- Deep learning is an active research field. In particular, recurrent neural network (RNN) with long short-term memory (LSTM) units has achieved state-of-the-art performances in many time series modeling tasks [Kuznetsov and Mariet, 2019]. Li et al. [2017] proposed an end-to-end deep learning framework to predict cascade size on Twitter. They found it is critical to learn how the information is diffused (graph embedding) but not merely who shares the information (node embedding). Zhu and Laptev [2017] used a Bayesian neural network model for both point and uncertainty estimation in Uber trip data. Although deep learning techniques may obtain superior modeling results, the fact that they often fall short of interpretability is undesirable in providing practical instructions to the stakeholders.

Our model in Chapter 5 takes a mixed approach – we extract features from both time series and network. To our knowledge, no prior work has attempted to predict video popularity with fine-grained content recommendation network due to the difficulty in constructing such network. Although the regression-based model is relatively simple, we show that once integrating the network structure, it is possible to beat more sophisticated time series and deep learning models.

2.5 Content recommendation networks

The goals of recommender systems can be summarized as two related yet distinct tasks. The first task is user-centric, i.e., given users' profiles and past activities, finding a collection of items that might interest them [Konstan and Riedl, 2012]. The resulting recommendations, often shown in user homepage feed, can be regarded as the entry point for the user action sequence (also known as a user session). The second task is item-centric, i.e., given the currently visited item, finding a ranked list of relevant items [Zhang et al., 2012; Gomez-Uribe and Hunt, 2016]. This can be regarded as recommending the next item in a sequence of actions.

In the same vein, we conceptualize and explain the behaviors on online platforms – users start the action sequences by latent interests, and their subsequent actions are driven by network effects. The items that are connected by the recommender systems, form a backbone content network of user navigation pathways.

Recommender systems on YouTube. Recommender systems, along with YouTube search, have been shown as the two dominant factors driving user attention on YouTube [Zhou et al., 2010]. In 2010, Davidson et al. [2010] reported the usage of a collaborative filtering method in the YouTube recommender systems, i.e., videos are recommended by counting the number of co-watches. This approach works well for videos with many views, however, it is less applicable for newly uploaded videos or least watched videos. Bendersky et al. [2014] proposed two methods to enhance the collaborative filtering approach by embedding the video topic representation into the recommender. Covington et al. [2016] applied deep neural networks and indicated that the final recommendation is a top-K sample from a large candidate set generated by taking into the account content relevance, past watch and search activities, etc. Other enhancements include incorporating contextual data [Beutel et al., 2018]. Chen et al. [2019] and Ie et al. [2019] showed success in applying reinforcement learning techniques in YouTube recommender systems. Most recently, fairness and responsibility in recommender systems have received significant amount of attention [Wilhelm et al., 2018; Beutel et al., 2019; Yi et al., 2019].

Our work does not deal with designing a recommender system, nor does it attempt to reverse engineer the YouTube recommender. Instead, we concentrate our analysis on the impacts of the recommender systems by presenting large-scale measurements of the content recommendation network.

Measuring the effects of recommender systems. Contrasting the extensive literature on evaluating the accuracy of recommendation [Zhang et al., 2012; Lalmas et al., 2015; Li et al., 2018], we focus on prior work that connects network structure with content consumption. Gal Oestreicher-Singer and her collaborators have presented a series of work on Amazon book recommendation network. Firstly, Oestreicher-Singer and Sundararajan [2012] demonstrated the demand effects of recommendation networks. Dhar et al. [2014] further showed the effectiveness of using the recommendation network in predicting item demands. Lastly, Carmi et al. [2017] reported how the book sales react to exogenous demand shocks – not only does the sales increase for the featured item, but the increase also propagates a few hops away by following the links created by the recommender systems.

Su et al. [2016] linked the aggregate effects of recommendations and network structure, and found that popular items profit substantially more than the average ones. However, Sharma et al. [2015] stressed the difficulty of inferring causal relations based on observational data in recommender systems. Zannettou et al. [2018] investigated a set of clickbait videos on YouTube. Surprisingly, their analysis suggested that YouTube recommender fails to take into account the factors of clickbait.

Cheng et al. [2008] are among the first to study the popularity statistics of YouTube recommender systems. They scraped video webpages to construct the video network at a weekly interval. Airoldi et al. [2016] followed the video suggestions on YouTube to construct one static network snapshot for a random collection of music videos.

Note that both studies adopt a snowball sampling technique to construct the network, whereas in Chapter 5 of this thesis, we have the complete trace of an easily identifiable group of Vevo artists. The data collection method allows us to find the less connected group in the network. We capture the dynamics of network snapshots at a much finer daily granularity. Since presenting an item does not guarantee the consumption of item, our work is significant as we link the content network with the attention dynamics. We discover the popularity bias in YouTube recommender systems – videos are disproportionately recommended to more popular videos. We also quantify the unequal attention allocation that the largest strongly connected component (23% of videos) attract 82% of all attention.

Chapter 3

Quantifying sampling effects of online social data

A comprehensive understanding of data quality is the cornerstone of measurement studies in social media research. Many researchers rely on public and free application programming interfaces provided by the hosting platforms to curate datasets. However, the process should be taken with caution, since defective data collection pipelines may introduce noises and potential biases in social data, effectively altering the observations and conclusions.

In this chapter, we present 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 in Section 3.2. 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 (Section 3.3). 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 (Section 3.4). For networks on Twitter, their structures are altered significantly and some components are more likely to be preserved (Section 3.5). For retweet cascades, we observe changes in distributions of tweet inter-arrival time and user influence, which will consequently affect models that rely on these features (Section 3.6). Altogether, the work in this chapter provides a few practical tools to measure Twitter sampling effects.

3.1 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 noises 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[2], 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 chapter, we focus on empirically quantifying the data noises resulted from the sampling in the filtered stream and its impacts on common measurements.

Our 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 et al., 2014; Pfeffer et al., 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

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

[2] We list the papers and their used Twitter APIs in Section A.1.

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 messages closely approximate the volume of missing data. We also find that sampling rates have distinct temporal variations across different timescales, especially at the level of hour and millisecond.

Addressing the second question, we measure the effects of Twitter data sampling across different subjects, e.g., the entity frequency, entity ranking, user-hashtag bipartite graph, retweet network, and retweet cascades. We find that (1) the Bernoulli process with a uniform rate can approximate the empirical entity distribution well; (2) the ranks of top entities are distorted; (3) the true entity frequency and ranking can be inferred based on sampled observations; (4) the network structures change significantly with some components more likely to be preserved; (5) sampling compromises the quality of diffusion models as the distributions of tweet inter-arrival time and user influence are substantially skewed. We remark that this work only studies the effects of Twitter sampling mechanism, but does not intend to reverse engineer it.

The main contributions of this chapter include:

- We show that Twitter rate limit message is an accurate indicator for the volume of missing tweets.
- A set of measurements on the Twitter sampling effects across different timescales and different subjects.
- We show how to estimate the entity frequency and ranking of the complete data using only the sample data.
- We release a software package "Twitter-intact-stream" for constructing the complete data streams on Twitter[3].

[3] The data collection package is available at https://github.com/avalanchesiqi/twitter-intact-stream. Analysis code and collected data are available at https://github.com/avalanchesiqi/twitter-sampling.

	Cyberbullying		YouTube	
	complete	sample	complete	sample
#collected tweets	114,488,537	60,400,257	53,557,950	49,087,406
#rate limit messages	3,047	1,201,315	3,061	320,751
#estimated missing tweets	42,623	54,175,503	77,055	4,542,397
#estimated total tweets	114,531,160	114,575,760	53,635,005	53,629,803
mean sampling rate	99.96%	52.72%	99.86%	91.53%

Table 3.1: Summary of Cyberbullying and YouTube datasets.

3.2 Datasets and Twitter rate limit messages

Constructing complete Twitter data streams. We collect two datasets, using two sets of keywords employed in recent large-scale studies that use Twitter. We choose these works because they are high volume and informative for important social science problems (cyberbullying [Cheng et al., 2020] and online content sharing). We use ρ to denote the sampling rate – i.e., the probability that a tweet is present in the collected (sampled) dataset. We use subscripts to differentiate sampling rates that vary over time ρ_t, users ρ_u, networks ρ_n, and cascades ρ_c. The datasets are collected using the Twitter filtered streaming API and are summarized in Table 3.1.

- Cyberbullying [Nand et al., 2016]: This dataset tracks all tweets that mention any of the 25 recommended keywords from psychology literature. The keywords include nerd, gay, loser, freak, emo, whale, pig, fat, wannabe, poser, whore, should, die, slept, caught, suck, slut, live, afraid, fight, pussy, cunt, kill, dick, bitch. The collection period is from 2019-10-13 to 2019-10-26.
- YouTube [Rizoiu et al., 2017b]: This dataset tracks all tweets that contain at least one YouTube video URL by using the rule "youtube OR (youtu AND be)". The collection period is from 2019-11-06 to 2019-11-19.

The streaming client is a program that receives streaming data via Twitter API. The client will be rate limited if the number of matching tweets exceeds a preset threshold – 50 tweets per second as of May, 2020 [twitter.com, 2020b]. When we use only one client to track all keywords, we find that both datasets trigger rate limiting. We refer to the crawling results from a single client as the *sample set*.

Id	Keywords	Languages
1	should	en
2	should	all\en
3	live	en
4	live	all\en
5	kill, fight, poser, nerd, freak, pig	all
6	dick, suck, gay, loser, whore, cunt	all
7	pussy, fat, die, afraid, emo, slut	all
8	bitch, wannabe, whale, slept, caught	all
complete	subcrawler 1-8	all
sample	all 25 keywords	all

Id	#collected tweets	#rate limit	#est. missing	sampling rate
1	29,647,814	1,357	7,324	99.98%
2	801,904	0	0	100.00%
3	16,526,226	1,273	25,976	99.84%
4	7,926,325	233	7,306	99.91%
5	15,449,973	16	108	100.00%
6	13,164,053	15	125	100.00%
7	21,333,866	89	1,118	99.99%
8	14,178,366	64	666	100.00%
complete	114,488,537	3,047	42,623	99.96%
sample	60,400,257	1,201,315	54,175,503	52.72%

Table 3.2: Subcrawler configurations for Cyberbullying dataset. "all" indicates all 66 language codes on Twitter, and "all\en" is all languages excluding "en".

We develop a software package "Twitter-intact-stream" for constructing the complete data streams on Twitter. The package splits the filtering predicates into multiple subsets, and tracks each set with a distinct streaming client. The Cyberbullying and YouTube datasets are crawled by 8 and 12 clients based on different combinations of keywords and languages[4]. Full specifications for all streaming clients are listed in Table 3.2 and Table 3.3, respectively. We remove the duplicate tweets and sort the distinct tweets chronologically. We refer to the crawling results from multiple clients as the *complete set*.

In very occasional cases, the complete sets also encounter rate limiting. Estimated from the rate limit messages (detailed next), 0.04% and 0.14% tweets in the complete sets are missing, which are negligible comparing to the volumes of missing tweets in the sample sets (47.28% and 8.47%, respectively). For rigorous comparison, we obtain a 30 minutes complete sample from Twitter Firehose and find the difference with our collected data is trivial. Hence, for the rest of this chapter, we treat the complete sets as if they contain no missing tweets.

[4] Twitter currently has 66 language codes: en, es, ja, ko, und, ar, pt, de, tl, fr, cs, it, vi, in, tr, pl, ru, sr, th, el, nl, hi, zh, da, ro, is, no, hu, fi, lv, et, bg, ht, uk, lt, cy, ka, ur, sv, ta, sl, iw, ne, fa, am, te, km, ckb, hy, eu, bn, si, my, pa, ml, gu, kn, ps, mr, sd, lo, or, bo, ug, dv, ca.

Id	Keywords	Languages
1	youtube	en
2	youtube	ja
3	youtube	ko
4	youtube	es
5	youtube	und
6	youtube	all\{en,ja,ko,es,und}
7	youtu AND be	en
8	youtu AND be	ja
9	youtu AND be	ko
10	youtu AND be	es
11	youtu AND be	und
12	youtu AND be	all\{en,ja,ko,es,und}
complete	subcrawler 1-12	all
sample	youtube OR (youtu AND be)	all

Id	#collected tweets	#rate limit	#est. missing	sampling rate
1	10,312,498	323	3,582	99.97%
2	6,620,927	118	3,211	99.95%
3	714,992	36	1,339	99.81%
4	2,106,474	0	0	100.00%
5	1,418,710	0	0	100.00%
6	5,264,150	20	169	100.00%
7	11,188,872	530	10,328	99.91%
8	8,389,060	619	9,657	99.89%
9	4,560,793	1,193	43,584	99.05%
10	2,271,712	27	829	99.96%
11	2,856,415	37	1,556	99.95%
12	7,351,671	158	2,800	99.96%
complete	53,557,950	3,061	77,055	99.86%
sample	49,087,406	320,751	4,542,397	91.53%

Table 3.3: Subcrawler configurations for YouTube dataset.

Validating Twitter rate limit messages. When the streaming rate exceeds the threshold, Twitter API emits a rate limit message that consists of a timestamp and an integer. The integer is designed to indicate the cumulative number of missing tweets since the connection starts [twitter.com, 2020e]. Therefore, the difference between 2 consecutive rate limit messages should estimate the missing volume in between.

We empirically validate the rate limit messages. We divide the datasets into a list of segments where (a) they contain no rate limit message in the complete set; (b) they are bounded by 2 rate limit messages in the sample set. This yields 1,871 and 253 segments in the Cyberbullying and YouTube datasets, respectively. The lengths of segments range from a few seconds to several hours, and collectively cover 13.5 days out of the 14-day crawling windows. In this way, we assure that the segments in the complete set have no tweet missing since no rate limit message is received.

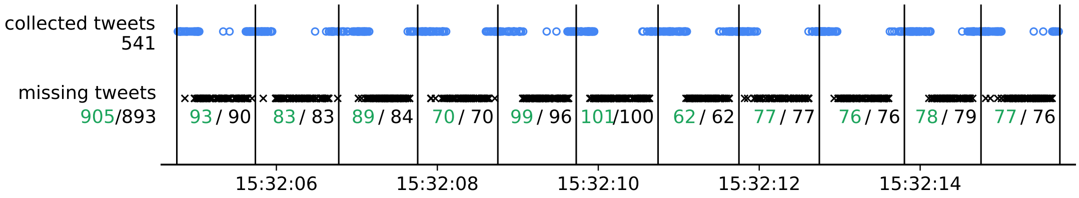


Figure 3.1: Collected and missing tweets in an 11-second interval. blue circle: collected tweet; black cross: missing tweet; black vertical line: rate limit message. green number: estimated missing volume from rate limit messages; black number: count of missing tweets compared to the complete set.

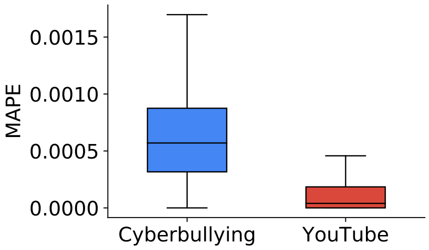


Figure 3.2: MAPE of estimating the missing volumes in the rate limit segments.

Consequently, for each segment we can compute the volume of missing tweets in the sample set by either computing the difference of the two rate limit messages bordering the segment, or by comparing the collected tweets with the complete set.

Figure 3.1 illustrates the collected and missing tweets in an 11-second interval. The estimated missing volumes from rate limit messages closely match the counts of the missing tweets in the complete set. As shown in Figure 3.2, the median error in estimating the missing volume using rate limit messages is less than 0.0005, measured by mean absolute percentage error (MAPE). We thus conclude that the rate limit message is an accurate indicator for the number of missing tweets. Note that it only approximates the volume of missing tweets, but not the content.

Our observations contradict those from Sampson et al. [2015], who used the same keyword-splitting approach, yet found that the rate limit messages give inaccurate estimations. They consistently retrieved more distinct tweets (up to 2 times) than the estimated total volume, i.e., the number of collected tweets plus the estimated missing tweets. In contrast, our datasets only have a small deviation (0.08% and 0.13%, comparing the number of collected tweets in the complete set to the number of estimated total tweets in the sample set in Table 3.1). This discrepancy is due to a different implementation choice back in 2015 – instead of having 1 rate limit message for each second, the rate limit messages were spawned across 4 threads, resulting in

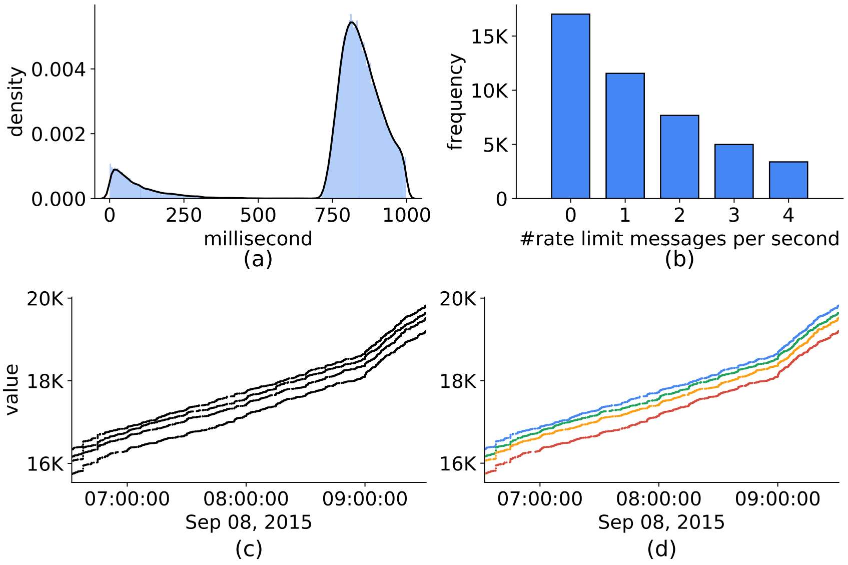


Figure 3.3: (a) The density distribution of milliseconds in the received 55,420 rate limit messages. **(b)** The histogram of the number of rate limit messages received in each second. **(c)** Scatter plot of rate limit messages. x-axis: timestamp; y-axis: values in the rate limit messages. **(d)** Coloring rate limit messages into 4 monotonically increasing threads by using Algorithm 1. All figures are produced based on the SAMPLED '15 dataset.

up to 4 messages per second (detailed next).

A detailed comparison with Sampson et al. [2015] on the rate limit messages. To understand the contradiction, we investigate a sampled dataset crawled within a 30 minutes interval on Sep 08, 2015 (dubbed SAMPLED '15). This dataset contains 66K tweets mentioning YouTube video URLs, and it is also publicly available in our Github repository. We believe that the design of rate limit messages in SAMPLED '15 is the same with that measured in [Sampson et al., 2015]. Differing from the observations we make in the 2019 CYBERBULLYING and YOUTUBE datasets, we notice 2 major differences in SAMPLED '15 dataset, which imply that the rate limit messages were implemented differently back in 2015.

Firstly, SAMPLED '15 receives up to 4 rate limit messages for each second. Figure 3.3(a) shows the milliseconds in the rate limit messages are not uniformly distributed – 89% rate limit messages are emitted between millisecond 700 to 1,000. With a total of 55,420 rate limit messages, Figure 3.3(b) shows that 0 to 4 rate limit messages can be received every second. On the contrary, we obtain at most 1 rate limit message per second in the datasets crawled in 2019. Based on a 5-year tweet tracking dataset we collect, the change was made on Nov 30, 2018.

Secondly, the integers in the rate limit messages are not increasing monotonically. This contradicts Twitter's official documentation that *"Limit notices contain*

```
input  : rate limit messages, a list of intergers $R = [R_1, R_2, \dots, R_n]$.
output: a list of list $A = [A_1, A_2, \dots, A_n]$, in which each list $A_i$ is increasing
          monotonically.
initialize $A_1 = [R_1]$ and $A = [A_1]$;
while not at $R_n$ do
    read the next integer $R_i$ from $R$;
    read the last element of all existing lists $[A_1, A_2, \dots, A_j]$ from $A$ into a list $T$;
    if $R_i \leq min(T)$ then
        append a new list $A_{j+1} = [R_i]$ to $A$;
    else
        find the index $k$ between 1 and $j$, so that $R_i$ has the smallest increment
          against $T[k]$;
        append integer $R_i$ to list $A_k$;
end
```

Algorithm 1: Mapping a list of rate limit messages into multiple monotonically increasing lists.

a total count of the number of undelivered Tweets since the connection was opened [twitter.com, 2020e]". Figure 3.3(c) shows the scatter plot of rate limit messages with the timestamp on the x-axis and the associated integer value on the y-axis. The above observations prompt us to believe that the rate limit messages (and streaming clients) are split into 4 parallel threads rather than 1. When the received messages are less than 4, one explanation could be some threads have streamed all tweets within them.

To estimate the total number of undelivered tweets, we propose Algorithm 1 for mapping rate limit messages to multiple monotonically increasing lists. Note that Algorithm 1 is elastic – if the rate limit messages are monotonically increasing, Algorithm 1 will output only one thread. We color the mapping results in Figure 3.3(d), which shows that 4 separated threads are presented. From the 4,500 rate limit messages received between 2015-09-08 06:30 UTC and 2015-09-08 09:30 UTC, we estimate that 85,720 tweets are missing.

We have noticed that Algorithm 1 would fail when the values in one rate limit thread are constantly smaller than those in another thread. For example, assuming that we have two interweaved threads $A_1 = [1, 3, 5, 7]$ and $A_2 = [2, 4, 6, 8]$, the resulting observed rate limit messages could be $R = [1, 2, 3, 4, 5, 6, 7, 8]$. In this case, the estimated missing number is 8 while the ground-truth missing number is 15. In practice, these interweaved threads occur when the streaming span is extremely short (e.g., the streaming client disconnects and reconnects every few seconds due to unstable Internet). Nonetheless, the short rate limit message streams mean that the discrepancy caused by this limitation would not significantly affect our estimation.

To validate the 4-thread counters for rate limit messages, we obtain a complete

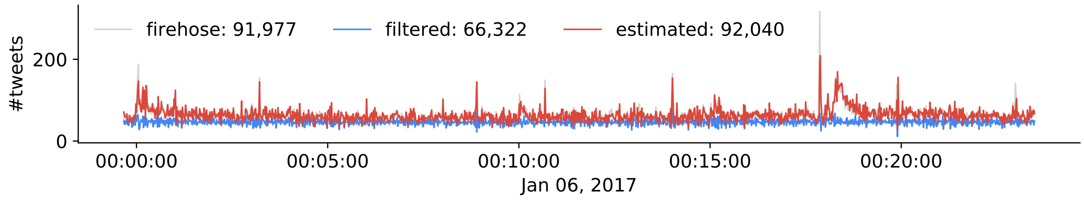


Figure 3.4: Comparing Twitter public filtered stream to the Firehose stream. MAPE is 0.68% with estimated missing tweet volume from rate limit messages.

data sample from a Twitter data reseller **discovertext.com**[5]. This company provides access to the Firehose service at a cost. We started two streaming clients simultaneously on 2017-01-06, one with the Twitter filtered streaming API, the other with the Firehose. We track the temporal tweet volumes at three levels: (1) public filtered stream (blue); (2) filtered stream plus the estimated missing volume from rate limit messages (red); (3) the complete tweet stream from Firehose (grey). The missing volume is estimated by using Algorithm 1, which maps the integers in rate limit messages onto 4 parallel counters. As shown in Figure 3.4, the red line and grey line almost overlap each other (MAPE is 0.68%). The small discrepancy may come from duplicate tweets, or counter-mismatches. Our experiments show that the 4-thread counters are accurate measures for the missing tweet volume in early 2017. Altogether, the collected tweets from public filtered stream plus the estimated missing tweets are close to both Firehose stream and tweets crawled from multiple subcrawlers. This confirms our core assumption that multiple subcrawlers capture the underlying universe of Twitter Firehose stream.

Sampson et al. [2015] used the 1-thread counter to compute the missing volume, in which they only considered one fourth of all rate limit messages. This approach reduces the estimated missing volume to about 25%, and consequently underestimates the total volume. This is the reason Sampson et al. [2015] observed discrepancies when compared the estimated volume from Twitter rate limit messages to the ground-truth Firehose stream.

Although it is unknown when Twitter will adjust its mechanism for signifying missing tweets in the future, as long as the total number of missing tweets are provided as part of a sampled stream, the methodology in this thesis can still be used to estimate sampling bias in different measurements.

[5] https://discovertext.com/

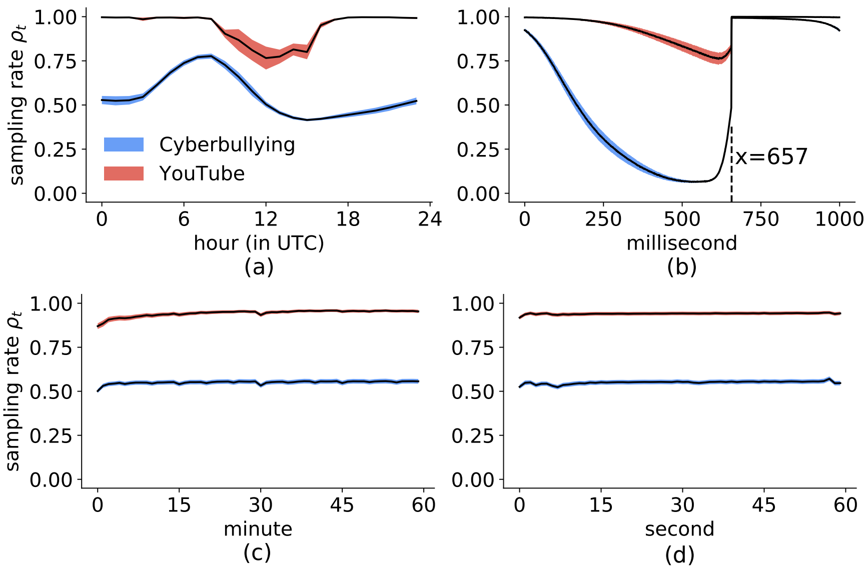


Figure 3.5: Sampling rates are uneven **(a)** in different hours or **(b)** in different milliseconds. black line: temporal mean sampling rates; color shades: 95% confidence interval. **(c)** Minutely sampling rates. **(d)** Secondly sampling rates.

3.3 Are tweets missing at random?

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

Tweet timestamps. Figure 3.5(a) plots the hourly sampling rates. Cyberbullying dataset has the highest sampling rate (ρ_t=78%) at UTC-8. The lowest sampling rate (ρ_t=41%) occurs at UTC-15, about half of the highest value. YouTube dataset is almost complete (ρ_t=100%) apart from UTC-8 to UTC-17. The lowest sampling rate is 76% at UTC-12. We posit that the hourly variation is related to the overall tweeting dynamics and the rate limit threshold (i.e., 50 tweets per second): higher tweet volumes yield lower sampling rates. Figure 3.5(b) shows the sampling rate at the millisecond level, which curiously exhibits a periodicity of one second. In Cyberbullying dataset, the sampling rate peaks at millisecond 657 (ρ_t=100%) and drops monotonically till millisecond 550 (ρ_t=6%) before bouncing back. YouTube dataset follows a similar trend with the lowest value (ρ_t=76%) at millisecond 615. The temporal variations are much less prominent at the minutely and secondly levels, as

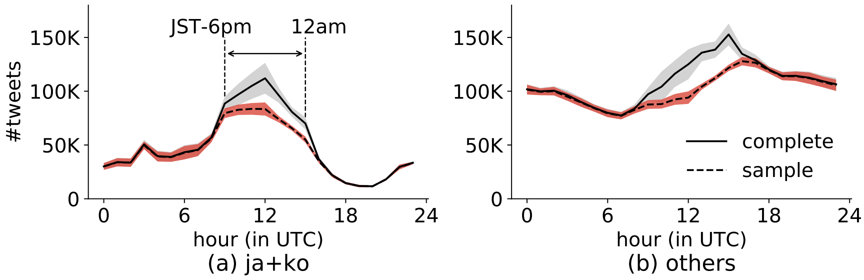


Figure 3.6: Hourly tweet volumes in YouTube dataset. **(a)** Japanese+Korean; **(b)** other languages. black line: temporal mean tweet volumes; color shades: 95% confidence interval.

shown in Figure 3.5 (c,d).

This artifact leaves the sample set vulnerable to automation tools. Users can deliberately schedule tweet posting time within the high sampling rate period for inflating their representativeness, or within the low sampling rate period for masking their content in the public API. In Section 3.4.3, we identify a set of users who may already exploit the sampling artifact.

Tweet languages. Some languages are mostly used within one particular timezone, e.g., Japanese and Korean[6]. The temporal tweet volumes for these languages are related to the daily activities in the corresponding countries. We break down the hourly tweet volumes of YouTube dataset into Japanese+Korean and other languages. The results are shown in Figure 3.6. Altogether, Japanese and Korean account for 31.4% tweets mentioning YouTube URLs. The temporal variations are visually different – 48.3% of Japanese and Korean tweets are posted in the evening of local time (JST-6pm to 12am), while tweets in other languages disperse more evenly. Because of the high volume of tweets in this period, sampling rates within UTC-9 to UTC-15 are lower (see Figure 3.5a). Consequently, "ja+ko" tweets are less likely to be observed (89.0% in average, 80.9% between JST-6pm and 12am) than others (92.9% in average).

Tweet types. Twitter allows the creation of 4 types of tweets. The users create a *root tweet* when they post new content from their home timelines. The other 3 types are interactions with existing tweets: *retweets* (when users click on the "Retweet" button); *quotes* (when users click on the "Retweet with comment" button); *replies* (when users click on the "Reply" button). The relative ratios of different types of tweets are distinct for the two datasets (see Table 3.4). Cyberbullying has higher ratios of retweets, quotes, and replies than YouTube, implying more interactions among users. However, the ratios of different types are very similar in the sampled versions of both datasets (max deviation=0.41%, retweets in YouTube dataset). We conclude that Twitter data sampling is not biased towards any tweet type.

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

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

Table 3.4: The sampling ratios of the 4 tweet types (root tweet, retweet, quote, and reply).

3.4 Impacts on Twitter entities

In this section, we study how the data sampling affects the observed frequency and relative ranking of Twitter entities, e.g., users, hashtags, and URLs. We first use a Bernoulli process to model the Twitter data sampling (Section 3.4.1). Next, we show how the entity statistics for one set (e.g., the complete) can be estimated using the other set (the sample, Section 3.4.2). Finally, we measure the distortions introduced in entity ranking by sampling and how to correct them (Section 3.4.3). The analyses in this section, Section 3.5, and Section 3.6, are done with Cyberbullying dataset since its sampling effects are more prominent.

3.4.1 Twitter sampling as a Bernoulli process

We examine how well we can use a Bernoulli process to approximate the Twitter sampling process. Assuming that tweets are sampled identically and independently, the Twitter sampling can be be seen as a simple Bernoulli process with the mean sampling rate $\bar{\rho}$. We empirically validate this assumption by plotting the complementary cumulative density functions (CCDFs) of user posting frequency (the number of times a user posts) and hashtag frequency (the number of times a hashtag appears) in Figure 3.7. The black and blue solid lines respectively show the CCDFs of the complete and the sample sets, while the black dashed line shows the CCDF in a synthetic dataset constructed from the complete set using a Bernoulli process with rate $\bar{\rho}$=52.72%. Firstly, we observe that the CCDF of the sample set is shifted left, towards the lower frequency end. Visually, the distributions for the synthetic (black dashed line) and for the observed sample set (blue solid line) overlap each other. Furthermore, following the practices in [Leskovec and Faloutsos, 2006], we measure the agreement between these distributions with Kolmogorov-Smirnov D-statistic, which is defined as

$$D(G, G') = \max_x\{|G(x) - G'(x)|\} \quad (3.1)$$

where G and G' are the cumulative distribution functions (CDFs) of two distribu-

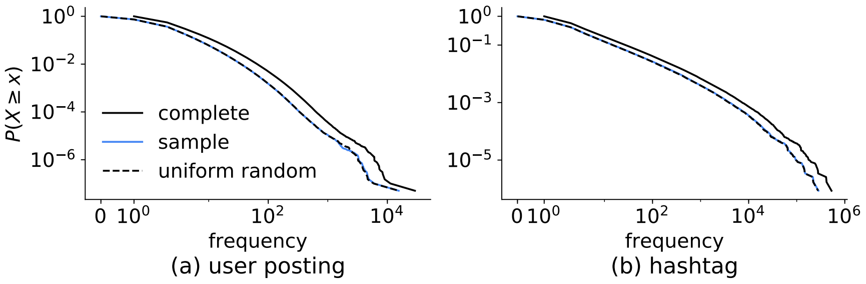


Figure 3.7: The frequency distributions of **(a)** user posting and **(b)** hashtag. The x-axis starts at 0 rather than 1, as the sample set and uniform random sample both have missing entities.

tions. With a value between 0 and 1, a smaller D-statistic implies more agreement between two measured distributions. The results show high agreement between entity distributions in the synthetic and the observed sample sets (0.0006 for user posting and 0.002 for hashtag). This suggests that despite the empirical sampling rates not being unique over time, a Bernoulli process of constant rate can model the observed entity frequency distribution well[7].

3.4.2 Entity frequency

We investigate whether the statistics on one set (complete or sample) can be estimated using only the statistics of the other set and the Bernoulli process model. We use n_c to denote the frequency in the complete set, and n_s the frequency in the sample set ($n_c \geq n_s$). More precisely, we ask these three questions: What is the distribution of n_s given $n_c = k$? What is the distribution of n_c given $n_s = k$? How many entities are missing altogether given the distribution of n_s?

Modeling sample frequency from the complete set. For a user who posts n_c times in the complete set, their sample frequency under the Bernoulli process follows a binomial distribution $B(n_c, \bar{\rho})$. Specifically, the probability of observing the user n_s times in the sample set is

$$\Pr(n_s | n_c, \bar{\rho}) = \binom{n_c}{n_s} \bar{\rho}^{n_s} (1-\bar{\rho})^{n_c - n_s} \tag{3.2}$$

We compute the empirical distribution and binomial distribution for n_c from 1 to 100. This covers more than 99% users in our dataset. Figure 3.8(a) shows the D-statistic between two distributions as a function of complete frequency n_c. The binomial distribution models the empirical data better when n_c is smaller. Figure 3.8(b)

[7] We do not choose the goodness of fit test (e.g., Kolmogorov-Smirnov test) because our sample sizes are in the order of millions. And trivial effects can be found to be significant with very large sample sizes. Instead we report the effect sizes (e.g., D-statistic). Alternative distance metrics (e.g., Bhattacharyya distance or Hellinger distance) yield qualitatively similar results.

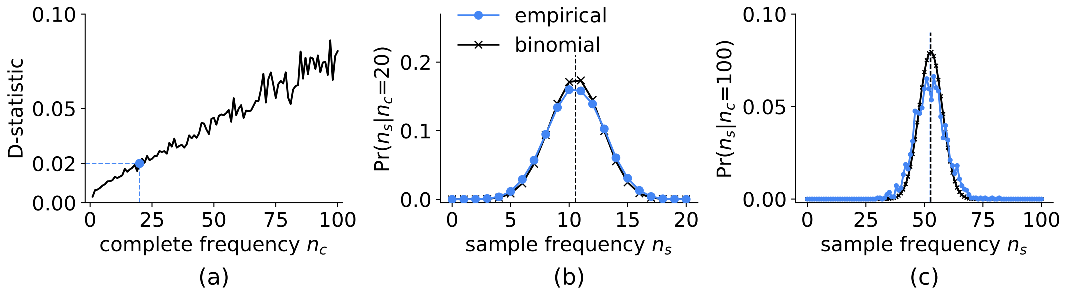


Figure 3.8: (a) D-statistic between empirical distribution and binomial distribution. **(b)** The probability distribution of observing n_s tweets in the sample set when a user posts 20 tweets in the complete set (mean value: 10.54). **(c)** The probability distribution of n_s when n_c=100.

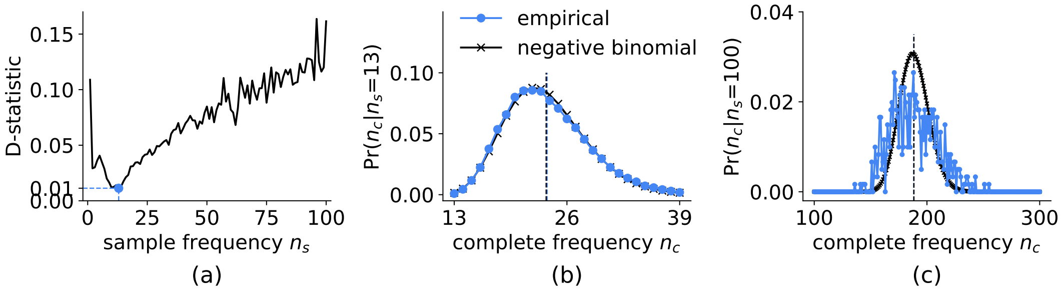
illustrates an example when the binomial distribution closely approximates the empirical distribution (n_c=20). Their mean sample frequencies (dashed vertical lines) are the same up to two decimal places (10.54). Figure 3.8(c) shows an example of n_c=100. Although the D-statistic is relatively large, binomial distribution still captures the general trend of the empirical distribution.

Inferring complete frequency from the sample set. Under the Bernoulli process, for users who are observed n_s times in the sample set, their complete frequencies follows a negative binomial distribution $NB(n_s, \bar{\rho})$. The negative binomial distribution models the discrete probability distribution of the number of Bernoulli trials before a predefined number of successes occurs. In our context, given n_s tweets ($n_s \geq 1$) are successfully sampled, the probability of having n_c tweets in the complete set is

$$\Pr(n_c|n_s, \bar{\rho}) = \binom{n_c-1}{n_s-1}\bar{\rho}^{n_s}(1-\bar{\rho})^{n_c-n_s} \tag{3.3}$$

We compute the empirical distribution and negative binomial distribution for n_s from 1 to 100. Figure 3.9(a) shows the D-statistic as a function of sample frequency n_s. Negative binomial distributions models the best when the number of observed tweets is between 9 and 15 (D-statistic<0.02). Figure 3.9(b) shows both distributions for n_s=13, where the minimal D-statistic is reached. The negative binomial distribution closely resembles the empirical distribution. Their estimated mean complete frequencies are very similar (23.60 vs. 23.72, shown as dashed vertical lines). Figure 3.9(c) shows an example of n_s=100. The approximation is visually more noisy. This implies that for users who tweet excessively, their user sampling rates may deviate from the mean sampling rate $\bar{\rho}$=52.72%. This observation motivates us to investigate the entity rank changes for the most active users post sampling (Section 3.4.3).

Estimating missing volume from the sample set. In data collection pipelines, the obtained entities from the filtered stream are sometimes used as seeds for the second



(a) (b) (c)

Figure 3.9: (a) D-statistic between empirical distribution and negative binomial distribution. **(b)** The probability distribution of posting n_c tweets in the complete set when observing a user posts 13 tweets in the sample set (mean value: 23.60). **(c)** The probability distribution of n_c when n_s=100.

	complete	sample	%missing	est. %missing
#users	19,802,506	14,649,558	26.02%	26.12%
#hashtags	1,166,483	880,096	24.55%	24.31%
#URLs	467,941	283,729	39.37%	38.99%

Table 3.5: Empirical and estimated missing rates for entities in Cyberbullying dataset, average tweet missing rate $1-\bar{\rho}$=47.28%.

step crawling, such as constructing user timelines based on user ids [Wang et al., 2015], or querying YouTube statistics based on video URLs [Wu et al., 2018]. However, some entities may be completely missing due to Twitter sampling. We thus ask: can we estimate the total number of missing entities given the entity frequency distribution of the sample set?

Table 3.5 shows the missing rates for different entities. Compared to the average tweet missing rate ($1-\bar{\rho}$=47.28%), the empirical missing rates of entities become much lower (24.55% to 39.37%) is because entities occurring multiple times are less likely to be missed. For example, users who tweet 5 times have 95.9% chance of not being missing. The different missing rates across entities is because the frequency distributions are different. For example, 30.6% hashtags are tweeted at least 5 times while 6.4% URLs are tweeted at least 5 times. These rationales also explain why the entity sampling ratios are higher than the mean sampling rate in Table 3.6.

We formulate the problem of estimating missing entity volume as solving a matrix equation with constraints. We use the symbol $\mathbf{F}$ to denote the entity frequency vector. $\mathbf{F}[n_s]$ represents the number of entities that occurs n_s times in the sample set. We want to estimate the frequency vector $\hat{\mathbf{F}}$ of the complete set. For any n_s, its sample frequency $\mathbf{F}[n_s]$ satisfies

$$\mathbf{F}[n_s] = \sum_{k=n_s}^{\infty} \Pr(n_s|k,\bar{\rho}) * \hat{\mathbf{F}}[k] \tag{3.4}$$

Equation (3.4) can be written using either binomial distribution (Equation (3.2)) or negative binomial distribution (Equation (3.3)). We choose binomial distribution because it fits better on empirical data (comparing Figure 3.8a to Figure 3.9a). We constrain $\hat{\mathbf{F}}$ to be non-negative numbers and to decrease monotonically since the frequency distribution is usually heavy-tailed in practice (Figure 3.7). We use the frequency vector for $n_s{\in}[1,100]$. The above matrix equation can be solved as a constrained optimization task. The constraints force all the frequencies to be non-negative. For users who post n_c times in the complete set, the probability of their tweets completely missing is $\Pr(n_s{=}0;n_c,\bar{\rho}){=}(1{-}\bar{\rho})^{n_c}$. Altogether, the estimated missing volume is $\hat{\mathbf{F}}[n_s{=}0] = \sum_{n_c=1}^{\infty}(1{-}\bar{\rho})^{n_c}\hat{\mathbf{F}}[n_c]$ for the whole dataset.

We apply this method to the distributions of users, hashtags and URLs. We show the estimated missing volume in Table 3.5. The relative errors (MAPE) are smaller than 0.5% for all entities. This suggests that the volume of missing entities can be accurately estimated if the frequency distribution of the sample set is observed.

Summary. Although the empirical sample rates have clear temporal variations, we show that we can use the mean sampling rate to estimate some entity statistics, including the frequency distribution and the missing volume. This reduces the concerns on assuming the observed data stream is a result of uniform random sampling [Joseph et al., 2014; Morstatter et al., 2014; Pfeffer et al., 2018].

3.4.3 Entity ranking

Entity ranking is important for many social media studies. One of the most common strategies in data filtering is to keep entities that rank within the top x, e.g., most active users or most mentioned hashtags [Morstatter et al., 2013; González-Bailón et al., 2014]. We measure how the Twitter data sampling distorts entity ranking for the most active users, and whether the ground-truth ranking in the complete set can be inferred from the sample ranking. Note that in this subsection, we allow the sampling rates to be time-dependent ρ_t and user-dependent ρ_u – as the sampling with a constant rate would preserve the ranking between the complete and the sample sets.

Detecting rank distortion. Figure 3.10(a) plots the most active 100 users in the sample set on the x-axis, and their ranks in the complete set on the y-axis. Each circle is colored based on the corresponding user sampling rate ρ_u. The diagonal line indicates uniform random sampling, in which the two sets of ranks should be preserved.

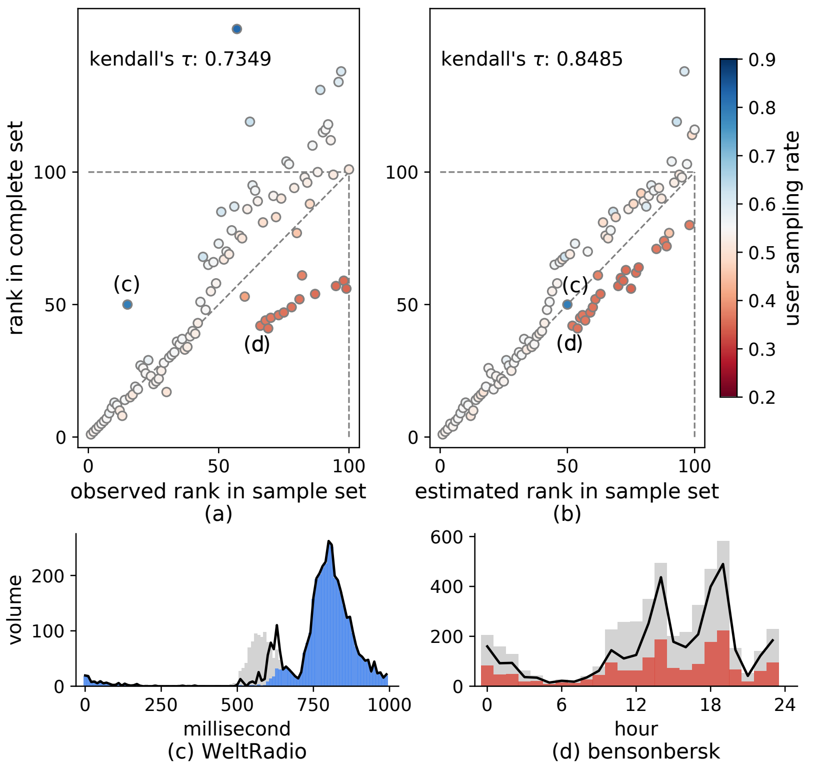


Figure 3.10: (a) Observed ranks in the sample set (x-axis) vs. true ranks in the complete set (y-axis). **(b)** Estimated ranks improve the agreement with the ground-truth ranks. **(c)** user *WeltRadio*, observed/true/estimated ranks: 15/50/50. **(d)** user *bensonbersk*, observed/true/estimated ranks: 66/42/52. blue/red shades: sample tweet volume; grey shades: complete tweet volume; black line: estimated tweet volume.

The users above the diagonal line improve their ranks in the sample set, while the ones below lose their positions.

Figure 3.10(c) highlights a user *WeltRadio*, who benefits the most from the sampling: it ranks 50th in the complete set, but it is boosted to 15th place in the sample set. Comparing the complete tweet volume, its volume (4,529) is only 67% relative to the user who actually ranks 15th in the complete set (6,728, user *thirdbrainfx*). We also find that *WeltRadio* tweets mostly in the very high sampling rate secondly period (millisecond 657 to 1,000), resulting in a high user sampling rate (ρ_u=79.1%). On the contrary, Figure 3.10(d) shows a user *bensonbersk* with decreased rank in the sample set and low sampling rate (ρ_u=36.5%). Examining his posting pattern, this user mainly tweets in the low sampling rate hours (UTC-12 to 19).

Estimating true ranking from the sample set. Apart from measuring the rank distortion between the complete and the sample sets, we investigate the possibility of estimating the ground-truth ranks by using the observations from the sample set. From

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

Table 3.6: Statistics of user-hashtag bipartite graph in Cyberbullying dataset. The sampling ratios (rightmost column) compare the values of the sample set against that of the complete set. The reason that ratios are higher than mean sampling rate ($\bar{\rho}$=52.72%) is due to entities that occur multiple times are more likely to be sampled.

the rate limit messages, we extract the temporal sampling rates that are associated with different timescales (hour, minute, second, and millisecond), i.e., $\rho_t(h, m, s, ms)$. Based on the negative binomial distribution, for a user who we observe n_s times at timestamp $\kappa{=}(h, m, s, ms)$, the expected volume is $n_s/\rho_t(\kappa)$. We compute the estimated tweet volumes for all users and select the most active 100 users. Figure 3.10(b) shows the estimated ranks on the x-axis and the true ranks on the y-axis. We quantify the degree of agreement using Kendall's τ, which computes the difference of concordant and discordant pairs between two ranked lists. With value between 0 and 1, a larger τ implies more agreement. The Kendall's τ is improved from 0.7349 to 0.8485 with our estimated ranks. The rank correction is important since it allows researchers to mitigate the rank distortion without constructing a complete data stream.

3.5 Impacts on networks

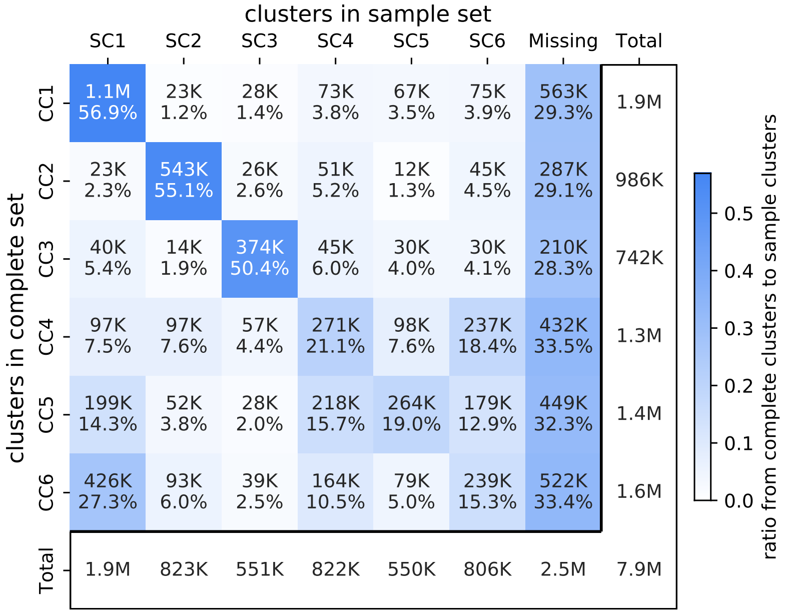
In this section, we measure the effects of data sampling on two commonly studied networks on Twitter: the user-hashtag bipartite graph, and the user-user retweet network.

3.5.1 User-hashtag bipartite graph

The bipartite graph maps the affiliation between two disjoint sets of entities. No two entities within the same set are linked. Bipartite graphs have been used in many social applications, e.g., mining the relation between scholars and published papers [Newman, 2001], or between artists and concert venues [Arakelyan et al., 2018]. Here we construct the user-hashtag bipartite graphs for both the complete and the sample sets. This graph links users to their used hashtags. Each edge has a weight – the number of tweets between its associated user and hashtag. The basic statistics for the bipartite graphs are summarized in Table 3.6.

Clustering techniques are often used to detect communities in such bipartite graphs. We apply spectral clustering [Stella and Shi, 2003] on the user-hashtag bipartite graph, with the number of clusters set at 6. The resulted clusters are summarized in Table 3.7, together with the most used 5 hashtags and a manually-assigned category. Apart from the cyberbullying keywords, there are significant amount of hashtags related to politics, live streaming, and Korean pop culture, which are considered as some of the most discussed topics on Twitter.

We further quantify how the clusters traverse from the complete set to the sample set in Figure 3.11. Three complete-set clusters (CC1, CC2, and CC3) are maintained in the sample-set clusters (respectively mapping to SC1, SC2, and SC3), since more than half of the entities preserve. The remaining three complete-set clusters disperse. Investigating the statistics for the complete-set clusters, the preserved ones have a larger average weighted degree, meaning more tweets between the users and hashtags in these clusters. Another notable observation is that albeit the entities (both users and hashtags) move to the sample-set clusters differently, all complete-set clusters have similar missing rates (28% to 34%, comparing the size of SCs to the size of CCs). It suggests that Twitter data sampling impacts the community structure. Denser structures are more resilient to sampling.



	SC1	SC2	SC3	SC4	SC5	SC6	Missing	Total
CC1	1.1M 56.9%	23K 1.2%	28K 1.4%	73K 3.8%	67K 3.5%	75K 3.9%	563K 29.3%	1.9M
CC2	23K 2.3%	543K 55.1%	26K 2.6%	51K 5.2%	12K 1.3%	45K 4.5%	287K 29.1%	986K
CC3	40K 5.4%	14K 1.9%	374K 50.4%	45K 6.0%	30K 4.0%	30K 4.1%	210K 28.3%	742K
CC4	97K 7.5%	97K 7.6%	57K 4.4%	271K 21.1%	98K 7.6%	237K 18.4%	432K 33.5%	1.3M
CC5	199K 14.3%	52K 3.8%	28K 2.0%	218K 15.7%	264K 19.0%	179K 12.9%	449K 32.3%	1.4M
CC6	426K 27.3%	93K 6.0%	39K 2.5%	164K 10.5%	79K 5.0%	239K 15.3%	522K 33.4%	1.6M
Total	1.9M	823K	551K	822K	550K	806K	2.5M	7.9M

Figure 3.11: The change of clusters from complete set to sample set. Each cell denotes the volume (top number) and the ratio (bottom percentage) of entities (users and hashtags) that traverse from a complete cluster to a sample cluster. Clusters are ordered to achieve maximal ratios along the diagonal.

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

	hashtags	**ps4live**
		10tv
		brexit
		afd
		demdebate
			bts
			mamavote
			blackpink
			pcas
			bts(ko)
				gay
				pussy
				sex
				horny
				porn
					mixch.tv(ja)
					bigil
					peckpalitchoke(th)
					reality_about_islam(hi)
					doki.live(ja)
						bigolive
						kamleshtiwari
						bb13
						biggboss13
						execution_rajeh_mahmoud(ar)
							Idolish7(ja)
							reunion
							Idolish7(ja)
							vixx
							vixx(ko)

Table 3.7: Statistics and the most used 5 hashtags in the 6 clusters of the user-hashtag bipartite graph. Three complete clusters maintain their structure in the sample set (**boldfaced**). The language code within brackets is the original language for the hashtag. ja: Japanese; ko: Korean; th: Thai; hi: Hindi; ar: Arabic.

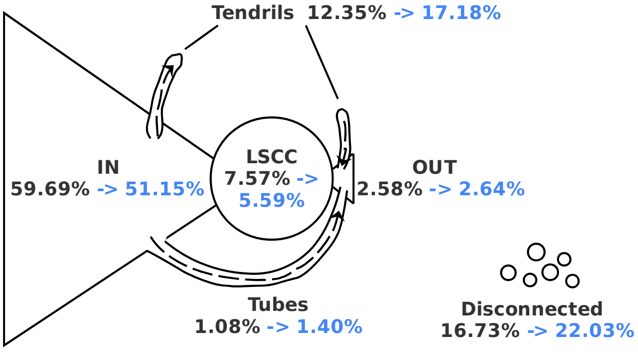


Figure 3.12: Visualization of bow-tie structure in complete set. The black number indicates the relative size of component in the complete set, blue number indicates the relative size in the sample set.

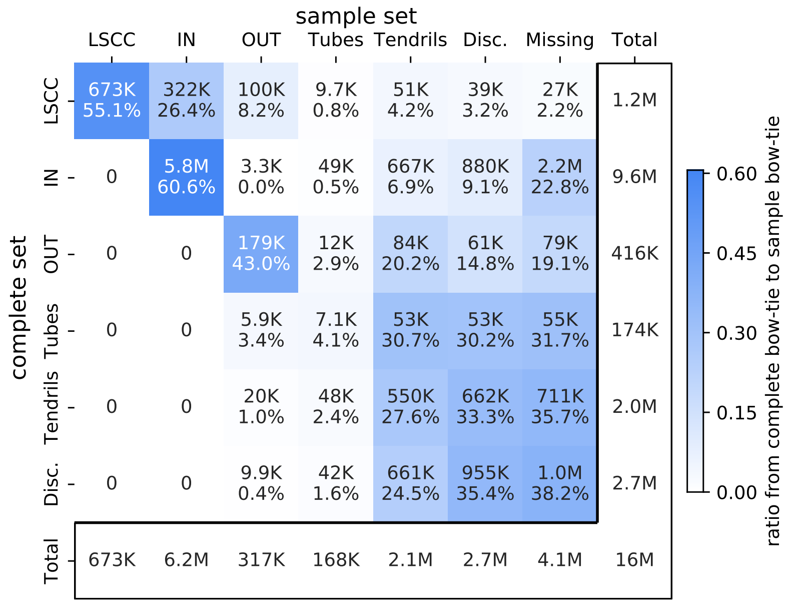
3.5.2 User-user retweet network

Retweet network describes the information sharing between users. We build a user-user retweet network by following the "@RT" relation. Each node is a user, and each edge is a directed link weighted by the number of retweets between two users. The user-user retweet network has been extensively investigated in literature [Sadikov et al., 2011; Morstatter et al., 2013; González-Bailón et al., 2014].

We choose to characterize the retweet network using the bow-tie structure. Initially proposed to measure the World Wide Web [Broder et al., 2000], the bow-tie structure was also used to measure the QA community [Zhang et al., 2007] or YouTube video networks [Wu et al., 2019]. The bow-tie structure characterizes a network into 6 components: (a) the largest strongly connected component (LSCC) as the central part; (b) the IN component contains nodes pointing to LSCC but not reachable from LSCC; (c) the OUT component contains nodes that can be reached by LSCC but not pointing back to LSCC; (d) the Tubes component connects the IN and OUT components; (e) the Tendrils component contains nodes pointing from In component or pointing to OUT component; (f) the Disconnected component includes nodes not in the above 5 components.

Figure 3.12 visualizes the bow-tie structure of the user-user retweet network, alongside with the relative size for each component in the complete and sample sets. The LSCC and IN components, which make up the majority part of the bow-tie, reduce the most in both absolute size and relative ratio due to sampling. OUT and Tubes are relatively small in both complete and sample sets. Tendrils and disconnected components enlarge 39% and 32% after sampling.

Figure 3.13 shows the node flow of each components from the complete set to the



complete set \ sample set	LSCC	IN	OUT	Tubes	Tendrils	Disc.	Missing	Total
LSCC	673K 55.1%	322K 26.4%	100K 8.2%	9.7K 0.8%	51K 4.2%	39K 3.2%	27K 2.2%	1.2M
IN	0	5.8M 60.6%	3.3K 0.0%	49K 0.5%	667K 6.9%	880K 9.1%	2.2M 22.8%	9.6M
OUT	0	0	179K 43.0%	12K 2.9%	84K 20.2%	61K 14.8%	79K 19.1%	416K
Tubes	0	0	5.9K 3.4%	7.1K 4.1%	53K 30.7%	53K 30.2%	55K 31.7%	174K
Tendrils	0	0	20K 1.0%	48K 2.4%	550K 27.6%	662K 33.3%	711K 35.7%	2.0M
Disc.	0	0	9.9K 0.4%	42K 1.6%	661K 24.5%	955K 35.4%	1.0M 38.2%	2.7M
Total	673K	6.2M	317K	168K	2.1M	2.7M	4.1M	16M

Figure 3.13: The change of bow-tie components from complete set to sample set. Each cell denotes the volume (top) and the ratio (bottom) of users that traverse from a component in complete set to a component in sample set.

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

3.6 Impacts on retweet cascades

Information diffusion is perhaps the most studied social phenomenon on Twitter. A retweet cascade consists of two parts: a root tweet and its subsequent retweets. A number of models have been proposed for modeling and predicting retweet cascades [Zhao et al., 2015; Mishra et al., 2016; Martin et al., 2016]. However, these usually make the assumption of observing all the retweets in cascades.

In this section, we analyze the impacts of Twitter sampling on retweet cascades and identify risks for existing models. We first construct cascades without missing tweets from the complete set. Next, we measure the sampling effects for some commonly used features in modeling retweet cascades, e.g., inter-arrival time and potential reach.

Constructing complete cascades. When using the filtered streaming API, if a root tweet is observed, the API should return all its retweets. This is because the API

	complete	sample	ratio
#cascades	3,008,572	1,168,896	38.9%
#cascades (≥50 retweets)	99,952	29,577	29.6%
average retweets per cascade	15.6	11.0	70.2%
median inter-arrival time (s)	22.9	105.7	461.6%

Table 3.8: Statistics of cascades in Cyberbullying dataset, mean sampling rate $\bar{\rho}$=52.72%. The numbers of sampled cascades (top 2 rows) are below 52.72% (see text for explanation).

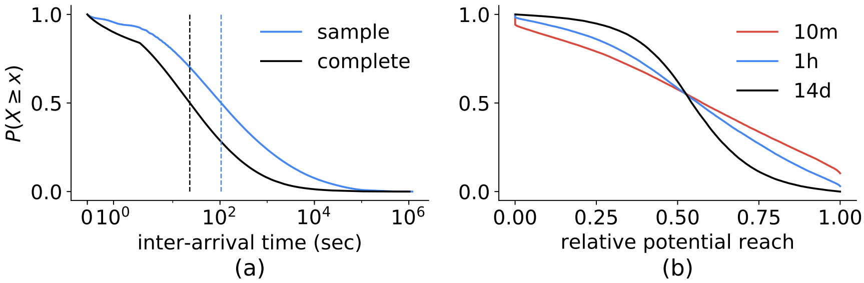


Figure 3.14: CCDFs of **(a)** inter-arrival time and **(b)** relative potential reach.

also tracks the keywords in the retweeted_status field of a tweet (i.e., the root tweet), which allows us to construct a set of complete cascades from the complete set. In the sample set, both the root tweet and any of its retweets could be missing. If the root tweet is missing, we miss the entire cascade. If some retweets are missing, we observe a partial cascade.

Table 3.8 lists the obtained cascades in the complete and the sample sets. Notably, there are 3M cascades in the complete set, but only 1.17M in the sample set (38.85%), out of which only 508k (16.88%) cascades are complete and their sizes are relatively small (max cascade size: 23, mean size: 1.37). Prior literature [Zhao et al., 2015] often concentrates on retweet cascades with more than 50 retweets. There are 99,952 such cascades in the complete set, but only 29,577 (29.6%) in the sample set, out of which none is complete. Assuming the root tweets are sampled independently with mean sampling rate 52.72%, theoretically at most 52.72% cascades can be observed because we require the root tweet must be sampled. With additional missing retweets, the sampling ratio for the cascades will be lower than 52.72%.

Inter-arrival time. One line of work models the information diffusion as point processes [Zhao et al., 2015; Mishra et al., 2016]. These models use a memory kernel as a function of the time gap Δt between two consecutive events, which is also known as inter-arrival time. Figure 3.14(a) plots the CCDFs of inter-arrival times in the complete and the sample sets. The distribution shifts right, towards larger values. This is expected as the missing tweets increase the time gap between two observed tweets.

The median inter-arrival time is 22.9 seconds in the complete set (black dashed line), meaning 50% retweets happen within 23 seconds from last retweet. After sampling, the median increases almost 5-fold to 105.7 seconds (blue dashed line). For research that uses tweet inter-arrival time, this presents the risk of miss-calibrating models and of underestimating the virality of the cascades.

Potential reach. Online influence is another well-studied phenomenon on Twitter, and one of its proxies is the number of followers of a user. We define potential reach as the total number of all observed retweeters' followers. This approximates the size of the potential audience for the root tweet. We compute the relative potential reach as the ratio of potential reach in the sample cascade against that in the complete cascade, and we plot the CCDFs in Figure 3.14(b). When observing cascades for as much as 14 days, 50% of the cascades have the relative potential reach below 0.544. This indicates that when using the sampled Twitter data, researchers can severely underestimate the size of the potential audience.

Another common setting is to perform early prediction, i.e., after observing 10 minutes or 1 hour of each retweet cascade. Figure 3.14(b) shows that the relative potential reach is more evenly distribution for shorter time windows – 21.0% cascades have relative potential reach below 0.25 and 33.7% cascades above 0.75 within 10 minutes span – comparing to the observation over 14 days (5.1% and 11.3%, respectively). Visually, the CCDF of longer horizons is curved and the area around mean sampling rate is the most dense. This is because the sampling rates across cascades stabilize around mean sampling rate and have smaller variance when the observation windows become longer.

3.7 Conclusion

This chapter presents a set of in-depth measurements on the effects of Twitter data sampling. We validate that Twitter rate limit messages closely approximate the volume of missing tweets. Across different timescales (hour, minute, second, millisecond), we find that the sampling rates have distinct temporal variations at each scale. We show the effects of sampling across different subjects (entities, networks, cascades), which may in turn distort the results and interpretations of measurement and modeling studies. For counting statistics such as number of tweets per user and per hashtag, we find that the Bernoulli process with a uniform rate is a reasonable approximation for Twitter data sampling. We also show how to estimate ground-truth statistics in the complete data by using only the sample data.

3.7.1 Limitations

These observations in this chapter apply to current Twitter APIs (as of May, 2020) and are subject to the changes of Twitter's proprietary sampling mechanisms. We are aware of that Twitter plans to release a new set of APIs in near future. Consistent with the current streaming APIs, the rate limit threshold for the new APIs is also set to 50 tweets per second [twitter.com, 2020b]. Therefore, we believe the observations of this work will hold.

3.7.2 Practical implications and future work

This work calls attention to the hidden noises and biases in social media data. We have shown effective methods for estimating ground-truth statistics, which allows researchers to mitigate the risks in their datasets without collecting the complete data. Our research provides methods and toolkits for collecting sampled and complete data streams on Twitter. Our findings also provide foundations to many other research topics using sampled data, such as community detection and information diffusion algorithms that are robust to data subsampling.

Future works include measuring a larger set of activity and network measurements under data sampling, generalizing the results of this work to other social media platforms and data formats, and quantifying the robustness of existing network and diffusion models against data sampling. To establish a benchmark, we release our collected complete and sampled retweet cascades that contain all required features (timestamp and user follower count) of existing diffusion models [Zhao et al., 2015; Mishra et al., 2016].

Chapter 4

Measuring and predicting engagement in online videos

In the previous chapter, we extensively discussed the impacts of Twitter data sampling. Since we use Twitter data as a proxy to curate YouTube video datasets, we can now use the methods in Chapter 3 to measure the sampling effects in our datasets.

In this chapter, we discuss engagement of online videos in detail. Most current research focuses on modeling video viewership, but we argue that video engagement, or time spent watching is a more appropriate measure for resource allocation problems in attention, network, and promotion activities.

We present the first large-scale measurement of video-level aggregate engagement on a collection of 5.3 million YouTube videos published over two months in 2016. We study a set of metrics including time and percentage of a video being watched. In Section 4.3, we propose a new metric, relative engagement, which is calibrated against video properties and strongly correlated with recognized notions of quality. Moreover, we find engagement measures of videos stable over time. In Section 4.4, we find that aggregate engagement metrics are predictable from a cold-start setup, having most of their variance explained by video context, topics and channel information. Channel past success is the most predictive feature while video topics seem to have a non-trivial effect. In Section 4.5, we link daily watch time to external sharing of a video using a self-exciting Hawkes Intensity Process, and find that we can forecast daily watch time more accurately than daily views.

This chapter provides a set of new yardsticks for measuring online content including video and other length-constrained media such as songs and podcasts. The observations here imply several prospective usages of engagement metrics – choosing engaging topics for video production, or promoting engaging videos in recommender systems.

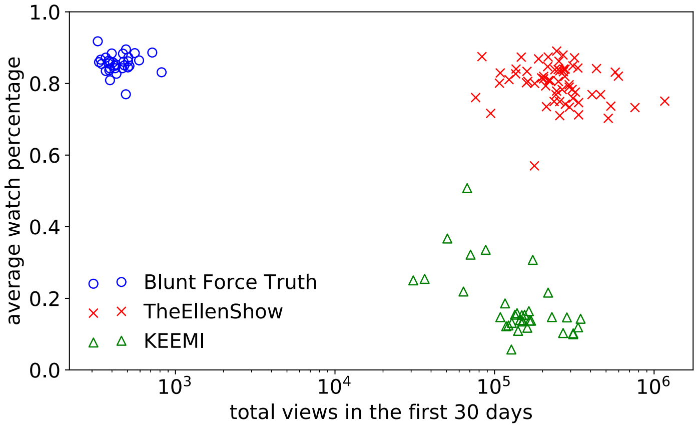


Figure 4.1: Scatter plot of videos from three YouTube channels: Blunt Force Truth (political entertainment, blue) circle), KEEMI (cooking vlog, green) triangle), and TheEllenShow (comedy, red) cross). x-axis: total views in the first 30 days; y-axis: average watch percentage.

4.1 Introduction

Attention is a scarce resource in the modern world. There are many metrics for measuring attention received by online content, such as page views for webpages, listen counts for songs, view counts for videos, and the number of impressions for advertisements. Although these metrics describe the human behavior of *choosing* one particular item, they do not describe how users *engage* with this item [Van Hentenryck et al., 2016]. For instance, an audience may become immersed in the interaction or quickly abandon it – the distinction of which will be clear if we know how much time the user spent interacting with this given item. Hence, we consider popularity and engagement as different measures of online behavior.

This chapter studies online videos using publicly available data from the largest video hosting site YouTube. On YouTube, popularity is characterized as the willingness to click a video, whereas engagement is the watch pattern after clicking. While most research has focused on modeling popularity [Pinto et al., 2013; Rizoiu et al., 2017b], engagement of online videos is not well understood, leading to key questions such as: How to measure video engagement? Does engagement relate to popularity? Can engagement be predicted? Once understood, engagement metrics will become relevant targets for recommender systems to rank the most valuable videos.

In Figure 4.1, we plot the number of views against the average percentage watched for 128 videos in 3 channels. While the entertainment channel Blunt Force Truth has the least views on average, the audience tend to watch more than 80% of each video. On the contrary, videos from the cooking vlogger KEEMI have on average 159,508 views, but they are watched only 18%. This example illustrates that videos with a high number of views do not necessarily have high watch percentages, and prompts us to investigate other metrics for describing engagement.

Recent progress in understanding video popularity and the availability of new datasets allow us to address four open questions about video engagement. Firstly, **on an aggregate level, how to measure engagement?** Most engagement literature focuses on the viewpoint of an individual user, such as recommending relevant products [Covington et al., 2016], tracking mouse gestures [Arapakis et al., 2014] or optimizing search results [Drutsa et al., 2015]. Since user-level data is often unavailable, defining and measuring average engagement is useful for content producers on YouTube. Secondly, **within the scope of online video, can engagement help measure content quality?** As shown in Figure 4.1, video popularity metric is inadequate to estimate quality. One early attempt to measure online content quality was taken by Salganik et al. [2006], who studied music listening behavior in an experimental environment. For a large number of online contents, measuring quality from empirical data still remains unexplored. Thirdly, **in a cold-start setup, can engagement be predicted?** For engagement, Park et al. [2016] showed the predictive power of collective user reactions. However, these features require monitoring the system for a period of time. In contrast, if engagement can be predicted before content is uploaded, it will provide actionable insights to content producers. Lastly, **on forecasting future performance, how predictable is engagement comparing to popularity?** Online popularity is known to be difficult to forecast [Martin et al., 2016; Hofman et al., 2017]. Yet, no one has tried to forecast the future trend of watch time, nor compared the difference of predictability between engagement and popularity.

We address the first question by constructing a new dataset that contains more than 5 million tweeted YouTube videos and 3 datasets that contain quality videos. We build two 2-dimensional maps that visualize the internal bias of existing engagement metrics – average watch time and average watch percentage – against video length. Building upon that, we derive a novel metric relative engagement, as the duration-calibrated rank of average watch percentage.

Addressing the second question, we demonstrate that relative engagement is stable over time, and strongly correlates with established quality measures in Music and News categories, such as Billboard songs, Vevo artists, and top news channels. It implies that relative engagement can be a target for recommender systems to prioritize quality videos, and for content producers to create engaging videos.

Addressing the third question, we predict engagement metrics in a cold-start setting, using only video topical content and channel features. With off-the-shelf machine learning algorithms, we achieve R^2=0.77 for predicting average watch percentage. We consider this as a significant result that shows the predictability of engagement metrics. Furthermore, we explore the predictive power of video topics and find some topics are strong indicators for engagement.

Addressing the last question, we adopt a variant of stochastic point processes, namely Hawkes Intensity Process (HIP) [Rizoiu et al., 2017b]. We model and forecast both daily watch time and daily views. The forecast error on engagement metric is 4.93 percentile points, better than that of popularity metric (5.43 percentile points).

The main contributions of this chapter include:

- We conduct a large-scale measurement study of engagement on 5.3 million videos over two-month period, and publicly release 4 new datasets and the engagement benchmarks[1].
- We measure a set of engagement metrics for online videos, including average watch time, average watch percentage, and a novel metric – relative engagement, which is calibrated with respect to video length, stable over time, and correlated with video quality.
- We predict relative engagement and watch percentage from video context, topics, and channel reputation in a cold-start setting (i.e., before the video gathers any view or comment), achieving R^2=0.45 and 0.77 respectively.
- We explain and forecast daily watch time. The self-exciting HIP model achieves an average error of 4.93 percentile points.
- We release a software package "YouTube-insight" for collecting metadata and historical data for videos on YouTube.[2].

4.2 Data

We curate the YouTube Engagement '16 datasets that consist of 4 new publicly available video datasets, as summarized in Table 4.1 and Table 4.2. We also discuss the Twitter sampling effects on our datasets. We conclude with an introduction of three daily series available for all videos: shares, views and watch time.

4.2.1 YouTube Engagement '16 datasets

Tweeted videos dataset contains 5,331,204 videos published between July 1st and August 31st, 2016 from 1,257,412 channels. The notion of *channel* on YouTube is

[1] The code and datasets are available at https://github.com/avalanchesiqi/youtube-engagement.

[2] The package is available at https://github.com/avalanchesiqi/youtube-insight. However, it is no longer working as designed since YouTube deprecated the insight endpoint around Nov, 2018.

Dataset	#videos	#channels	est. #videos	est. #channels
Tweeted Videos	5,331,204	1,257,412	5,976,111	1,362,140
Vevo Videos	67,649	8,685	N/A	N/A
Billboard Videos	63	47	N/A	N/A
Top News Videos	28,685	91	N/A	N/A

Table 4.1: Overview of 4 new video datasets. The estimated number of the complete videos and channels are provided due to Twitter data sampling, see computing methods described in Section 3.4.2. Video sampling rate is 89.2% for the Tweeted Videos dataset.

category	#videos	est. #videos	category	#videos	est. #videos
People	1,265,805	1,438,604	Comedy	138,068	153,563
Gaming	1,079,434	1,188,898	Science	110,635	122,448
Entertainment	775,941	867,469	Auto	84,796	96,916
News	459,728	518,365	Travel	65,155	73,739
Music	449,314	502,288	Activism	58,787	66,009
Sports	243,650	272,859	Pets	27,505	31,478
Film	194,891	219,467	Show	1,457	1,599
Howto	192,931	216,245	Movie	158	185
Education	182,849	205,857	Trailer	100	115

Table 4.2: Breakdown of Tweeted Videos by category.

analogous to that of *user* on other social platforms, since every video is published by a channel and belonging to one user account. Using Twitter mentions to sample a collection of YouTube videos has been used in previous works [Yu et al., 2014; Rizoiu et al., 2017b]. 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). This yields 244 million tweets over the two-month period. In each tweet, we search the extended_urls field and extract the associated YouTube video id. This results in 36 million unique video ids and over 206 million tweets. For each video, we extract its metadata and three attention-related dynamics, as described in Section 4.2.3. A non-trivial fraction (45.82%) of all videos have either been deleted or their statistics are not publicly available. This leaves a total of 19.5 million usable videos.

We further filter videos based on recency and the level of attention. We remove videos that are published prior to this two-month period to avoid older videos, since being tweeted a while after being uploaded may indicate higher engagement. We also filter out videos that receive less than 100 views within their first 30 days after upload, which is the same filter used by Brodersen et al. [2012]. Videos that do not appear on Twitter, or have extremely low number of early views are unlikely to accumulate a large amount of attention [Pinto et al., 2013; Rizoiu and Xie, 2017], therefore, they do not provide enough data to reflect collective watch patterns. Our proposed measures can still be computed on these removed videos, however the results might have limited relevance given the low level of user interaction with them. Table 4.2 shows a detailed category breakdown of TWEETED VIDEOS.

Quality videos datasets. We collect three datasets containing videos deemed of high quality by domain experts, two of which are on Music category and one is on News. These datasets are used to link engagement and video quality (Section 4.3.4).

- **Vevo Videos.** Vevo is a multinational video hosting service which syndicates licensed music clips from three major record companies on YouTube [wikipedia.com, 2020b]. VEVO artists usually come from professional music background, and their videos are professionally produced. We consider VEVO VIDEOS to be of higher quality than the average Music videos in the TWEETED VIDEOS dataset. We collect all the YouTube channels that contain the keyword "Vevo" in the title and a "verified" status badge on the profile webpage. In total, this dataset contains 8,685 Vevo channels with 67,649 music clips, as of August 31st, 2016.

- **Billboard Videos.** Billboard acts as a canonical ranking source in the music industry, aggregating 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 being of high quality. We collect 63 videos from 47 artists based on the 2016 Billboard Hot 100 chart[3].

- **Top News Videos** features a list of top 100 most viewed News channels, as reported by an external ranking source[4]. This list includes traditional news broadcasting companies (e.g., CNN), as well as popular politic talk shows (e.g., The Young Turks). For each channel, we retrieve its last 500 videos published before Aug 31st, 2016. This dataset contains 91 publicly available News channels and 28,685 videos.

4.2.2 Twitter sampling effects

We apply methods described in Chapter 3 to measure the sampling effects on the YOUTUBE ENGAGEMENT '16 datasets. The overall tweet sampling rate between 2016-07-01 and 2016-08-31 is 76.67% (230,290,042 collected tweets out of 300,347,371 estimated total tweets). Next, we infer the complete number of videos (channels) by feeding the sampled tweetcount distribution per video (channel) into Equation (3.4). The estimated volumes are shown in Table 4.1. Additionally, Table 4.2 lists the estimated volume for each category. The video sampling rates for different categories vary between 85.4% and 90.8%.

[3] https://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_2016

[4] https://vidstatsx.com/youtube-top-100-most-viewed-news-politics

For the three Quality videos datasets, there is no need to estimate the complete volume because they are not constructed from collected tweets. Hence, the Twitter sampling effects do not apply on them.

4.2.3 Video metadata and attention dynamics

For each video, we use the YouTube Data API to retrieve video metadata information – video id, title, description, upload time, category, duration, definition, channel id, channel title, and associated Freebase topic ids, which we resolve to entity names using the latest Freebase data dump[5].

We develop a software package "YouTube-insight" to extract three daily series of video attention dynamics: daily volume of shares, view counts and watch time. Throughout this chapter, we denote the number of shares and views that a video receives on the t^{th} day after upload as $s[t]$ and $x_v[t]$, respectively. Similarly, $x_w[t]$ is the total amount of time of video being watched on the t^{th} day. Each attention series is observed for at least 30 days, i.e., $t{=}1, 2, \ldots 30$. Most prior research on modeling video popularity dynamics [Szabo and Huberman, 2010; Figueiredo et al., 2016] studies only view counts. To the best of our knowledge, our work is the first to perform large-scale measurements on video watch time. The YouTube Engagement '16 datasets are also one of the two publicly available datasets including information of video watch time. The other is our Vevo Music Graph dataset (see Section 5.2.1).

4.3 Measures of video engagement

In this section, we measure the interplay between view count, watch time, watch percentage and video duration. We first examine their relation in a new visual presentation – *engagement map*, then we propose *relative engagement*, a novel metric to estimate video engagement (Section 4.3.3). We show that relative engagement calibrates watch patterns for videos of different lengths, demonstrates correlation to external notions of video quality (Section 4.3.4), and remains stable over time (Section 4.3.5).

4.3.1 Discrepancy between views and watch time

Figure 4.1 illustrates that watch patterns (e.g., average percentage of video watched) can be very different for videos with similar views. We examine the union set of top n videos in Tweeted Videos dataset, respectively ranked by total views and total watch time at the age of 30 days. For n varying from 100 to 1000, we measure their

[5] https://developers.google.com/freebase. The Freebase corpus is not updated any more but data dump is still available.

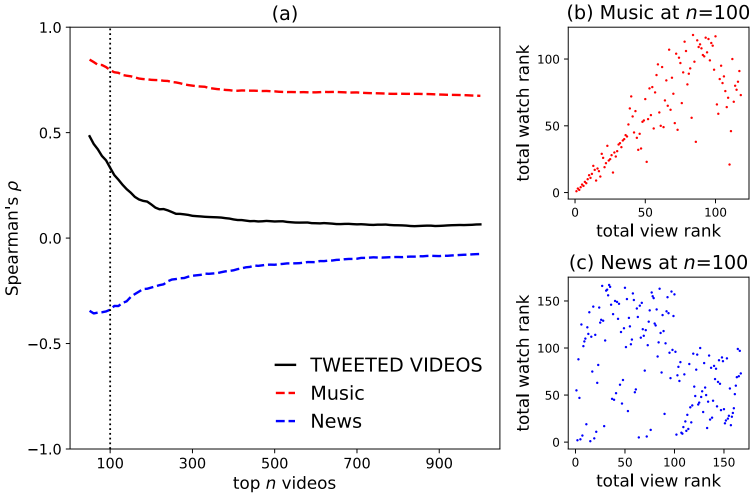


Figure 4.2: (a) Disagreement between the union set of top n most viewed and top n most watched videos in TWEETED VIDEOS at the age of 30 days, measured with Spearman's ρ. **(b-c)** Scatter plots of video ranking in view and in watch at n=100 in Music (ρ=0.80) and News (ρ= − 0.34).

agreement using Spearman's ρ. With value between -1 and +1, a positive ρ implies that as the rank in one variable increases, so does the rank in the other variable. A ρ of 0 indicates that no correlation exists in these two ranked variables. Figure 4.2(a) shows that in TWEETED VIDEOS, video ranks in total view count and total watch time correlate at the level of 0.48 when n is 50, but this correlation declines to 0.08 when n increases to 500 (solid black line). Furthermore, the level of agreement varies across different video categories: for Music, a video that ranks high in total view count often ranks high in total watch time ($\rho = 0.80$ at $n = 100$, Figure 4.2b); for News, the two metrics have a weak negative correlation ($\rho = -0.34$ at $n = 100$, Figure 4.2c). The difference results from the varied duration distributions across categories (shown in the upper panels of Figure 4.4) – Music videos are often from 3 to 5 minutes, while News videos can be as short as a few seconds (e.g., user-generated clips by phones), or as long as a few hours (e.g., live streaming reports by news outlets).

This observation suggests that total view count and total watch time provide different aspects of how audience interact with YouTube videos. One recommender system optimizing for view count may generate remarkably different results with one that drives watch time [Yi et al., 2014]. In the next section, we analyze their interplay to construct more diverse set of measures for video engagement.

4.3.2 New tool – engagement map

For a given video, we compute two aggregate metrics:

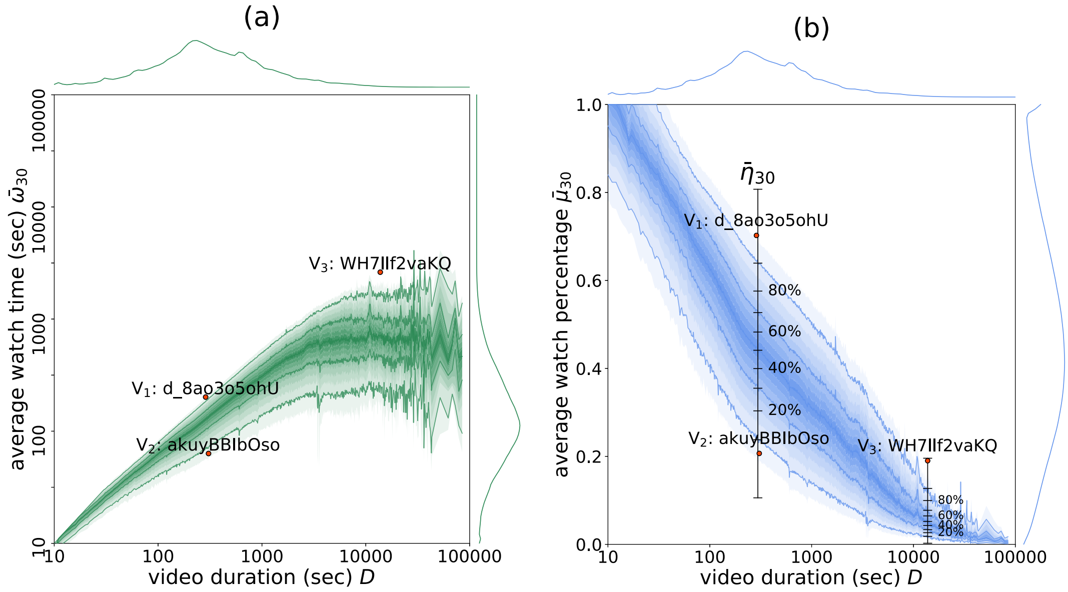


Figure 4.3: Video engagement in the Tweeted videos dataset at the age of 30 days. (a) video duration D vs. average watch time $\bar{\omega}_{30}$; (b) the engagement map: video duration D vs. average watch percentage $\bar{\eta}_{30}$.

- *average watch time* $\bar{\omega}_t$: the total watch time $x_w[1:t]$ divided by the total view count $x_v[1:t]$ up to day t

$$\bar{\omega}_t = \frac{\sum_{i=1}^{t} x_w[i]}{\sum_{i=1}^{t} x_v[i]} \tag{4.1}$$

- *average watch percentage* $\bar{\mu}_t$: the average watch time $\bar{\omega}_t$ normalized by video duration D

$$\bar{\mu}_t = \frac{\bar{\omega}_t}{D} \tag{4.2}$$

$\bar{\omega}_t$ is a positive number bounded by the video length, whereas $\bar{\mu}_t$ takes values between 0 and 1 and represents the average percentage of video watched.

We observe that video duration is an important covariate on watch percentage. In the Tweeted videos dataset, duration alone explains more than 58% of the variance of watch percentage. Intuitively, longer videos are less likely to be fully watched compared to shorter videos due to the limited human attention span.

We construct two 2-dimensional maps, where the x-axis shows video duration D, and the y-axis shows average watch time $\bar{\omega}_{30}$ (Figure 4.3a) and average watch percentage $\bar{\mu}_{30}$ (Figure 4.3b) over the first 30 days. We project all videos in the Tweeted videos dataset onto both maps. The x-axis is split into 1,000 equally wide bins in log scale. We choose 1,000 bins to trade-off enough data in each bin and having enough bins. We have also tried discretizing to smaller or larger number of bins, but the results are visually similar. We merge bins containing a very low number of videos (<50) to nearby bins. Overall, each bin contains between 50 and 38,508 videos. The color shades correspond to data percentiles inside each bin: the darkest color corresponds to the median value and the lightest correspond to the extremes (0% and 100%). Both maps calibrate watch time and watch percentage against video duration: highly-watched videos are positioned towards the top of allocated bin, while barely-watched videos are at the bottom compared to other videos with similar length.

Those two maps are logically identical because the position of each video in Figure 4.3(b) can be obtained by normalizing with its duration in Figure 4.3(a). It is worth noticing that a linear trend exists between average watch time and video duration in the log-log space, with an increasing variance as duration grows. In this work, we predominantly use the map of watch percentage (Figure 4.3b) given its y-axis is bounded between [0,1], making it easier to interpret. We denote this map as the engagement map.

Note that our method of constructing the engagement map resembles the idea of non-parametric quantile regression, which essentially computes a quantile regression fit in an equally spaced span [Koenker and Hallock, 2001]. For smaller datasets, using quantile regression may result in a smoother mapping. We tried quantile regression on Tweeted videos dataset, and we found that the values on both tails are inaccurate as the polynomial fits do not accurately reflect nonlinear trends. In contrast, our binning method works better in this case. Finally, we remarks that the engagement map can be constructed at different ages, which allows us to study the temporal evolution of engagement (Section 4.3.1).

4.3.3 New metric – relative engagement

Based on the engagement map, we propose the relative engagement $\bar{\eta}_t \in [0, 1]$, defined as the rank percentile of video in its duration bin. This is an average engagement measure in the first t days. Figure 4.3(b) illustrates the relation between video duration D, watch percentage $\bar{\mu}_{30}$, and relative engagement $\bar{\eta}_{30}$ for three example videos. Video v_1 (*d_8ao3o5ohU*) shows kids doing karate and v_2 (*akuyBBIbOso*) is about teaching toddlers colors. They are both about 5 minutes, but have different watch percentages, $\bar{\mu}_{30}(v_1)$= 0.70 and $\bar{\mu}_{30}(v_2)$=0.21. These amount to very different values of the relative engagement: $\bar{\eta}_{30}(v_1)$=0.96, while $\bar{\eta}_{30}(v_2)$=0.07. Video v_3 (*WH7llf2vaKQ*) is a much longer video (D=3 hours 49 minutes) showing a live fighting show. It has a relatively low watch percentage ($\bar{\mu}_{30}(v_3)$=0.19), similar to v_2. However, its relative engagement $\bar{\eta}_{30}(v_3)$ amounts to 0.99, positioning it among the most engaging videos in its peer group.

We denote the mapping from watch percentage $\bar{\mu}_t$ to relative engagement $\bar{\eta}_t$ as f, and its inverse mapping as f^{-1}. Here f is implemented as a length 1,000 look up table with a maximum resolution of 0.1% (or 1,000 ranking bins). For a given video with duration D, we first map it to corresponding bin on the engagement map, then return the engagement percentile by watch percentage. Equation (4.3) describes the mapping between relative engagement and average watch percentage using engagement map.

$$\bar{\eta}_t = f(\bar{\mu}_t, D) \Leftrightarrow \bar{\mu}_t = f^{-1}(\bar{\eta}_t, D) \tag{4.3}$$

While researchers have observed that watch percentage is affected by video duration [Guo et al., 2014; Park et al., 2016], to the best of our knowledge, this work is the first to quantitatively map its non-linear relation with video duration and present measurements in a large-scale dataset.

We choose day 30 to construct the engagement map and compute relative engagement because we need a reasonably long window for the watching statistics to stabilize. Theoretically, one can pick any duration to compute such metrics but the results will have larger variance. In fact, we vary the time window to compute η_7 to η_{30} for investigating the temporal dynamics of relative engagement in Section 4.3.5.

4.3.4 Linking relative engagement and video quality

Recent studies show that the quality of a digital item is linked to the audience's decision to continue watching or listening after first opening it [Salganik et al., 2006; Krumme et al., 2012]. Therefore, the average amount of time that the audience spend on watching a video should be indicative of video quality.

We examine the relation between relative engagement and video quality. We place the QUALITY VIDEOS datasets (Section 4.3.4) on the engagement map. Figure 4.4(a) plots the engagement map of all Music videos in the TWEETED VIDEOS (blue), that of the VEVO VIDEOS (red), and the videos in the BILLBOARD VIDEOS as a scatter plot (black dots). Similarly, Figure 4.4(b) plots the engagement map of all News videos in the TWEETED VIDEOS in blue and that of the TOP NEWS VIDEOS in red. All the maps are built from observations in the first 30 days.

Visibly, the QUALITY VIDEOS are skewed towards higher relative engagement values in both figures. Most notably, 44 videos in the BILLBOARD VIDEOS dataset (70% of the dataset) possess a high relative engagement of over 0.9. The other 30% of videos have an average $\bar{\eta}_{30}$ of 0.83 with a minimum of 0.54. For QUALITY VIDEOS, the

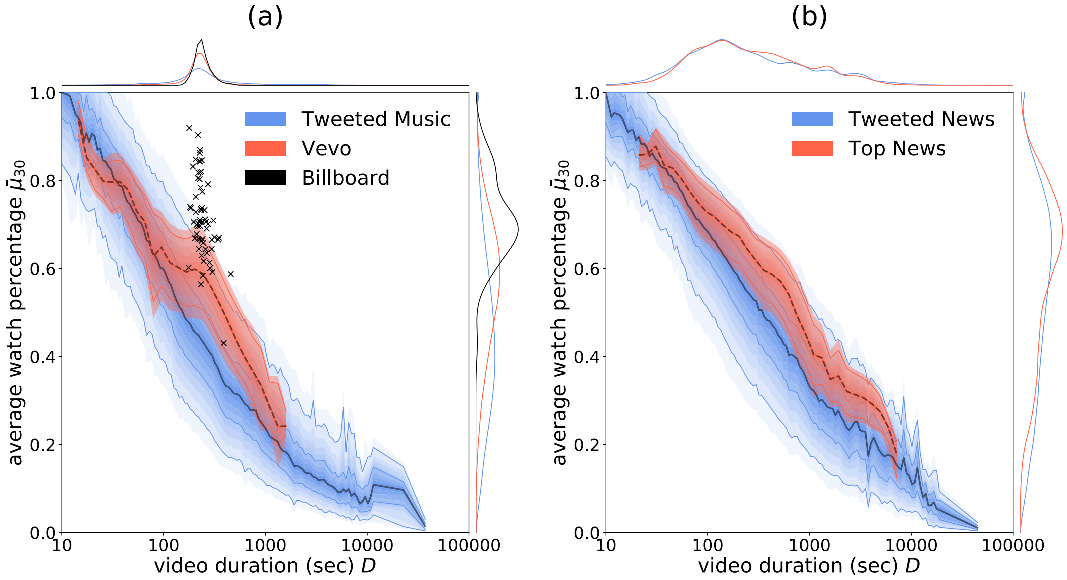


Figure 4.4: Relative engagement and video quality for Music **(a)** and News **(b)**. Videos in QUALITY VIDEOS dataset are shifted towards higher relative engagement compared to that in TWEETED VIDEOS. Best viewed in colors.

1-dimensional density distribution of average watch percentage $\bar{\mu}_{30}$ also shifts to the upper end as shown on the right margin of Figure 4.4. Overall, relative engagement values are high for content judged to be high quality by experts and the community. Thus, relative engagement is one plausible surrogate metric for content quality.

Relative engagement within channel. Figure 4.5 shows the engagement mapping results of 25 videos within one channel (*PBABowling*). This channel uploads sports videos about Professional Bowlers Association with widely varying lengths – from 2-minute player highlights to 1-hour event broadcasts. Video length has a signifi-

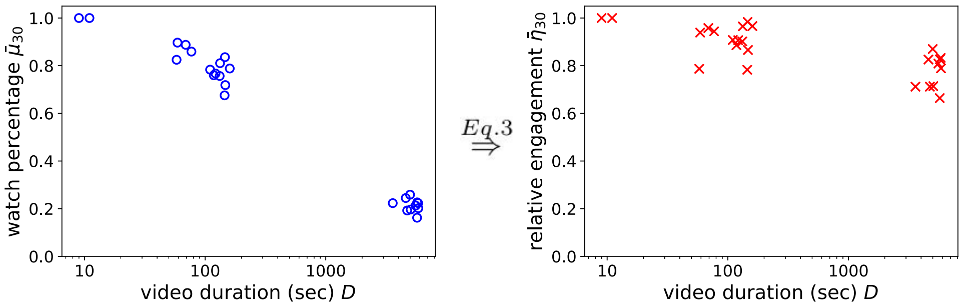


Figure 4.5: Watch percentage $\bar{\mu}_{30}$ (left) and relative engagement $\bar{\eta}_{30}$ (right) for videos in channel PBABowling. While it appears that $\bar{\mu}_{30}$ has a linear relation with the logarithmic duration $\log_{10} D$, $\bar{\eta}_{30}$ can be reasonably explained by only using the mean value of $\bar{\eta}_{30}$.

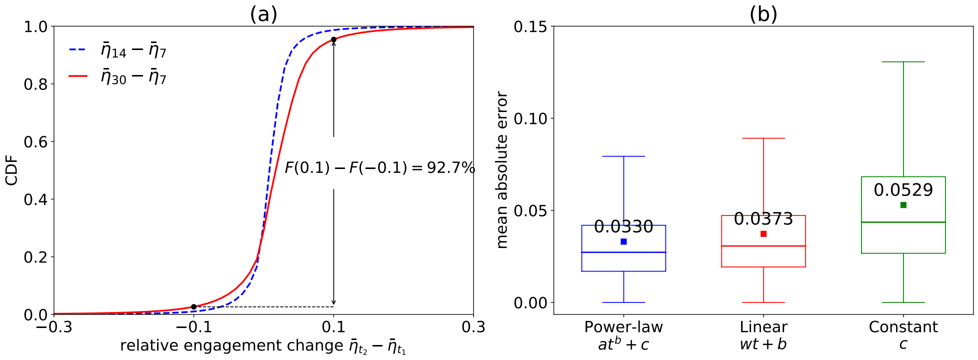


Figure 4.6: Relative engagement is stable over time. **(a)** CDF of temporal change in relative engagement of day 7 vs. day 14 (blue), day 7 vs. day 30 (red). **(b)** Fitting error of power-law model (blue), linear regressor (red) and constant function (green) in Tweeted videos.

cant impact: the short video cluster has mean average watch percentage $\bar{\mu}_{30}$ of 0.82, whereas the long video cluster has mean $\bar{\mu}_{30}$ of 0.21. However, after mapping to relative engagement, those two clusters have mean $\bar{\eta}_{30}$ of 0.92 and 0.78 – much more consistent within this channel than measured by watch percentage. Overall, the mean relative engagement of channel *PBABowling* is 0.86, which suggests this channel is likely to produce more engaging videos than an average YouTube channel, regardless of the video length. This example illustrates that video relative engagement tends to be stable within the same channel, and sheds some light on using past videos to predict future relative engagement.

4.3.5 Temporal dynamics of relative engagement

How does engagement change over time? This question is important because popularity dynamics tend to be bursty and hard to predict [Yang and Leskovec, 2011; Matsubara et al., 2012]. If engagement dynamics can be shown to be stable, it is useful for content producers to understand watch patterns from early observation. Note that the method for constructing the engagement map is the same, but one can use data at different ages t to build different mapping function $f(\bar{\mu}_t, D)$.

Engagement metrics are stable over time. We examine the temporal change of relative engagement at two given days t_1 and t_2 ($t_1 < t_2$) in Tweeted videos. We denote the cumulative distribution function (CDF) as $F_x(\Delta\bar{\eta})$, where $x = \bar{\eta}_{t_2} - \bar{\eta}_{t_1}$. This computes the fraction of videos with relative engagement changing *less* than $\Delta\bar{\eta}$ during t_1 to t_2. Figure 4.6(a) shows $\Delta\bar{\eta}$ distribution of day 7 *vs* day 14 and day 7 *vs* day 30. There are 4.6% of videos that increase more than 0.1 and 2.7% that decrease more than 0.1, yielding 92.7% of the videos with an absolute relative engagement change of less than 0.1 between day 7 and day 30. Such a small change results from the fact

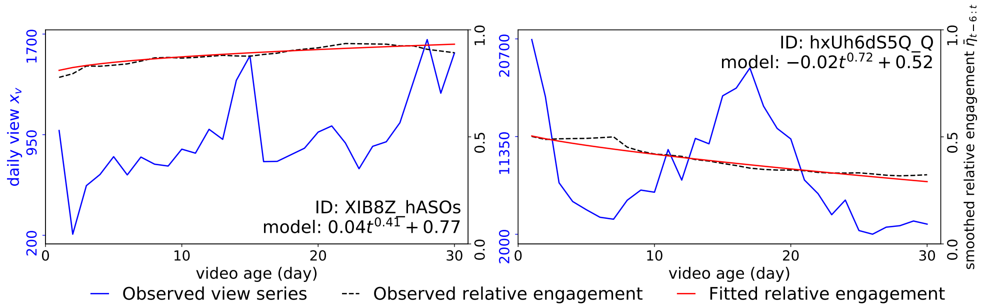


Figure 4.7: Temporal view series (blue) and smoothed daily relative engagement (black dashed) fitted by generalized power-law model $at^b + c$ (red).

that relative engagement $\bar{\eta}_t$ is defined as average measure over the past t days. It suggests that future relative engagement can be predicted from early watch patterns within a small margin of error. Similarly, this observation extends to both average watch percentage $\bar{\mu}_t$ and average watch time $\bar{\omega}_t$.

Next, we examine relative engagement on a daily basis. To avoid days with zero views, we use a 7-day sliding window, i.e., changing the summations in Equation (4.1) to between $t-6$ and t, yielding a smoothed daily watch percentage $\bar{\mu}_{t-6:t} = \frac{\sum_{i=t-6}^{t} x_w[i]}{D\sum_{i=t-6}^{t} x_v[i]}$. We then convert $\bar{\mu}_{t-6:t}$ to smoothed daily relative engagement $\bar{\eta}_{t-6:t}$ via the corresponding engagement map. For $t{<}7$, we calculate relative engagement from all prior days before t.

Figure 4.7 shows the daily views and smoothed relative engagement over the first 30 days of two example videos. While the view series has multiple spikes (blue), relative engagement is stable with only a slightly positive trend for video *XIB8Z_hASOs* and a slightly negative trend for *hxUh6dS5Q_Q* (black dashed). View dynamics have been shown to be affected by external sharing behavior [Rizoiu and Xie, 2017], the stability of relative engagement can be explained by the fact that it measures the average watch pattern but not how many people view the video.

Fitting relative engagement dynamics. We examine the stability of engagement metrics across the entire Tweeted Videos dataset. If the engagement dynamics can be modeled by a parametric function, one can forecast future engagement from initial observations. To explore what best describes the gradual change of relative engagement $\bar{\eta}_t$, we examine 3 functions: generalized power-law model ($at^b + c$), linear regressor ($wt + b$), and constant (c) function. For videos in Tweeted videos, we fit each of the three functions to smoothed daily relative engagement series $\bar{\eta}_{t-6:t}$ over the first 30 days. Figure 4.6(b) shows that power-law function fits best on the dynamics of relative engagement, with an average mean absolute error of 0.033.

To sum up, we observe that engagement metrics (relative engagement, watch percentage, and watch time) are stable throughout a video's lifetime, which implies that early watching patterns are strong predictors for future engagement and makes them attractive prediction targets. It is desirable to predict them in a cold-start setting, i.e., before videos get uploaded or any viewing behavior is observed.

4.4 Predicting aggregate engagement

In this section, we predict relative engagement and average watch percentage of a video in a cold-start setting. We further analyze the importance of each video feature on predicting engagement metrics.

4.4.1 Experimental setup

Prediction targets. We setup two regression tasks to predict average watch percentage $\bar{\mu}_{30}$ and relative engagement $\bar{\eta}_{30}$. Watch percentage is intuitively useful for content producers, while relative engagement is designed to calibrate watch percentage against duration as detailed in Section 4.3.3. It is interesting to see whether such calibration changes prediction performance. We report three evaluation results: predicting relative engagement and watch percentage directly, and predicting relative engagement then mapping to watch percentage via engagement map by using Equation (4.3). We do not predict average watch time because it can be deterministically computed by multiplying watch percentage and video duration.
Training and test data. We split TWEETED VIDEOS at 5:1 ratio over publish time. We use the first 51 days (2016-07-01 to 2016-08-20) for training, containing 4,455,339 videos from 1,132,933 channels; and the last 11 days for testing (2016-08-21 to 2016-08-31), containing 875,865 videos from 366,311 channels. 242,017 (66%) channels in the test set have appeared in training set, however, none of the videos in the test set is in the training set. The engagement map between watch percentage and relative engagement is built on the training set over the first 30 days. We split the dataset in time to ensure that learning is on past videos and prediction is on future videos.
Evaluation metrics. Performance is measured with two metrics:

- Mean Absolute Error $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$
- Coefficient of Determination $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}$

Here y is the true value, $\hat{y}$ the predicted value, $\bar{y}$ the average; i indexes samples in the test set. MAE is a standard metric for average error. R^2 quantifies the proportion of

Control variable (D)	
Duration	Logarithm of duration in seconds
Context features (C)	
Definition	Binary, high definition or not
Category	One hot encoding of 18 categories
Language	One hot encoding of 55 languages
Freebase topic features (T)	
Freebase topics	One hot sparse representation of 405K topics
Channel reputation features (R)	
Activity level	Mean number of daily upload
Past engagement	Mean, std and five points summary of previously uploaded videos
Channel specific predictor (CSP)	
One predictor for each channel using available features	

Table 4.3: Overview of features for predicting engagement.

the variance in the dependent variable that is predictable from the independent variable [Allen, 1997], and is often used to compare different prediction problems [Martin et al., 2016]. A lower MAE is better whereas a higher R^2 is better.

4.4.2 Features

We describe each YouTube video with 4 types of features as summarized in Table 4.3.
Control variable. Because video duration is the primary source of variation for engagement (Figure 4.3), we use *duration* as a control variable and include it in all predictors. In Tweeted Videos dataset, durations vary from 1 second to 24 hours, with a mean value of 12 minutes and median of 5 minutes. We take the logarithm (base 10) of duration to account for the skewness.
Context features. Context features are provided by video uploader. They describe basic video properties and production quality [Hessel et al., 2017].

- *Definition:* "1" represents high definition (720p or 1080p) and "0" represents low definition (480p, 360p, 240p or 144p). High definition yields better perceptual quality and encourages engagement [Dobrian et al., 2011].

- *Category:* broad content identifications assigned by video producers, the full list is shown in Table 4.2. Here we encode it as an 18-dimensional one-hot vector.

- *Language:* we run langdetect package on the video description and choose the most likely language. langdetect implements a Naive Bayes classifier to detect 55 languages with high precision [Shuyo, 2010]. The language is indicative of audience demographics.

Freebase topics features. YouTube labels videos with Freebase entities [Bollacker et al., 2008]. These labels incorporate user engagement signals, video metadata and content analysis, and are built upon a large amount of data and computational resources. With the recent advances in computer vision and natural language processing, there may exist more accurate methods for annotating videos. However, one can not easily build such an annotator at scale, and finding the best video annotation technique is beyond the scope of this work. On average, each video in the Tweeted Videos dataset has 6.16 topics. Overall, there are 405K topics and 98K of them appear more than 6 times. These topics vary from broad categories (*Song*), to specific object (*Game of Thrones*), celebrities (*Adele*), real-world events (*2012 Seattle International Film Festival*) and many more. Such fine-grained topics are descriptive of video content. While learning embedding vectors can help predict engagement [Covington et al., 2016], using raw Freebase topics enables us to interpret the effect of individual topic (Section 4.4.4).

Channel reputation features. Prior research shows that user features are predictive for product popularity [Martin et al., 2016; Mishra et al., 2016]. Here we compute feature from a channel's history to represent its reputation. We could not use social status indicators such as the number of subscribers, because it is a time-varying quantity and the value when a video is uploaded can not be retrospectively obtained. Thus, we compute two proxies for describing channel features.

- *Activity level:* mean number of daily published videos by channels in the training data. Intuitively, channels with higher upload rates reflect better productivity.
- *Past engagement:* relative engagement of previously uploaded videos from the same channel in the training set. Here we compute mean, standard deviation and five points summary: median, 25th and 75th percentile, min and max.

Several features used in prior works are interesting, but they do not apply in our setting. Network traffic measurement [Dobrian et al., 2011] requires access to the hosting backend. Audience reactions such as likes and comments [Park et al., 2016] can not be obtained before a video's upload.

4.4.3 Methods

We use linear regression with L2-regularization to predict engagement metrics, $\bar{\eta}_{30}$ and $\bar{\mu}_{30}$, both lie between 0 and 1. Since the dimensionality of Freebase topics features is high (4M x 405K), we convert the feature matrix to a sparse representation, allowing the predictor to be trained on one workstation. We adopt a fall-back strategy to deal with missing features. For instance, we use the context predictor for videos for

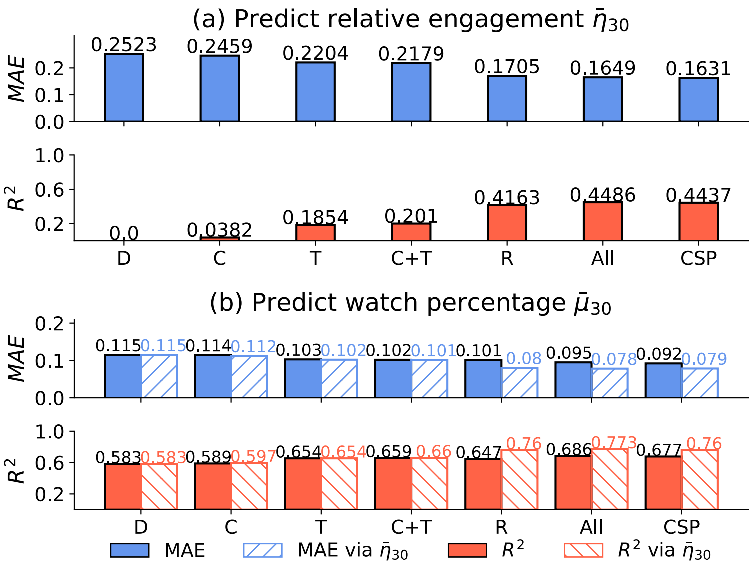


Figure 4.8: Summary of engagement prediction with two metrics, *MAE*: lower is better; R^2: higher is better. (a): Performance for predicting $\bar{\eta}_{30}$ in different feature combinations. (b): Performance for predicting $\bar{\mu}_{30}$ in different feature combinations, directly (solid bars, left) or via relative engagement $\bar{\eta}_{30}$ (shaded bars, right). Predicting watch percentage via converting relative engagement performs better than predicting watch percentage directly in all predictors.

which the channel reputation features are unavailable. The fall-back setting usually results in a lower prediction performance, however, it allows to predict engagement for *any* video. We have also tried KNN regression and support vector regression, but they do not yield better performances. All the models are implemented using the Scikit-learn Python package.

Channel specific predictor (CSP). In addition to the shared predictor, we train a separate predictor for each channel that has at least 5 videos in the training set. This fine-grained predictor covers 61.4% videos in the test data and may capture the "on-topic" effect within channel [Martin et al., 2016]. Intuitively, a channel might have specialty on certain topics and videos about those attract the audience to watch longer. For the remaining 38.6% videos, we use the shared linear regressor with all available features.

4.4.4 Results and analysis

Figure 4.8(a) bottom summarizes the results of predicting the relative engagement $\bar{\eta}_{30}$. Context (**C**) and Freebase topics (**T**) alone are weak predictors, explaining 0.04 and 0.19 variance of $\bar{\eta}_{30}$ in the test set. Combining the two (**C+T**) yields a slight gain over Freebase topics. Channel reputation (**R**) is the strongest feature, achieving

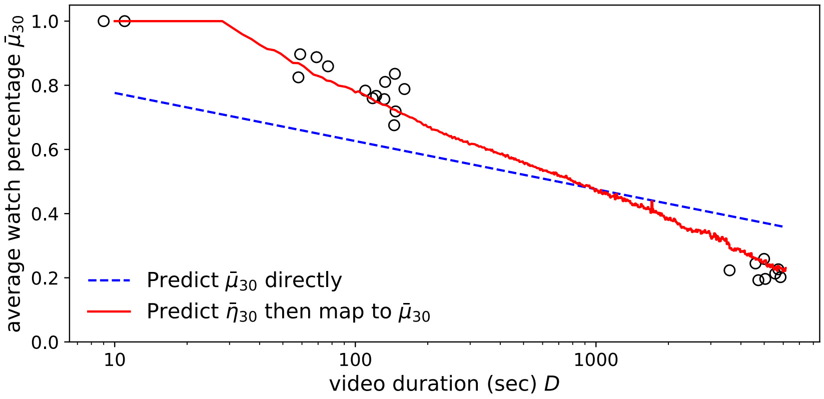


Figure 4.9: Shared linear regressors with channel reputation features on channel PBABowling, for predicting $\bar{\mu}_{30}$ (blue dashed) and predicting $\bar{\eta}_{30}$ then mapping to $\bar{\mu}_{30}$ (red solid).

R^2=0.42, and is slightly improved by adding context and Freebase topics. Channel-specific predictor (**CSP**) performs similarly to the All-feature predictor (**All**), suggesting that one can use a shared predictor to achieve similar performance with finer-grained per-channel model for this task. The analysis using MAE (Figure 4.8a top) also illustrates the same observations.

Average watch percentage $\bar{\mu}_{30}$ is easier to predict, achieving R^2 up to 0.69 (Figure 4.8b bottom) by using all features. Interestingly, predicting $\bar{\eta}_{30}$ then mapping to $\bar{\mu}_{30}$ consistently outperforms direct prediction of $\bar{\mu}_{30}$, achieving R^2 of 0.77. This shows that removing the influence of video duration via engagement map is beneficial for predicting engagement.

To understand why predicting via $\bar{\eta}_{30}$ performs better, we examine the shared linear regressors in both tasks. For simplicity, we include video duration and channel reputation features as covariates, and exclude the (generally much weaker) context and Freebase topics features for this example. In Figure 4.9, we visualize the two shared channel reputation predictors (**R**) at different video lengths for channel *PBABowling* (also shown in Figure 4.5): one predicts $\bar{\mu}_{30}$ directly (blue dashed), and the other predicts $\bar{\eta}_{30}$, then maps to $\bar{\mu}_{30}$ via the engagement map (red solid). The engagement map captures the non-linear effect for both short and long videos. In contrast, predicting $\bar{\mu}_{30}$ directly does not capture the bimodal duration distribution here: it overestimates for longer videos and underestimates for shorter videos.

Analysis of failed cases. We investigate the causes of failed prediction for each predictor. The availability of channel information seems important – for most poorly predicted videos, their channels have only one or two videos in the training set. Moreover, some topics appear more difficult to predict than others. For example, videos that are labeled with *music* obtain a MAE score of 0.175 ($\bar{\eta}_{30}$ using the All-feature predictor). This amounts to an error increase of 28% compared to videos labeled with *obama* (MAE=0.136). Lastly, the prediction performance varies considerably even for videos from the same channel and identically labeled. For example, the channel *Smyth Radio* (id: *UC79quCUqSgHyAY9Kwt1V6mg*) released a series of videos about "United States presidential election", 8 of which are in our dataset: 6 are in the training set and 2 are in the test set. These videos have similar lengths (3 hours) and they are produced in a similar style. The 6 videos in training set are watched on average between 3 and 10 minutes, yielding a $\bar{\eta}_{30}$ of 0.08. However, the 2 videos in the test set achieve considerable attention – 1.5 hours watch time on average, projecting $\bar{\eta}_{30}$ at 1.0. One possible explanation is that the videos in the test set discuss conspiracy theories and explicitly lists them in the title.

Overall, engagement metrics are predictable from context, topics and channel information in a *cold-start* experiment setting. Although channel reputation information is the strongest predictor, Freebase topics features are also somewhat predictive.

4.4.5 Are Freebase topics informative?

In this section, we analyze the Freebase topics features in detail and provide actionable insights for producing videos. Firstly, we group videos by Freebase topic and extract the most frequent 500 topics. Next we measure the amount of information gain with respect to relative engagement conditional entropy, defined in following equation:

$$H(Y|X_i{=}1) = -\sum_{y \in Y} P(y|x_i{=}1) \log_2 P(y|x_i{=}1) \tag{4.4}$$

Each topic is represented as a binary variable $x_i{\in}\{0, 1\}$, for $i{=}1, \ldots, 500$. We divide relative engagement into 20 bins, and y is the discretized bin. A lower conditional entropy indicates the presence of current topic is informative for engagement prediction (either higher or lower). Here we calculate $H(Y|X = 1)$ rather than $H(Y|X)$, because $X = 0$ represents the majority of videos for most topics and the corresponding term will dominate. Using $H(Y|X = 1)$ quantifies its effect only when the topic is in presence [Sedhain et al., 2013]. Figure 4.10 is a scatter plot of topic size and conditional entropy. Here large topics such as *book* (3.2M videos) or *music* (842K videos) have high conditional entropy and mean relative engagement close to 0.5, which suggests they are not informative in predicting engagement. All informative topics (e.g., with conditional entropy 4.0 and lower) are relative small (e.g., appearing around 10K times in the training set).

Figure 4.10 (inset) plots two example topics that are very informative on engage-

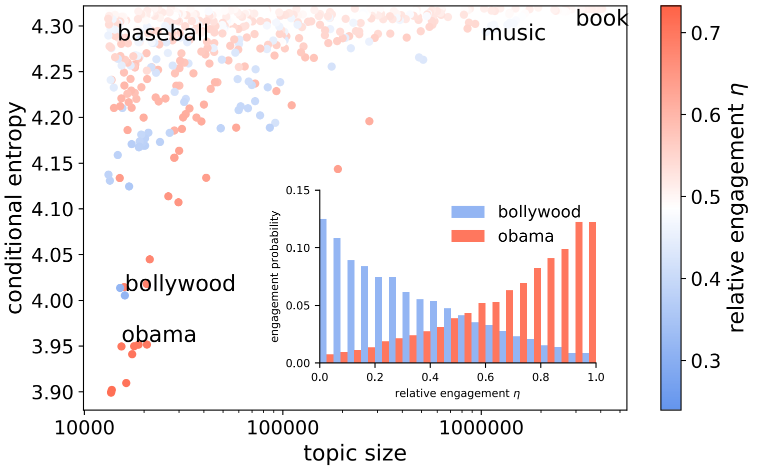


Figure 4.10: Informativeness for the most frequent 500 Freebase topics, measured by conditional entropy.

ment, from which we observe that videos about *bollywood* are more likely to have a low relative engagement while topic *obama* tends to keep audience watching longer. However, not all small topics are informative. A counter-example is *baseball*, which has a small topic size but a high condition entropy value.

In summary, watch percentage and relative engagement are predictable in a cold-start setting, before any behavioral data is collected. A few content-based semantic topics are predictive of low- or high- engagement. Such observation can help content producers choose engaging topics in video production.

4.5 Forecasting temporal engagement

While average engagement metrics are relatively stable over time, the amount of views and watch time fluctuate from day to day. In this section, we leverage the state of the art method on forecasting YouTube video viewership – Hawkes Intensity Process [Rizoiu et al., 2017b] – to forecast the time series of daily watch time.

4.5.1 Experimental setup

Forecasting targets. The daily watch time of a video, denoted as $x_w[t]$, is a time series with no regular shape. It can go through multiple long-term rising and falling phases [Yu et al., 2015], or have noisy short-term fluctuations from day to day [Matsubara et al., 2012]. Figure 4.11 illustrates the daily watch time of 2 videos (in hours, dashed black line). Albeit complicated, watch time is reported to be the central metric in YouTube recommender system [youtube.com, 2012; Covington et al., 2016].

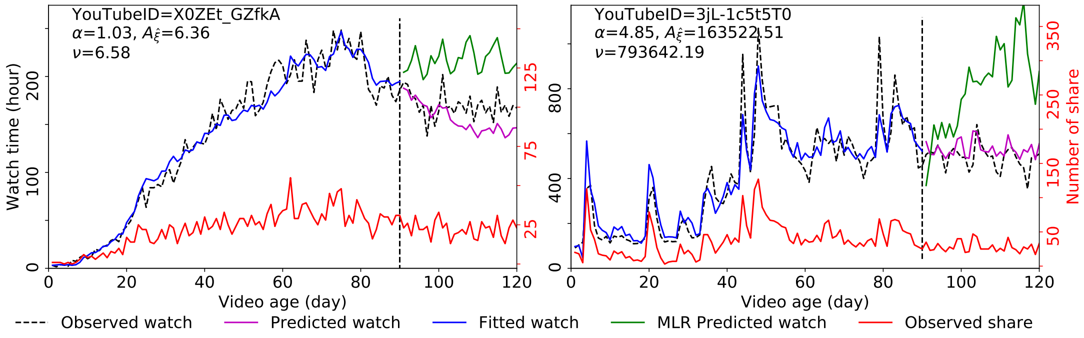


Figure 4.11: HIP fitting and forecasting for music video *X0ZEt_GZfkA* and *3jL-1c5t5T0*.

Our target is to forecast the future values of $x_w[t]$, given the observed watch time series in the past. To contrast the predictability, we also replicate the original task of forecasting video daily views, denoted as $x_v[t]$.

Dataset and splitting. We obtain the released Active videos '14 dataset from [Rizoiu et al., 2017b], and further enhance it with the watch time information with our software "YouTube-insight". This dataset contains 13,738 YouTube videos that are uploaded in 2014 and are tweeted at least 100 times within 120 days of their onsets. For the data construction details, we refer to the original paper [Rizoiu et al., 2017b].

We adopt the same data splitting strategy: we use the first 90 days of each video's viewing, watching, and sharing history to estimate model parameters. The learned parameters are later used to forecast future values of $x_v[t]$ and $x_w[t]$ between day 91 and day 120.

Evaluation metric. We use the absolute percentile error (APE) to measure forecasting results. APE compares the percentile difference of true value $\sum_{t=91}^{120} x[t]$ and predicted value $\sum_{t=91}^{120} \hat{x}[t]$. This metric effectively computes the ranking position change of each video against the entire dataset in terms of predicted targets. A lower APE indicates better forecasting performance.

4.5.2 Methods - Hawkes Intensity Process (HIP)

It is notable that temporal attention metrics (popularity and engagement) are highly unpredictable [Cheng et al., 2014; Martin et al., 2016] because the external social environment may change suddenly. One encouraging recent result from Rizoiu et al. [2017b] shows the contrary. Their key insights are:

1. Attention metrics are responses to promotions in the external world. It will become more predictable once the promotions are accounted for.

2. Content quality and other scaling factors can be estimated from a modest amount of attention history.

We use these insights, and adopt their proposed Hawkes Intensity Process (HIP) to model daily views and watch time on YouTube.

HIP extends the well-known Hawkes (self-exciting) process [Hawkes, 1971] to describe the volume of activities within a fixed time interval (e.g. daily). This is done by taking expectations over the stochastic event history. Specifically, this model describes a self-exciting phenomenon that is commonly observed in online social network [Zhao et al., 2015; Mishra et al., 2016; Rizoiu and Xie, 2017]. It models the target quantity $x[t]$ as a self-consistent equation into three parts: the unobserved external influence, the effects of external promotions, and the influence from historical events. Formally, it can be written as

$$x[t] = \gamma \mathbb{1}[t = 0] + \eta \mathbb{1}[t > 0] + \alpha s[t] + C \sum_{\tau=1}^{t} x[t - \tau](\tau + c)^{-(1+\theta)} \tag{4.5}$$

The first two terms represent unobserved external influences. γ and η model the strengths of an initial impulse and a constant background rate, respectively. in the middle component, α is the sensitivity to external promotion, $s[t]$ is the volume of promotion, and $\alpha s(t)$ is the instantaneous response to promotion. In the last component, θ is the exponent of a power-law memory kernel $(\tau + c)^{-(1+\theta)}$. c is a nuisance parameter for keeping the kernel bounded, and C accounts for the latent content quality. Overall, this last component models the impact over its own event history $x[\tau]$ for τ=1 : t−1.

In our case, $x[t]$ is the time series of daily views $x_v[t]$ or daily watch time $x_w[t]$ (in hours). $s[t]$ is the daily number of shares (as tracked by YouTube). The parameter set $\{\gamma, \eta, \alpha, C, c, \theta\}$ is estimated from the first 90-day interval of each video using the constrained L-BFGS algorithm in SciPy Python package.

Baseline. Another widely used method for volume forecasting is Multivariate Linear Regression (MLR) [Szabo and Huberman, 2010; Pinto et al., 2013], i.e., estimating a weighted linear combination of historical volumes of attention and promotion on a set of training videos, and then applying the learned weights to new videos.

$$x[t] = \mathbf{w}^{t-1}\mathbf{x}[1 : t - 1] + \mathbf{w}_s^t \mathbf{s}[1 : t] \tag{4.6}$$

Here $x[t]$ is the prediction target on day t; $\mathbf{x}[1 : t{-}1]$ is historical signals from day 1 to day $t{-}1$; $\mathbf{s}[1{:}t]$ is the sharing (promotions) series on the item; $\mathbf{w}^{t-1}$ and $\mathbf{w}_s^t$ are weights estimated from training data. The MLR is a stronger baseline than other linear models such as auto-regressive moving average (ARMA), since the size of the

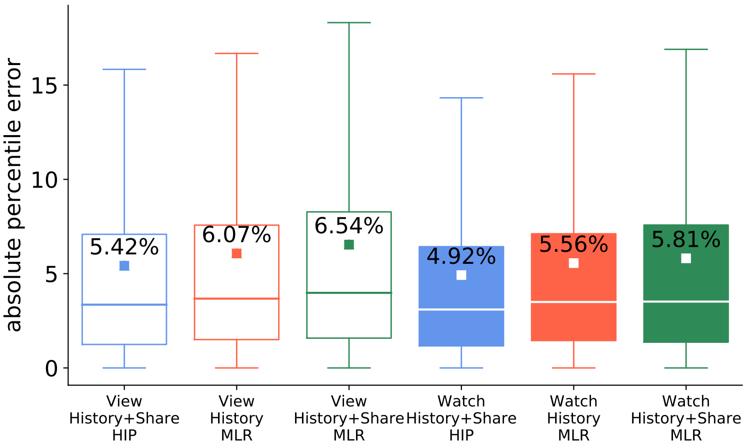


Figure 4.12: Forecasting errors of HIP and MLR on daily views (hollow) and watch time (solid) using historical attention and sharing dynamics.

moving window in MLR is the entire available history.

4.5.3 Results and analysis

We report results on the Active videos '14 dataset [Rizoiu et al., 2017b]. Using this dataset allows us to contrast the tasks of forecasting daily views and watch time. For both HIP and MLR, we estimate model parameters over the first 90 days and perform forecasting between day 91 and day 120. HIP is estimated using both attention history and sharing series, while MLR forecasts future value with historical data only, or with additional sharing information.

Our forecasting errors are shown in Figure 4.12, measured with APE. We observe that HIP consistently outperforms MLR in forecasting daily watch time (and views, reproducing earlier results in [Rizoiu et al., 2017b]). Moreover, the watching signal (engagement metric) seems more predictable than the viewing signal (popularity metric). The best HIP model of forecasting watch time (APE=4.92%) is better than that of forecasting daily views (APE=5.42%).

One possible explanation for this discrepancy can be attributed to the robustness of engagement metric. As users may be enticed by clickbait snippet, the amount of views is naturally more fluctuated than the watch time. Once users click on a video, they will quickly exit if they realize the content is inconsistent with the title, or spend more time watching if they perceive the video is of high quality. The inherent quality effectively plays a role in calibrating the amount of attention each video receives. However, the calibration only applies to watch time, but not view count.

HIP-view vs. HIP-watch. Two key quantities derived from the HIP model are the viral score α and the viral potential ν. The viral score α quantifies the degree of instantaneous response to external promotions. Figure 4.11 contains two videos with results of fitting and forecasting from both HIP (magenta) and MLR (green). The latino music video (*X0ZEt_GZfkA*) has a lower viral score than a Vevo video by Ricky Martin (*3jL-1c5t5T0*). This suggests that the latino music video needs more promotions to attract a similar amount of attention when compared to the Vevo video.

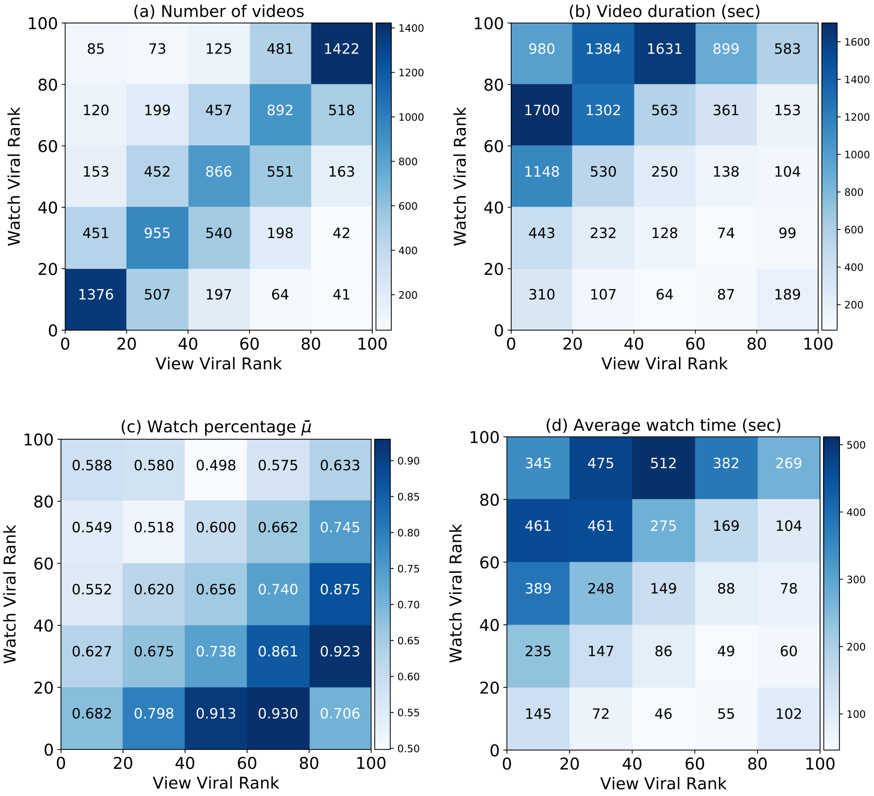


Figure 4.13: Correlation plots of viral rank (0–100 percentile) of HIP-view (x-axis) versus HIP-watch (y-axis). (a)-(d): number of videos, average video duration, average watch percentage and average watch time.

The viral potential ν quantifies how much attention will be gained for each impulse received. Here we introduce a nuanced concept *impulse*, denoted as $\hat{\xi}[t]$. For each share, it will generate α impulses. The viral potential ν is computed as a numeric integration of response to a unit of impulse from 0 to 10,000 time steps. At each time step, $\hat{\xi}[t]$ accounts for all past impulses $\hat{\xi}[t-\tau]$ over the power-law kernel.

$$\begin{aligned} \hat{\xi}[0] = 1; \quad \hat{\xi}[t] &= C \sum_{\tau=1}^{T} \hat{\xi}[t-\tau](\tau+c)^{-(1+\theta)} \text{ for } t \geq 1; \\ \nu &= \sum_{t=0}^{10000} \hat{\xi}[t] \end{aligned} \tag{4.7}$$

We compare viral potential ν from the HIP models on views and on watch time, by looking at correlation matrices after rank-normalizing their respective values of ν, dubbed *viral rank*. Figure 4.13(a) shows that the viral ranks of HIP-view and HIP-watch highly correlate with each other – 50.4% of videos are in the same 20-percentile bucket in both models (summing over the secondary diagonal), and the ranks of another 36.2% differ by one.

For each cell in Figure 4.13(a), we examine some video-level properties – duration, watch percentage, and average watch time. We visualize the results in Figure 4.13(b-d). The number within each cell denotes the mean value of the property for videos in corresponding cell. Videos with a high viral rank in watch but low rank in views (top left buckets) tend to have longer length (Figure 4.13b) and longer watch time (Figure 4.13c), whereas those with a low viral rank in watch but high rank in views (bottom right buckets) tend to have a very high watch percentage (generally >0.90, see Figure 4.13d).

In summary, self-exciting model can describe the attention dynamics of YouTube videos well. The task of forecasting future watch time is relatively easy compared with forecasting future view counts. The HIP model is a useful tool to quantity the gained attention given a known volume of social promotion, or identify videos that have the potential to become viral.

4.6 Conclusion

In this chapter, we measure a set of aggregate engagement metrics for online videos, including average watch time, average watch percentage, and a new metric, relative engagement. We study the proposed metrics on a public dataset of 5.3 million videos. We show that relative engagement is stable over the video lifetime, and is strongly correlated with established notions of video quality. We further predict the average watch percentage, and forecast the volume of future watch time. We find that average engagement metrics are quite predictable, even in a cold-start setting. Although forecasting future attention dynamics is difficult, we find the series of watch time is more predictable than view count.

4.6.1 Limitations

Our observations are only on publicly available videos. It is possible that untweeted, private, and unlisted videos behave differently. The attention data used are aggregated over all viewers of a video. Therefore, our observations here are more limited than those from content hosting sites, which have access to individual user attributes and behavior logs. Our results are item-centric and cannot directly translate to user-centric engagement study.

4.6.2 Practical implications and future work

The observations in this work provide content producers with a new set of tools to create engaging videos and forecast user behavior. For video hosting sites, engagement metrics can be used to optimize recommender systems and advertising strategies, as well as to detect potential clickbaits. For future work, one open problem is to quantify the gap between aggregate and individual measurements. Another is to extract more sophisticated features and to apply more advance techniques for improving the prediction and forecasting performance.

Chapter 5

Measuring and modeling online recommendation networks

Chapter 4 examined the engagement patterns of online videos at the aggregate level. In this chapter, we shift our focus to the popularity measure and its interplay with the recommender systems. At present, most video hosting sites implement recommender systems, which connect the videos into a directed network and build pathways for users to navigate from one video to another. However, little is known about how the collective human attention is allocated over such large-scale networks, and about the impacts of the recommender systems on driving overall traffic.

In Section 5.2, we 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. From Section 5.3 to Section 5.5, we present a set of measurements from the macroscopic, microscopic, and temporal perspectives. These measurements 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. In Section 5.6, 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, the observations and models in this chapter provide a new set of tools to better understand the impacts of recommender systems on collective social attention.

5.1 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 [Su et al., 2016], discovering interesting communities on LinkedIn [Sharma and Yan, 2013], and recommending similar goods on Amazon [Oestreicher-Singer and Sundararajan, 2012; Dhar et al., 2014]. In the domain of multimedia, service providers (e.g., YouTube, Netflix, and Spotify) use recommender systems to suggest related videos or songs [Davidson et al., 2010; Covington et al., 2016; Gomez-Uribe and Hunt, 2016; Zhang et al., 2012; Celma and Cano, 2008], and to generate non-stop media playlist [Bendersky et al., 2014; Chen et al., 2019]. Much effort has been on generating more accurate recommendations, but relatively little is said about the effects of recommender systems on overall attention, such as their effects on item popularity ranking, the estimated strength of item-to-item links, and global patterns on the attention gained due to being recommended. This chapter aims to answer such questions for online videos, using publicly available recommendation networks and attention time series.

Consistent with Chapter 4, we use the term *attention* to refer to a broad range of user activities with respect to an online item, such as clicking, commenting, sharing, or watching. In contrast, the term *popularity* is used to denote observed attention statistics that are often used to rank online items against each other. Our measurement and estimation are carried out on the largest online video platform YouTube, and we specifically quantify popularity using the number of daily views for each video. Note that though the validation in this chapter is limited to popularity, our outlined methods may well apply to other deeper forms of user engagement such as watch time.

We illustrate the goals of this chapter through an example. Figure 5.1(a) shows the recommendation network for six videos from the artist Adele[1]. It is a directed network and the directions imply how users can navigate between videos by following the recommendation links. Some videos are not directly connected but reachable within a few hops. For example, "Skyfall" is not on the recommended list of "Hello", but a user can visit "Skyfall" from "Hello" by first visiting "Rolling in the deep". Figure 5.1(b) plots the daily view series since the upload of each of the six videos. When "Hello" was released, it broke the YouTube debut records by attracting 28M views in the first 24 hours[2]. Simultaneously, we observe a traffic spike in all of her other

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

[2] https://www.billboard.com/articles/news/6745062/adele-hello-biggest-youtube-debut-this-year

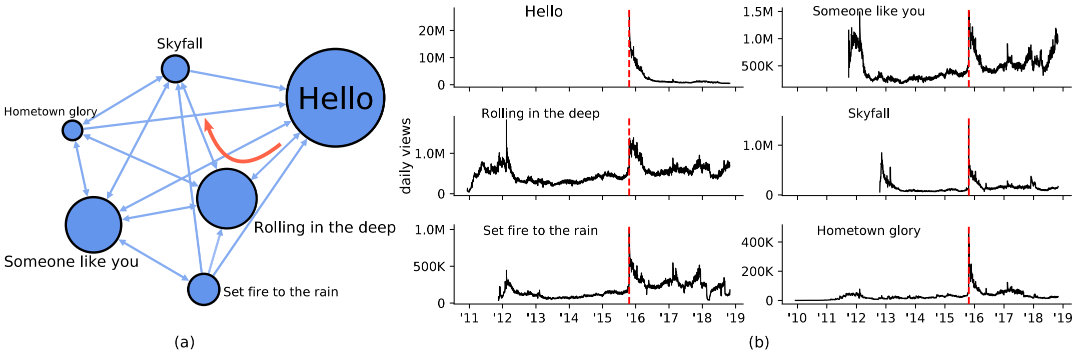


Figure 5.1: Observing the effects of recommendation network on video popularity. **(a)** A directed network consists of six videos by the artist Adele. The node size is proportional to the video's cumulative view counts till Nov 02, 2018. The red arrow highlights one possible route that users visit "Skyfall" from "Hello" in 2 hops. **(b)** View series for the six videos shown in **(a)**. x-axis is calendar year. Visually we observe a simultaneous spike across all videos when "Hello" was uploaded on Oct 22, 2015, denoted by red dashed vertical line.

videos, even in three videos that were not directly pointed by "Hello". This example illustrates that the viewing dynamics of videos connected directly or indirectly through recommendation links may correlate, and it prompts us to investigate the patterns of attention flowing between them.

In this chapter, we bridge two gaps in the current literature. The first gap measures and estimates the effects of recommender systems in complex social systems. The main goal of recommender systems is maximizing the chance that a user clicks on an item in the next step [Davidson et al., 2010; Covington et al., 2016; Bendersky et al., 2014; Yi et al., 2014] or in a longer time horizon [Beutel et al., 2018; Chen et al., 2019; Ie et al., 2019]. However, recommendation in social systems remains as an open problem for two reasons: (1) a limited conceptual understanding of how finite human attention is allocated over the network of content, in which some items gain popularity at the expense of, or with the assistance of others; (2) the computational challenge of jointly recommending a large collection of items. The second gap comes from a lack of fine-grained measurements on the attention captured by items structured as a network. There are recent measurements on the YouTube recommendation networks [Airoldi et al., 2016; Cheng et al., 2008], but their measurements are not connected to the attention patterns on content. Similarly, measurement studies on YouTube popularity [Zhou et al., 2010] quantify the overall volume of views directed from recommended links. However, no measurement that accounts for both the network structure and the attention flow is available for online videos.

This chapter tackles three research questions:

1. How to measure video recommendation network from public information?
2. What are the characteristics of the video recommendation network?
3. Can we estimate the attention flow in the video recommendation network?

We address the first question by curating a new YouTube dataset consisting of a large set of Vevo artists. This is the first dataset that records both the temporal network snapshots of a recommender system, and the attention dynamics for items in it. Our observation window lasts 9 weeks. We present two means to construct the non-personalized recommendation network, and we discuss the relation between them in detail (Section 5.2).

Addressing the second question, we conceptualize the global structure of the network as a bow-tie [Broder et al., 2000] and we find that the largest strongly connected component accounts for 23.11% of the videos while occupying 82.6% of the attention. Surprisingly, videos with high indegree are mostly songs with sustained interests, but not the latest released songs with high view counts. We further find that the network structure is temporally consistent on the macroscopic level, however, there is a significant link turnover on the microscopic level. For example, 50% of the videos with an indegree of 100 on a particular day will gain or lose at least 10 links on the next day, and 25% links appear only once during our 9-week observation window (Section 5.3 to Section 5.5).

Answering the third question, we build a model which employs both the temporal and network features to predict video popularity, and we estimate the amount of views flowing over each link. Our networked model consistently outperforms the autoregressive and neural network baseline methods. For an average video in our dataset, we estimate that 31.4% of its views are contributed by the recommendation network. We also find the evidence of YouTube recommender system boosting the popularity of some niche artists (Section 5.6).

The new methods and observations in this chapter can be used by content owners, hosting sites, and online users alike. For content owners, the understanding of how much traffic is driven among their own content or from/to other content can lead to better production and promotion strategies. For hosting sites, such understanding can help avoid social optimization, and shed light on building a fair and transparent content recommender systems. For online users, understanding how human attention is shaped by the algorithmic recommendation can help them be conscious of the relevance, novelty and diversity trade-offs in the content they are recommended to.

The main contributions of this chapter include:

- We curate a new YouTube dataset, called VEVO MUSIC GRAPH dataset[3], which contains the daily snapshots of the video recommendation network over a span of 9 weeks, and the associated daily view series for each video since upload.
- We perform, to our knowledge, the first large-scale measurement study that connects the structure of the recommendation network with video attention dynamics.
- We propose an effective model that accounts for the network structure to predict video popularity and to estimate the attention flow over each recommendation link.

5.2 Data

In this section, we first introduce our newly curated VEVO MUSIC GRAPH dataset (Section 5.2.1). Next, we detail the data collection strategy (Section 5.2.2) and analyze the relation between two types of non-personalized video recommendation lists (Section 5.2.3).

5.2.1 Vevo Music Graph dataset

The VEVO MUSIC GRAPH dataset consists of the verified Vevo artists who are active in six English-speaking countries (United States, United Kingdom, Canada, Australia, New Zealand, and Ireland), together with their complete record of videos uploaded on YouTube from the launch of Vevo (Dec 8, 2009) until Aug 31, 2018. Our dataset contains 4,435 Vevo artists and 60,740 music videos. For each video, we collect its metadata (e.g., title, description, uploader), its view count time series, and its recommendation relations with other videos. The videos and their recommendation relations form a dynamic directed network, which we capture daily between Sep 1, 2018 and Nov 2, 2018 (63 days, 9 weeks).

Why Vevo? Vevo[4] is the largest syndication hub that provides licensed music videos from major record companies to YouTube [Wikipedia.com, 2020b]. We choose to study the networked attention flow on Vevo for several reasons. First, Vevo is an ecosystem of its own that attracts tremendous attention — 94 of all-time top 100 most viewed videos on YouTube are music, and 64 of which are distributed via Vevo [wikipedia.com, 2020a]. On average, our dataset accounts for 310 millions views and 9.1 millions watch hours every day. Second, many users utilize YouTube as their

[3] The code and datasets are available at https://github.com/avalanchesiqi/networked-popularity

[4] The Vevo website was shut down on May 24, 2018. However, videos syndicated on YouTube before are still embedded with a "VEVO" watermark on their thumbnails. See screenshot in Figure 5.2 for illustration.

music streaming player, listening to non-stop playlists generated by the recommender systems. After the completion of the current video, YouTube automatically plays the "Up next" video – the video in the first position of the recommended list, as illustrated in Figure 5.2. This usage pattern for music videos makes the network effects of YouTube recommender systems more significant for directing user attention from one video to another. Third, Vevo artists and their videos form a tightly connected network. The average degree in the Vevo video network is 10, compared to 3.2 in the YouTube video network collected by Airoldi et al. [2016] via snowball sampling (see Section 5.2.3). The nodes are homogeneous in terms of content – they are all music videos from artists based in English-speaking countries. Lastly, the Vevo artists are easily identifiable – they include the keyword "VEVO" in the channel title, they possess a verification badge on the channel page, and they publish licensed videos with a "VEVO" watermark.

5.2.2 Data collection strategy

We identify Vevo artists starting from Twitter. We capture every tweet that mentions YouTube videos by feeding the rule "youtube" OR ("youtu" AND "be") into the Twitter Streaming API [twitter.com, 2020a]. Our Twitter crawler has been running continuously since Jun 2014. From the "extended_urls" field of each tweet, we extract the associated YouTube video ID, and we use our open-source tool "YouTube-insight" [Wu et al., 2018] to retrieve the video's metadata, daily view count series and the ranked list of relevant videos. Next, we select the Vevo artists by keeping only the channels that have the keyword "VEVO" in the channel title and a "verified" status badge on the channel homepage. Note that a channel refers to a user who uploads videos on YouTube. We query an open music database MusicBrainz[5] to retrieve more features about each artist, such as the music genres and the geographical area of activities. We retain the artists who are active in the six aforementioned English-speaking countries, and the videos that are classified into the "Music" category. For completeness, we also implement a snowball-like procedure to retrieve further artists and their videos by following the recommendation relations from the tweeted videos. However, this procedure only adds 2 more artists (out of the 4,435 Vevo artists in our dataset) and 5 more videos (out of the 60,740 music videos). This is not surprising, considering most artists would promote their works on social media platforms. One data limitation is that artists who are not affiliated with Vevo will not appear in our collection, such as Ed Sheeran and Christina Perry.

[5] https://musicbrainz.org

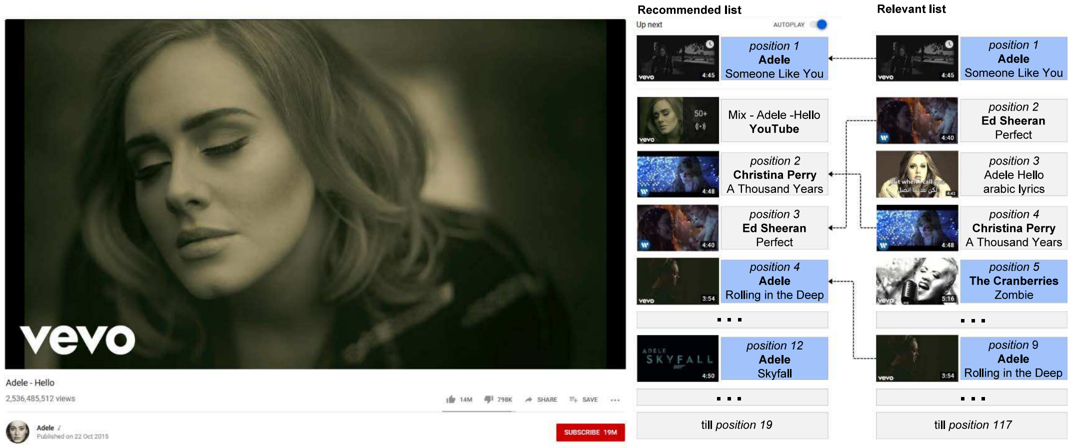


Figure 5.2: YouTube webpage layout for the video "Adele - Hello", together with its recommended and relevant lists. Vevo videos are colored in blue and included in our dataset. (Screenshot source: https://www.youtube.com/watch?v=YQHsXMglC9A)

5.2.3 The network of YouTube videos

For any YouTube video, there are two publicly accessible sources of recommendation relations. The first is the right-hand panel of videos that YouTube displays on its interface. We denote this as the *recommended list* (visualized in Figure 5.2). The second is from the YouTube Data API[6], which retrieves a list of videos that are relevant to the query video, ranked by the relevance. We denote this as the *relevant list*. We retrieve both the recommended and the relevant lists for every video in our dataset. We construct the recommended list by simulating a browser to access the video webpage and scraping the list on the right-hand panel. We retrieve the first 20 videos from the panel, which are the default number of videos shown to the viewers on YouTube. Note that typically, YouTube customizes the viewers' recommendation panel based on their personal interests and prior interaction history. Here, we retrieve the non-personalized recommended list by sending all requests from a non-logged in client and by clearing the cookies before each request. We denote the networks of videos constructed using the recommended and the relevant lists as the *recommended network* and the *relevant network*, respectively.

From Sep 1, 2018 to Nov 2, 2018, we crawled both the recommended and the relevant lists for each of the 60,740 Vevo videos on a daily basis. The crawling jobs were distributed across 20 virtual machines, and took about 2 hours to finish. In this way, we obtain successive snapshots for both the recommended and the relevant networks over 9 weeks.

[6] https://developers.google.com/youtube/v3/docs/search/list

An illustrative example. Figure 5.2 illustrates the YouTube webpage layout for the video "Hello" by Adele, together with its recommended and relevant lists. Videos belonging to the Vevo artists are colored in blue (e.g., Adele and The Cranberries), while others are colored in grey (e.g., Ed Sheeran and Christina Perry). Visibly, not all videos on the recommended and relevant lists belong to the Vevo artists (e.g., "Ed Sheeran - Perfect"). Notice that for Music videos, a platform-generated playlist is always shown at the second position of the recommended list (here, "Mix - Adele - Hello"), effectively capping the size of this list at 19. The length of the relevant list often exceeds 100. We observe that not all relevant videos appear in the recommended list (e.g., "The Cranberries - Zombie"), nor all recommended videos originate from the relevant list (e.g., "Adele - Skyfall"). Also, the relative positions of two videos can appear flipped between the two lists (e.g., "Ed Sheeran - Perfect" and "Christina Perry - A Thousand Years").

Display probabilities from the relevant to the recommended list. We study the relation between the positions of videos on the relevant and on the recommended lists. We construct four bins based on the video position on the recommended list (position 1, position 2-5, position 6-10, and position 11-15). Figure 5.3(a) shows as stacked bars the probability that a video ends up in each of the bins, as a function of its position on the relevant list. The total height of the stacked bars gives the overall probability that a video originating from the relevant list appears at the top 15 positions on the recommended list. We observe videos that appear at an upper position on the relevant list are more likely to appear on the recommended list, and at an upper position. For example, the video at position 1 on the relevant list has 0.34 probability to appear at the first position and 0.84 probability to appear at the top 15 positions on the recommended list. The probability decays for videos that appear at lower positions. A relevant video appearing in position 41 to 50 has less than 0.05 probability to appear on the recommended list. We compute the probabilities of appearance between each pair of positions in the relevant and the recommended lists – denoted as *display probabilities* – using the 9-week dynamic network snapshots.

In Figure 5.3(b), we show the plot of the probability that a video at a given position on the recommended list originates from the relevant list. We observe videos that appear at an upper position on the recommended list are more likely to originate from the upper position of relevant list. The other notable observation is that the overall recall for recommended videos are high at over 0.8, meaning for any video on the recommended list, we are likely to see it on the relevant list.

YouTube video network density. Airoldi et al. [2016] used the first 25 videos on the relevant list to construct the relevant network, which had an average degree of 3.2. By comparison, our Vevo video network is much denser at the same cutoff, with

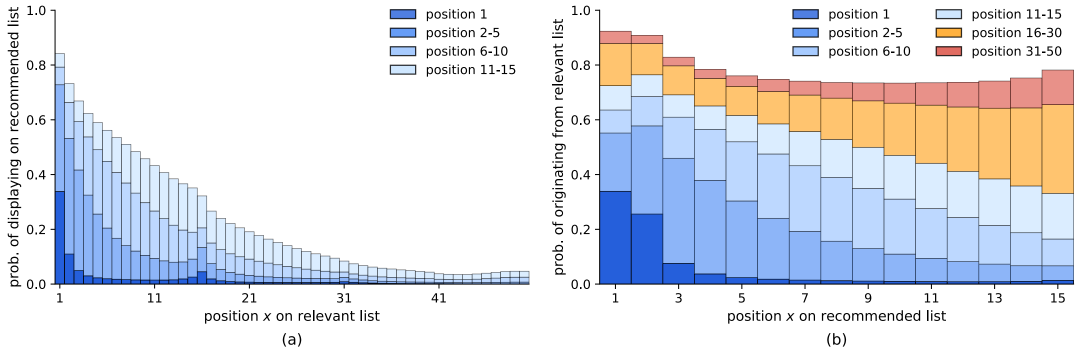
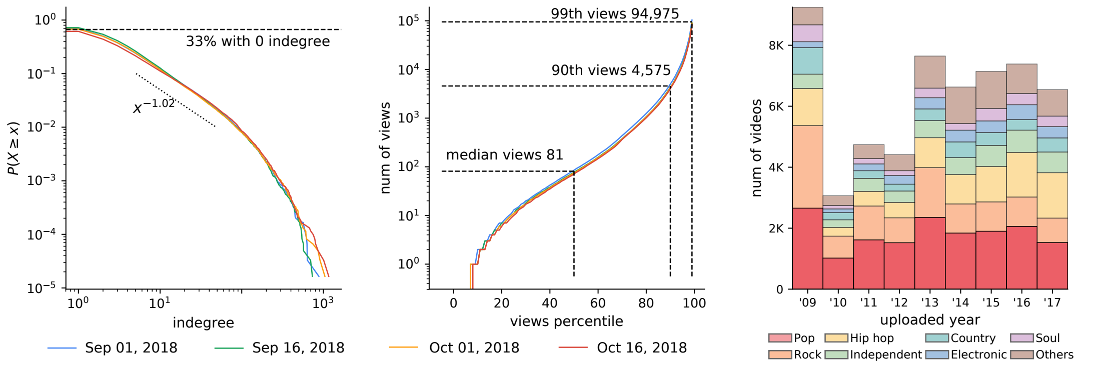


Figure 5.3: (a) Display probabilities of videos from the position x of the relevant list appear at different positions of the recommended list. **(b)** Probabilities of videos from the position x of the recommended list originate from different positions of the relevant list. **(a)** and **(b)** are similar but with the relevant and recommended lists along the x-axis and y-axis swapped.

an average degree of 10. One could expect that the relevant network becomes even denser when videos at lower position are included; however, the display probabilities also need to be considered. In this chapter and unless otherwise specified, we use the first 15 positions on the relevant list ($0.35 \leq P_{\text{display}} \leq 0.84$) to construct the relevant network. We denote this threshold as the *cutoff* and we study the impact of different cutoff values on the network structure in Section 5.3.3. Measurements with other cutoff values yield similar results and thus are omitted.

Discussion on the recommended and relevant lists. The notions of recommended and relevant lists have been previously adopted in the field of recommender systems [Herlocker et al., 2004]. The relevant list is usually hidden from the user-interface and ranked according to the semantic relevance between the query and the items. In contrast, the recommended list reflects the final recommendations in the user interface, i.e., displaying on the right-hand panel of the video webpage. On YouTube, the recommended list is a top-K sample from the concatenation of the relevant list, user demographics, watch history, search history, and spatial-temporal information [Covington et al., 2016]. All features, apart from the relevant list, are user-, time- and location-dependent. Hence, the displayed recommended list of the same video can be very different for two viewers, regardless of their logged-in state, location or viewing time. On the other hand, the relevant list is consistent for all requests, from any client during any period of time. We also observe the relevant list changes less frequently than the recommended list, which suggests it is more robust to the update of YouTube recommender systems. For these reasons, we use the relevant list to construct and measure YouTube video network.



(a) indegree distribution (b) daily views vs. its percentile (c) VEVO videos uploading trend

Figure 5.4: (a) The indegree distribution of the Vevo network for four snapshots. The x-axis is the (log) indegree, and the y-axis is the (log) CCDF. On average, 33% of all videos have no incoming links at the cutoff of 15. Best fitted power-law model is $x^{-1.02}$. **(b)** The average daily view counts distribution of Vevo network for the same four snapshots. The x-axis shows average daily view count percentiles, and the y-axis shows the raw number of view counts in log scale. Both **(a)** and **(b)** show that the macroscopic structure is temporally consistent. **(c)** The number of Vevo videos uploaded each year, broken down by genre.

5.3 Macroscopic measures

We first compute several basic statistics such as indegree distribution, view count distribution, and Vevo videos uploading trend (Section 5.3.1). Next, we study the connection between the network structure and video popularity (Section 5.3.2). Lastly, we use the bow-tie structure to characterize the Vevo network and we discuss the impact of different cutoff values (Section 5.3.3).

5.3.1 Basic statistics

Over-represented medium-size indegree videos. Here we study the indegree distribution of the Vevo network. Note that the outdegree of all nodes is bounded by the cutoff value on the relevant list and therefore not presented. We remove all links pointing to non-Vevo videos, resulting in an average of 363,965 edges each day, and an average degree of 6. Note that the average degree of 10 mentioned in Section 5.2 is obtained with a cutoff of 25, whereas here we study the relevant network constructed with a cutoff of 15, since the display probability of videos below position 15 appearing on recommended list is less than 0.32. Figure 5.4(a) shows the complementary cumulative density function (CCDF) of the indegree distribution for four different snapshots of the network, taken 15 days apart. We notice that the indegree distribution does not resemble a straight line in the log-log plot, meaning it is not power-law, unlike for other online networks, e.g., the World Wide Web [Broder et al., 2000; Meusel et al., 2014], the network of interaction in online communities [Zhang et al., 2007], and the follower/following network on social media [Kwak et al., 2010]. The medium-sized indegree videos are over represented than that in the best fitted power-law model ($\alpha = 2.02$, fitted by POWERLAW package [Alstott et al., 2014], resulting in $x^{-1.02}$ in CCDF [Clauset et al., 2009]). This result holds for all four snapshots.

Attention is unequally allocated. Figure 5.4(b) plots the average daily views against the view count percentile. The daily view count at median is 81, but it is 4,575 at the 90th percentile. These observations, together with a Gini coefficient of 0.946, indicate that the attention allocation in the Vevo network is highly unequal — the top 10% most viewed videos occupy 93.1% views. We also find a moderate correlation between view count and indegree value (details in Section 5.4).

Uploading trend by music genres. To date, our dataset is the largest digital trace of Vevo artists on YouTube, allowing us to study the production dynamics of the Vevo platform. Figure 5.4(c) shows the number of Vevo videos that are uploaded each year from 2009 to 2017, broken down by their genres. We omit year 2018 as we only observed 8 months for it (until August). There is a significantly higher number of uploads (9,277) in 2009 as it is the year when Vevo was launched, and when many all-time favorite songs were syndicated to the YouTube platform. Pop, Rock, and Hip hop music are the top 3 genres, accounting for 62.85% of all uploads. The Vevo videos upload rate is more or less constant around 7,000 since 2013. The flattening production dynamics is somewhat surprising given the overall growth of YouTube [youbube.com, 2017].

5.3.2 Linking network structure and popularity

Here, we investigate the connection between the relevant network structure and video view counts. Specifically, we divide the videos in the VEVO MUSIC GRAPH dataset into four equal groups by computing the view count quartiles. Each group contains 15,185 videos. Next, we count the number of edges that originate and end in each pair of groups. Figure 5.5 represents the four groups together with the number of links between them. The "top 25%" group contains the top 25% most viewed videos, while the "bottom 25%" contains the 25% least viewed videos. The width of the arrows is scaled by the number of the edges between the videos placed in the two groups. One can conceptualize that the edges act as conduits for the attention to flow between different groups and their thickness indicate the probability that a random user jumps from one group to the other. We observe that all four groups have the most links pointing to the "top 25%" group. In fact, every group disproportionately points towards more popular groups than towards the less popular ones. This means

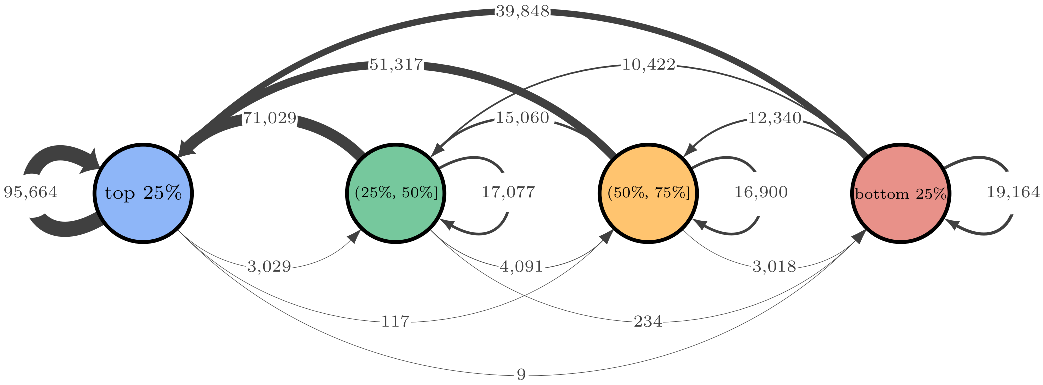


Figure 5.5: The four video groups are constructed based on view count percentiles and the connections between them. The arrow from group "(25%, 50%]" to group "top 25%" indicates the number of recommendation links from all videos in the second quartile to all videos in the top quartile. All groups disproportionately point to more popular groups.

the recommendation network built by the platform is likely to take a random viewer towards more popular videos and keep them there, therefore reinforcing the "rich get richer" phenomenon.

5.3.3 The bow-tie structure

The bow-tie structure was first proposed by Broder et al. [2000] to visualize the structure of the whole web. It classifies the complex web graph into five components: (a) the largest strongly connected component (LSCC) as the core; (b) the IN component which can reach the LSCC, but not the other way around; (c) the OUT component which can be reached from the LSCC, but not the other way around; (d) the Tendrils component which connect to either the IN or the OUT, bypassing the LSCC; (e) the Disconnected components which are disconnected from the rest of the components. The strongly connected component (SCC) can be easily computed in linear time by using Tarjan's algorithm [Tarjan, 1972]. For the Vevo network, we quantify the sizes of different components in the bow-tie structure using both the number of nodes (videos) and the amount of attention (views). Unlike the Web graph [Broder et al., 2000], we know the amount of views garnered by each video, this allows us to comparatively analyze the total attracted attention in each component. The bow-tie structure is a good conceptual description, because the directed edges exist only from the IN to the LSCC component (similarly, LSCC to OUT, and IN to OUT) but not the other way around, indicating that the attention in the network can only flow in a single direction from IN to LSCC (similarly, LSCC to OUT, and IN to OUT).

Component	LSCC	IN	OUT	Tendrils	Disconnected
Web '97 [Broder et al., 2000]	27.74%	21.29%	21.21%	21.52%	8.24%
Web '12 [Meusel et al., 2014]	51.28%	31.96%	6.05%	4.87%	5.84%
Forum [Zhang et al., 2007]	12.3%	54.9%	13.0%	17.9%	1.9%
Citation [Kim et al., 2012]	4.26%	54.93%	3.74%	24.76%	12.30%
Vevo videos	23.11%	68.54%	0.35%	2.47%	5.53%
Vevo attention	82.60%	12.74%	3.40%	1.13%	0.14%

Table 5.1: Comparison of bow-tie structure in prior studies and Vevo network.

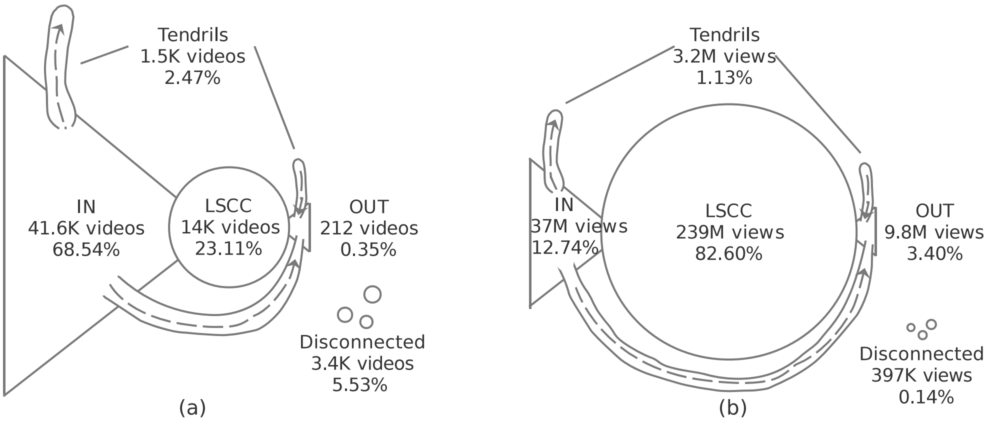


Figure 5.6: (a) The bow-tie structure of the Vevo network, using a snapshot on Oct, 1, 2018 and a cutoff of 15 on the relevant list. **(b)** The bow-tie structure, with each component resized by its corresponding view counts. The LSCC consumes the majority of attention in the Vevo network.

Table 5.1 compares the relative sizes of each component in prior literature and in our Vevo network. The Vevo network is quite different with respect to other previously studied online networks, e.g., the Web graph [Broder et al., 2000; Meusel et al., 2014] and user activity network in online community [Zhang et al., 2007; Kim et al., 2012]. It has a much larger IN component, encompassing 68.54% of all the videos. The OUT, Tendrils, and Disconnected components are all very small, accounting for a total of 8.35% videos. Figure 5.6(a) visualizes the bow-tie structure of the Vevo network. Unlike other graphs, our Vevo graph is the by-product of the recommender systems, which is subjected to the proprietary algorithm and its updating cycle. This suggests there may exist considerable temporal variation in the composition of the bow-tie components, see Section 5.5 for observations over time.

Figure 5.6(b) resizes each component of the Vevo bow-tie by the total view counts in it. Visibly the roles of LSCC and IN are reversed: the LSCC now occupies 82.6% attention (while accounting for only 23.11% of the videos), while the big IN component (68.54% of the videos) only attract 12.74% attention. This is consistent with

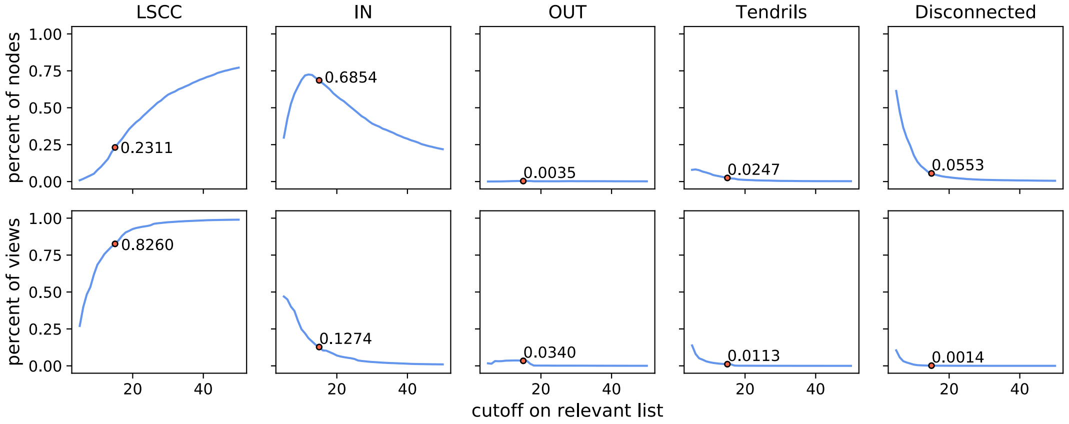


Figure 5.7: The relative size of the components in bow-tie structure, as a function of the cutoff on the relevant list. Red dots denote the statistics at the cutoff of 15.

the observation in Section 5.3.2 that the attention is unequally allocated in the Vevo network. Given the definition of the IN component, its 68.54% of videos contribute attention towards the LSCC, but not the other way around (there is no link from LSCC towards IN). As a result, the LSCC accumulates a large proportion of all attention. The OUT, Tendrils, and Disconnected components account for almost negligible attention (4.67% of the views altogether).

Impact of different cutoff values on the bow-tie structure. The Vevo network changes as we change the cutoff on the relevant list, as taking more edges into account densifies the network. Figure 5.7 shows how the relative size of the bow-tie component changes with varying cutoff values. As the cutoff increases, more edges are added to the network, especially for the videos in the Disconnected component. Backwards links are formed between videos in the LSCC and IN, and as a result, the LSCC absorbs parts of the IN component. Therefore, the LSCC increases, the IN decreases, while the other three components (OUT, Tendrils, and Disconnected) become negligible. At cutoff of 50, the Vevo network structures into 2 distinct components: a LSCC component consisting of 77% videos and 99% attention, and an IN component consisting of the remaining 23% videos and accounting for only 1% of the attention.

5.4 Microscopic measures

In this section, we jointly analyze the relation between video age, indegree, and popularity by examining overall correlation, as well as among top-ranked videos.

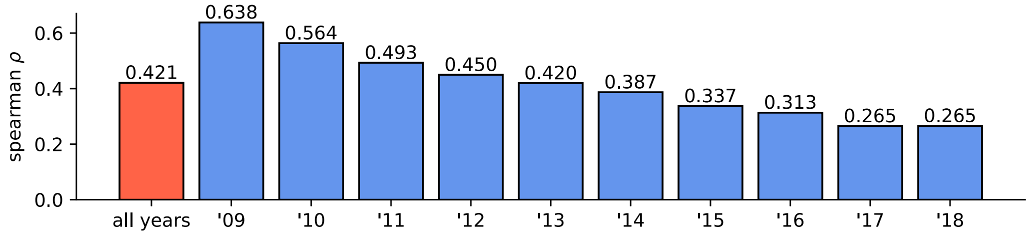


Figure 5.8: Spearman's ρ between video indegree and views, across all videos or disaggregated by uploaded year.

5.4.1 The disconnect between network indegree and video view count

We measure the correlation between video indegree and view count using Spearman's ρ – a measure of the strength of the association between two ranked variables, and which takes values between -1 and +1. A positive ρ implies the ranks of the two variables move together in the same direction. At the level of the entire dataset, we detect a moderate correlation between video indegree and view count (Spearman's $\rho = 0.421^{***}$, $p < 0.001$). Figure 5.8 shows the Spearman's ρ when we further break down the videos in the Vevo Music Graph based on their uploaded year. We observe that the strength of the correlation decreases for fresher videos. Videos uploaded in 2009 have a much stronger correlation ($\rho = 0.638^{***}$) than videos uploaded in 2018 ($\rho = 0.265^{***}$). This suggests that video age is an important confounding factor when one tries to estimate the effects of the recommendation network. Empirically, this may indicate the shift in what drives attention towards video consumption. Zhou et al. [2010] have measured that the two main drivers for video views are YouTube search and recommender. One explanation of our observation above is that as videos get older, the effects of recommendation become more pronounced.

5.4.2 A closer look at the top videos

Table 5.2 presents the top 20 videos with highest average daily indegree (top panel) and top 20 videos with highest average daily views (bottom panel). We observe a modest amount of discrepancy between these two dimensions, with only 5 videos being on both lists (shown in bold font). Most of the top-viewed videos are relatively new to the platform – 10 out 20 are published within one year and the top 5 are all within the past 7 months (relative to November 2018). In contrast, the videos with high indegree are mostly songs with sustained interests, some dating back to 10 years ago, such as "The Cranberries - Zombie" and "Bon Jovi - It's My Life". These two songs were respectively released in 1993 and 2000, having existed for a long time before being uploaded to YouTube. Currently, they still attract half a million views everyday after nearly 20 years, ranking 3rd and 17th on the most-linked video list, respectively. This may shed light onto why video popularity lifecycle exhibits a multi-phase pattern [Yu et al., 2015]. Our observations do not conflict with the design of YouTube recommender systems, which promote "reasonably recent and fresh" content [Davidson et al., 2010; Covington et al., 2016; Beutel et al., 2018]. Fresh videos can be recommended due to the relevance, novelty and diversity trade-offs [Konstan and Riedl, 2012; Ziegler et al., 2005]. Instead, our observed video relations are based on the content recommendation network [Carmi et al., 2017; Dhar et al., 2014].

Another group of interest is the videos that are highly viewed yet with low indegree. We find this pattern appears at the level of the artist. For instance, "Becky G" has 3 videos on the top 20 most-viewed list, ranking 2, 4, and 14. However, the indegrees for her videos are extremely low (rank 2411, 40040, and 958 respectively). Particularly, the video "Cuando Te Bese" attracts an average of 2.4M views every day for 9 consecutive weeks. However, it has only one video pointing to it from the rest of the 60,739 Vevo videos. A closer look reveals that "Becky G" is an American singer who often releases Spanish songs. The above observation shows that her videos are either recommended from non-English and/or non-Vevo videos, e.g., the Spanish songs community, or that recommendation network is not the main traffic driver for her videos.

Video title	Artist	Age	Indegree	-rank	Views	-rank
Girls Like You	**Maroon 5**	**155**	**870**	**1**	**7,167,077**	**1**
Rolling in the Deep	Adele	2,894	835	2	703,495	42
Zombie	The Cranberries	3,426	769	3	580,928	66
Something Just Like This	**Chainsmokers**	**618**	**732**	**4**	**1,840,077**	**6**
Counting Stars	**OneRepublic**	**1,981**	**714**	**5**	**1,632,001**	**9**
Uptown Funks	**Mark Ronson**	**1,444**	**587**	**6**	**1,724,938**	**7**
Here Without You	3 Doors Down	3,315	541	7	401,989	114
Someone Like You	Adele	2,591	514	8	954,981	26
Mr. Brightside	The Killers	3,426	509	9	197,313	255
The Pretender	Foo Fighters	3,317	480	10	266,610	182
I Want It That Way	Backstreet Boys	3,296	464	11	368,859	131
Unforgettable	French Montana	587	463	12	682,396	44
Dusk Till Dawn	**ZAYN**	**421**	**452**	**13**	**1,011,255**	**20**
Starboy	The Weeknd	765	450	14	519,343	77
Hello	Adele	1,106	447	15	518,429	78
Love The Way You Lie	Eminem	3,011	433	16	729,344	37
It's My Life	Bon Jovi	3,425	426	17	470,175	91
Cups	Anna Kendrick	2,030	419	18	140,614	386
Say You Won't Let Go	James Arthur	784	417	19	774,130	32
Pumped up Kicks	Foster The People	2,827	408	20	420,350	107
Girls Like You	**Maroon 5**	**155**	**870**	**1**	**7,167,077**	**1**
Sin Pijama	Becky G	196	28	2,411	3,988,681	2
Taste	Tyga	170	242	81	2,542,673	3
Cuando Te Bese	Becky G	92	0	40,040	2,373,613	4
Rise	Jonas Blue	140	40	1,603	1,937,467	5
Something Just Like This	**Chainsmokers**	**618**	**732**	**4**	**1,840,077**	**6**
Uptown Funks	**Mark Ronson**	**1,444**	**587**	**6**	**1,724,938**	**7**
No Tears Left to Cry	Ariana Grande	196	84	597	1,634,916	8
Counting Stars	**OneRepublic**	**1,981**	**714**	**5**	**1,632,001**	**9**
Thunder	Imagine Dragons	548	72	764	1,474,712	10
One Kiss	Calvin Harris	184	23	2,973	1,230,173	11
Natural	Imagine Dragons	70	8	7,747	1,186,965	12
Believer	Imagine Dragons	605	60	998	1,174,431	13
Mayores	Becky G	476	62	958	1,173,191	14
What's Up	4 Non Blondes	2,809	355	36	1,128,159	15
Sugar	Maroon 5	1,388	349	37	1,116,300	16
God's Plan	Drake	258	227	102	1,093,048	17
Sicko Mode	Travis Scott	91	64	913	1,076,694	18
Whatever It Takes	Imagine Dragons	386	118	347	1,016,394	19
Dusk Till Dawn	**ZAYN**	**421**	**452**	**13**	**1,011,255**	**20**

Table 5.2: Top 20 most-linked (top panel) and top 20 most-viewed videos (bottom panel). Both the indegree and the view counts are the average of daily values across 9 weeks. The age (in days) is calculated till Nov 2, 2018. Only 5 videos appear in both charts (**boldfaced**). Most high indegree videos are songs with sustained interests, whereas most highly viewed videos are recently uploaded.

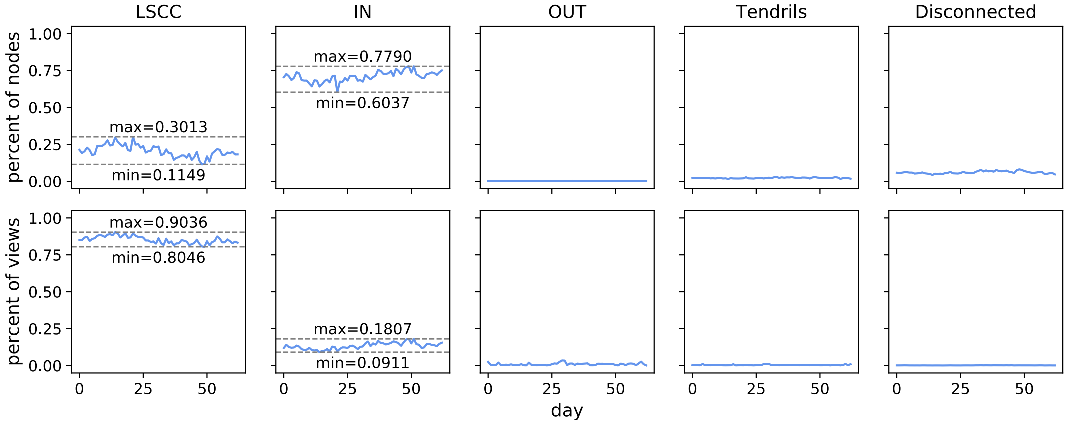


Figure 5.9: Temporal evolution of the bow-tie structure over 63 days.

5.5 Temporal patterns

Here, we study the dynamics of the Vevo network over 9 weeks, namely the appearance and disappearance of recommendation links between videos. We show that pairs of videos can have either ephemeral link or frequent link between them.

Macroscopic dynamics. Figure 5.4(a) and (b) show that both the indegree distribution and the view count distribution are temporally consistent. However, when we plot the size variation of the different components in the bow-tie structure, we obtain a more nuanced story. Figure 5.9 shows that the size of the LSCC ranges from 11.49% to 30.13%, while IN component from 60.37% to 77.9% over 9 weeks. Similarly, the percentage of total views in the LSCC ranges from 80.46% to 90.36%, while IN component from 9.11% to 18.07%. Given that the same set of videos is tracked throughout the observation period and no new video is added, the above observations imply a significant turnover in the recommendation links between videos. For example, the appearance of a link will allow a node to transition from the IN to the LSCC component; the disappearance of the same link would make it drop back into IN component.

Incoming ego-network dynamics. We study the link turnover using the incoming ego-network for each video. Ego network consists of an individual focal node and the edges pointed towards it. We only consider incoming edges, as the number of outgoing edges is capped by the relevant list cutoff (here the cutoff is 15). For each video, we first extract the days with at least 20 incoming links. Then for each day t, we compute the indegree change ratio between day t and day $t+1$ by dividing the indegree delta (positive or negative) by the value in day t. We obtain a number between -1 and 1, where -1 means that the video loses all of its incoming edges,

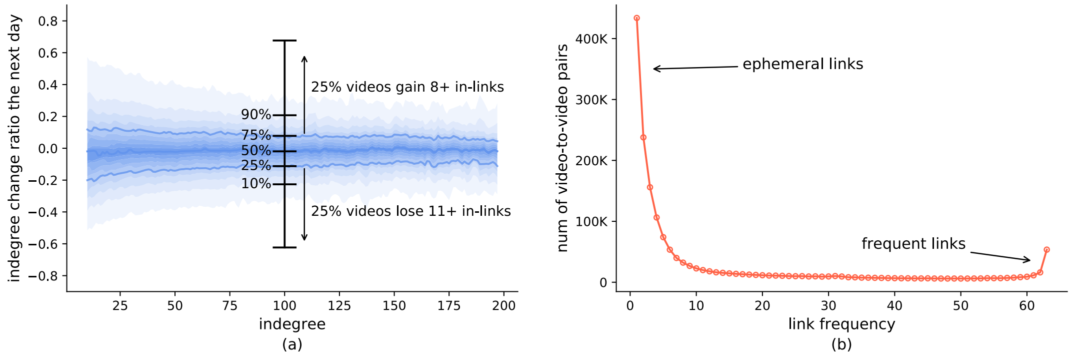


Figure 5.10: (**a**) Daily indegree change rate for videos that have at least 20 in-links. **(b)** Link frequency of video-to-video pairs. 434K (25.2%) links appear only once while 54K (3.1%) links appear every day of our 9-week observation windows.

and a value of 1 signifies that the video doubles the number of incoming edges. Figure 5.10(a) shows the indegree change ratio summarized as quantiles, broken down by the value of indegree. We highlight the 10th, 25th, median, 75th, and 90th percentile for the videos with an indegree of 100. 25% videos with an indegree of 100 will gain at least 8 in-links on the next day while another 25% lose at least 11 in-links. The median is around zero, meaning that there are as many videos that gain links as these that lose links. Overall, this suggests that videos have very dynamic incoming ego-networks, with a non-trivial number of edges prone to appear and disappear from one day to another.

Ephemeral links and frequent links. Given the rate at which links appear and disappear, here we ask the question if there exist videos that are frequently connected. For each pair of connected videos, we count the number of times that a link appears between them over the 63 daily snapshots. Figure 5.10(b) plots the link frequency (taking values between 1 and 63) on the x-axis and the number of video-to-video pairs with that link frequency on the y-axis. We find that many links are ephemeral – they appear several times, scattering in the 63 days time window. We count that 434K (25.2%) video-to-video links only appear once. On the other hand, there are links that appear in every snapshot — we count 54K (3.1%) such links. Ephemeral links may contribute to bursty popularity dynamics of YouTube videos, and to the generally perceived unpredictability in complex social systems [Martin et al., 2016; Rizoiu et al., 2017b, 2018]. Frequent links may hold the answer to understanding and predicting the attention flow in a network of content.

5.6 Estimating attention flow in recommendation network

The goal of this section is to estimate how well can the view counts of a video v at day t (denoted by $\mathbf{y}_v[t]$) be predicted, given (1) the view series of v in the past w days, $\mathbf{y}_v[t-w], \dots \mathbf{y}_v[t-1]$; (2) the view series, $\mathbf{y}_u[t-w], \dots \mathbf{y}_u[t]$, for the set of videos $\{u | (u \to v) \in G\}$ pointing to v.

To this end, we first define and extract a persistent network that contains links appearing throughout all the snapshots (Section 5.6.1). Next, we detail the setup of predicting video popularity with recommendation network information (Section 5.6.3). We analyze the prediction results and provide an analysis on the strength of each link (Section 5.6.5). Finally, we introduce a new metric – estimated network contribution ratio. We use it to identify the types of content that benefit most from being recommended in the network (Section 5.6.5).

5.6.1 Constructing a network with persistent links

In order to reliably estimate the effects of the recommendation network on the viewing behaviors, we apply two filters: (a) target videos should have at least 100 daily views on average; (b) the average daily views of the source videos should be at least 1% of those of the target videos as such videos cannot substantially influence their far more popular neighbors. One can adjust the filtering criteria – a lower threshold will attribute more variances to the less popular videos while a higher threshold focuses on the flows between more popular videos. In the resulting network, we further remove the *ephemeral links* that appear sporadically over time and correct for the *missing links* that appear frequently, but with scattered gaps in between their appearances. We assume that the missing links are likely to exist in the scattered gaps, and we use a majority smoothing method to find them (detailed next). Links appearing in all the 63 daily snapshots and the corrected missing links, both dubbed *persistent links*, make up the *persistent network*.

Finding persistent links. We use a moving window of length 7, same as the weekly seasonality, to extract the persistent structure of the Vevo network over the 63-day observation window. A link from video u to video v, $(u \to v)$, is maintained on day t if $(u \to v)$ appears in a majority (≥ 4) of the days in time window $[t-3, t+3]$. Likewise, if a link is missing on the current day t but it appears in the majority of surrounding 7-day window, we consider it is a missing link and add it back to the network. When $t-3$ is earlier than the first day of data collection, or $t+3$ later than the last day, we still apply the majority rule on the available days. The resulting graph has 52,758 directed links, pointing from 28,657 source videos to 13,710 target videos. Among them, 2,696 links are reciprocal, meaning two videos mutually recommend

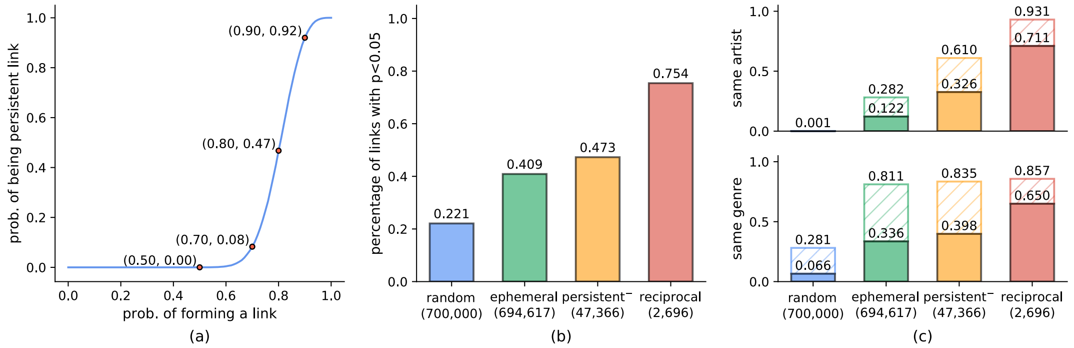


Figure 5.11: **(a)** The probability of forming a persistent link (y-axis) as a function of the probability of forming a link (x-axis). **(b)** Fraction of statistically correlated links in four groups at significance level of 0.05. The numbers in the brackets indicate the number of links in each group. **(c)** The numbers above the shaded bar indicate the fraction of links between videos from the same artist (top) or with the same genre (bottom). The numbers above the solid bar indicate the fraction of links connecting videos with the same artist/genre and whose popularity dynamics are statistically correlated at significance level of 0.05.

each other. We find significant homophily in the persistent network: 33,908 (64.3%) links have both the source and the target videos belonging to the same Vevo artist, and 44,154 (83.7%) links are between videos of the same music genre.

Validating persistent links via simulation. We illustrate the probability of persistent links by simulating a simple link presence/absence model. We assume a link is independently presented on each day with probability $p_l \in [0, 1]$, and absent with probability $1 - p_l$. We first simulate the link formulation for 63 times, then apply our 7-day majority smoothing to determine if it is persistent. We repeat the simulation for 100,000 times, and compute the probability of a link being persistent, denoted by ξ. In Figure 5.11(a), we plot the obtained ξ against varying p_l. For $p_l = 0.5$ the edge is never persistent ($\xi = 0$), whereas for $p_l = 0.9$ the edge is very likely to be persistent ($\xi = 0.92$). From the simulation results, we can see that our 7-day majority smoothing rule favors links that appear much more frequent than chance, and suppresses links that appear lower or closer to chance.

Videos connected by persistent links have correlated popularity dynamics. We use Pearson's r to measure the correlation between the popularity dynamics of two videos connected by a persistent link. It is known that the cross-correlation of time series data is affected by the within-series dependence. Therefore, we deseasonalize, detrend, and normalize the view count series by following the benchmark steps in the M4 forecasting competition [Competition, 2018]. This is to ensure that the residual time series data is stationary and to avoid spurious correlations. We compute the Pearson's r on the obtained residual data, and we perform a paired correlation test which we consider statistically significant for $p < 0.05$.

Figure 5.11(b) shows the fraction of links for which the correlation test is statistically significant over four groups of links. The *persistent*$^{-}$ group contains all the 52,758 persistent links we identified but excluding the 2,696 pairs of *reciprocal* links — resulting in 47,366 persistent yet non-reciprocal links. The *ephemeral* group consists of all links which have been deemed as non-persistent after applying the 7-day majority smoothing. The *random* group is constructed by randomly selecting pairs of unconnected videos and pretending that they have a link. All groups are filtered based on the same two criteria mentioned before. There are a total of 694,617 links in the ephemeral group and we sample 700,000 links in the random group. We find that 75.4% of the reciprocal links connect videos with statistically correlated popularity series. We include both positive and negative correlations as two user attention series may cooperate or compete with each other [Zarezade et al., 2017]. Combining the reciprocal and persistent^{-} groups, 26,460 (50.2%) links in our persistent network have correlated dynamics. This is much higher than the percentage for ephemeral links (40.9%) and that for unconnected random video pairs (22.1%).

We further examine the content similarity in the persistent links by grouping links that connect videos from the same artist or with the same music genre (described in Figure 5.4(c)). Figure 5.11(c) top shows that most reciprocal links (93.1%) connect videos from the same artist, while 71.1% of them have statistically correlated popularity dynamics. The percentages are slightly lower for the persistent^{-} group (61% from the same artist, and 32.6% with correlated popularity) and it drops even lower for ephemeral group (28.2% and 12.2%, respectively). The situation is slightly different when we study the links that connect videos of the same genre, as shown in Figure 5.11(c) bottom. We find that more than 80% of the links connect videos of the same genre, irrespective of whether they are sporadically or persistently connected. The percentages of statistically correlated links with the same genre follow the same trend as those from the same artist, i.e., highest for reciprocal (65%), followed by persistent^{-} (39.8%), ephemeral (33.6%) and lowest for random (6.6%). The above observations indicate that not all persistent links have the same effect on video popularity, and motivate us to build a prediction model for each of the links.

5.6.2 Problem statement

One important observation is that viewing dynamics exhibit a 7-day seasonality [Huang et al., 2018; Cheng et al., 2008]. In our temporal hold-out setting, we use the first 8 weeks (2018-09-01 to 2018-10-26) to train the model and we predict the daily view counts in the last week (2018-10-27 to 2018-11-02). This chronological split ensures that the training data temporally precedes the testing data. If at any point we are required to use the day $t+1$ to predict the day $t+2$ (when both $t+1$ and $t+2$ are in the testing period), we use the predicted value $\hat{\mathbf{y}}[t+1]$ instead of observed value $\mathbf{y}[t+1]$.

5.6.3 Experimental setup

Evaluation metric. The predicting performance is quantified using the symmetric mean absolute percentage error (SMAPE). SMAPE is an alternative to the mean absolute percentage error (MAPE) that can handle the case when the true value or the predicted value is zero. It is a scale-independent metric and suitable for our task in which the volume of views for different videos vary considerably. Formally, SMAPE can be defined as

$$\text{SMAPE}(v) = \frac{200}{\mathbf{T}} \sum_{t=1}^{\mathbf{T}} \frac{|\mathbf{y}_v[t] - \hat{\mathbf{y}}_v[t]|}{|\mathbf{y}_v[t]| + |\hat{\mathbf{y}}_v[t]|} \quad \text{or} \quad \text{SMAPE}(t) = \frac{200}{|G|} \sum_{v \in G} \frac{|\mathbf{y}_v[t] - \hat{\mathbf{y}}_v[t]|}{|\mathbf{y}_v[t]| + |\hat{\mathbf{y}}_v[t]|} \tag{5.1}$$

where $\mathbf{y}_v[t]$ is the true value for video v on day t, $\hat{\mathbf{y}}_v[t]$ is the predicted value, T is maximal forecast horizon, and G is the persistent network. SMAPE(v) averages the forecast errors over different horizons for an individual video v, while SMAPE(t) averages over different videos for a certain forecast horizon t. The overall SMAPE for each model is computed by taking the arithmetic mean of SMAPEs over different horizons and over all videos. SMAPE ranges from 0 to 200, while 0 indicates perfect prediction and 200 the largest error, when one of the true or the predicted values is 0. When the true and the predicted are both 0, we define SMAPE to be 0.

5.6.4 Methods

Baseline models. We use a few off-the-shelf time series forecasting methods from naive forecast to recurrent neural network. The baseline models are estimated on a per-video basis.

- Naive: The forecast at all future times is the last known observation.

  $$\hat{\mathbf{y}}_v[t] = \mathbf{y}_v[\mathrm{T}^*] \tag{5.2}$$

  where T* is the last day in the training phase.

- Seasonal naive (SN): The forecast is the corresponding observation in the last seasonal cycle. This method often works well for seasonal data. We observe that many videos in the VEVO MUSIC GRAPH dataset exhibit a 7-day seasonality. Therefore we set the periodicity length m* to be 7.

$$\hat{\mathbf{y}}_v[t] = \mathbf{y}_v[t - \mathrm{m}^*] \tag{5.3}$$

- Autogressive (AR): AR is one of the most commonly used model in time series forecasting. An AR model of order p describes the relation between each of the past p days and current day, formally defined as:

  $$\hat{\mathbf{y}}_v[t] = \sum_{\tau=1}^{p} \alpha_{v,\tau} \mathbf{y}_v[t - \tau] \tag{5.4}$$

  We choose the order p to be 7. $\alpha_{v,\tau}$ represents the relation between current day and τ days before.

- Recurrent neural network (RNN): RNN is a deep learning architecture that models temporal sequences. We implement RNN with long short-term memory (LSTM) units. LSTM-based approaches have been competitive in time series forecast tasks, mainly in a sequence-to-sequence (seq2seq) setup, see [Kuznetsov and Mariet, 2019] for detailed discussions.

Networked popularity model. Built on top of the AR model, we model the network effects by assigning a weight $\beta_{u,v}$ to each link $(u \to v)$ existing in the persistent graph G, which modulates the inbound traffic received via that link, defined as:

$$\hat{\mathbf{y}}_v[t] = \sum_{\tau=1}^{p} \alpha_{v,\tau} \mathbf{y}_v[t - \tau] + \sum_{(u,v) \in G} \beta_{u,v} \mathbf{y}_u[t] \tag{5.5}$$

$\beta_{u,v}$ can be explained as the probability that a generic user clicks on video v from video u, therefore, we impose the constraint $0 \leq \beta_{u,v} \leq 1$. We refer to this model as ARNet.

One way to interpret the ARNet is to conceptualize a YouTube watching session as a sequence of video clicking. We therefore categorize views on YouTube into two classes: *initial* views and *subsequent* views. The initial views start the clicking sequences. Some possible entry points include homepage feed, search results, or YouTube URLs on other social media. The subsequent views model the behaviors of users clicking by following the recommendation links. The session ends when the user navigates back to YouTube homepage, or quits the browser. Although in the dataset we cannot differentiate initial views from subsequent views, we consider that initial views are driven by the latent interest of users, modelled as autoregression of the past p days; in contrast, subsequent views are directed by the recommendation network, modelled as contribution from its incoming neighbours $\{u|(u \rightarrow v) \in G\}$ and mediated by estimated link strength $\beta_{u,v}$.

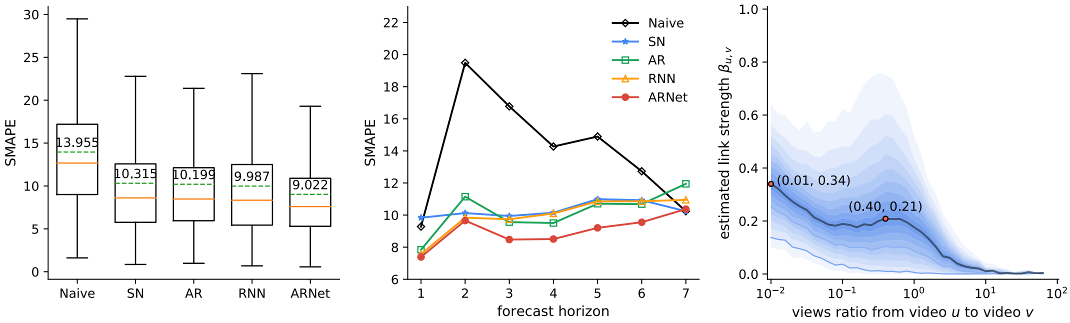
We use the STATSMODELS.TSA package for the AR model, KERAS package for the RNN, and build a customized optimization task with constrained L-BFGS for the ARNet. We use the SMAPE as objective function in both RNN and ARNet.

5.6.5 Results and analysis

Figure 5.12(a) summarizes the prediction errors achieved by the five methods defined in Section 5.6.3. The Naive model alone is a weak predictor, however accounting for the seasonal effects (SN model) yields a significant error decrease. It is worth noticing that the AR model yields similar performance as the advanced RNN model — due to the known result that future popularity of online videos correlates with their past popularity [Pinto et al., 2013]. We observe that using recommendation network information further improves the prediction performance: the ARNet model achieves a 9.66% relative error reduction compared to the RNN model[7]. This prediction task shows that one can better predict the view series for a video if the list of videos pointing to it is known. Next we study the prediction performance with respect to the forecast horizon, i.e., how many days in advance do we predict. We average the SMAPEs over all videos against predictions for a given forecast horizon t, computed as SMAPE(t) in Equation (5.1). Figure 5.12(b) shows a nuanced story: the prediction performances decrease for all models as the forecast horizon extends. Nevertheless, the ARNet model consistently outperforms other baselines across all forecast horizons, especially for larger horizons.

We posit two factors in preventing the models from obtaining even better results. Firstly, it is well known that the attention dynamics tend to be bursty when items are first uploaded [Rizoiu and Xie, 2017; Cheng et al., 2016; Martin et al., 2016], and the interest dissipates with time [Figueiredo et al., 2016]. Given that 56,845 (93.6%) videos in our dataset have been uploaded for more than one year and 9,277 (15.3%) videos for almost ten years, most of the videos have passed the phases of the initial attention burst. As a result, a large part of popularity variation comes from the weekly seasonality, rendering the simple seasonal naive model particularly competitive when compared to the more advanced RNN method. The second is data sparsity when we build the models on a per-video basis. RNN works best when it has ample volumes of data to train. However, we use a sliding 7-day windows to predict the views in the next 7 days as suggested in [Kuznetsov and Mariet, 2019],

[7] In a follow-up study from our lab, Tran et al. [2021] have showed a further performance improvement by using a multi-head attention model.



(a) (b) (c)

Figure 5.12: Summary of prediction results, SMAPE: lower is better. **(a)** Boxplots aggregate the prediction performances over the 13,710 videos in the test set. The dotted green line and the values show the mean SMAPE. **(b)** SMAPE for different forecast horizons (in days). **(c)** The distribution of estimated link strength $\beta_{u,v}$ (y-axis) against the ratio of views of source video to that of target video (x-axis, in log scale). It has a bi-modal shape.

therefore our data size is limiting the effectively training of the RNN model.

In our ARNet model, the estimated link strength $\beta_{u,v}$ can be used to quantify the influence from a video to its neighbours. In Figure 5.12(c), we plot the distribution of $\beta_{u,v}$ against the ratio of views of source video to that of target video. We split the x-axis into 40 equally wide bins in log scale. Within each bin, we compute the values at each percentile, and then connect the same percentile across all bins. The median line is highlighted in black. The lighter the color shades are, the further the corresponding percentiles are away from the median. We observe the distribution has a bi-modal shape with the first mode in 0.01 and second in 0.40 (for the median), meaning users are more likely to click a much more popular video (100 times more popular), or a moderate more popular video (2.5 times). In contrast, the estimated link strength towards a less popular video is very low. This observation, together with the measurement that videos disproportionately point to more popular videos (Section 5.3.2), further reinforces the "rich get richer" phenomenon.

The impacts of network on video popularity prediction.

From the ARNet model, we derive a metric called the estimated network contribution ratio η_v, which is defined as

$$\eta_v = \frac{\sum_{t=1}^{\mathbf{T}} \sum_{(u,v)\in G} \beta_{u,v} \mathbf{y}_u[t]}{\sum_{t=1}^{\mathbf{T}} \hat{\mathbf{y}}_v[t]} \tag{5.6}$$

η_v is the fraction of estimated inbound traffic from video v's neighbours against its own predicted popularity. As we constrain all coefficients in Equation (5.5) to be non-negative, η_v is bounded in $[0, 1]$. In our dataset, the mean η_v is 0.314. In other words, for an average video in the Vevo Music Graph dataset, 31.4% of its views

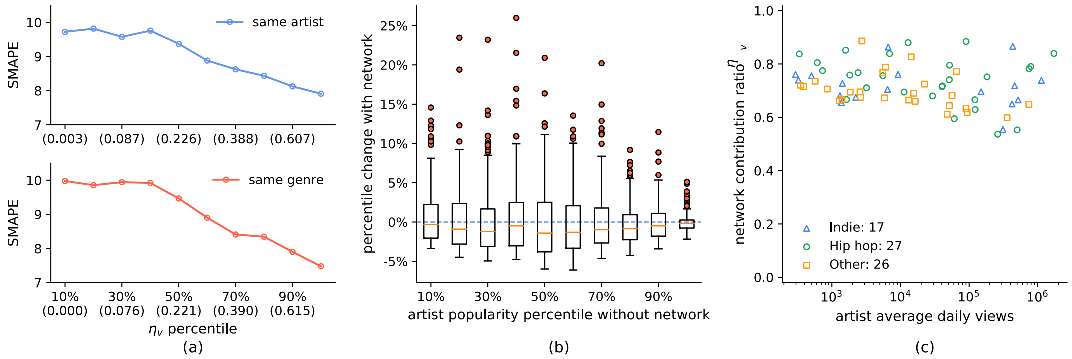


Figure 5.13: (a) SMAPE as a function of the network contribution η_v from videos with the same artist (top) or the same genre (bottom). We use the η_v percentile as x-axis. The numbers within the brackets indicate split values for each percentile, e.g., the right-most dots indicate top 10% videos with the highest percent of views from similar content, having η_v larger than 0.607 for the same artist (top) or 0.615 for the same genre (bottom). **(b)** Boxplot of artists' popularity percentile changes when adding the recommendation network. x-axis: popularity percentile if removing the network; y-axis: popularity percentile change with network. The outliers (red circles) denote the artists who gain the most popularity through the network among their cohort. **(c)** A closer look of artists identified in **(b)**. A group of Hip hop artists and Indie artists rely more on the recommendation network to become popular.

are estimated from the recommendation network. This value is slightly higher than the YouTube network contribution measured by Zhou et al. [2010] in 2010 (reported below 30%). We posit two potential reasons: (1) the Vevo network is more tightly connected than a random YouTube video network [Airoldi et al., 2016]; (2) traffic on recommendation links may have increased since then, signifying the advances of modern recommender systems. Furthermore, among the 31.4% networked views, 85.9% are estimated from the same artist, echoing the network homogeneity found by Airoldi et al. [2016]. On average, the 13,710 target videos in the persistent network attract 245.3M views every day. Our ARNet model estimates that 78.6M (32%) of these views are contributed via the recommendation network.

Firstly, we explore the relation between prediction performance and content similarity concerning the artist and music genre. In Figure 5.13(a), we compute η_v conditioned on that $(u, v) \in G$ and that u and v are from the same artist (top) or with the same genre (bottom). We then slice the x-axis into 20 bins, 5 percentiles apart, based on the artist/genre network contribution ratio. We compute the mean SMAPEs for the videos in each bin. Videos that are connected solely by videos from other artists/genres will be placed in the leftmost bin ($\eta_v = 0$). The plot shows that the SMAPE error decreases with the increasing percentage of views from videos with the same artist or genre.

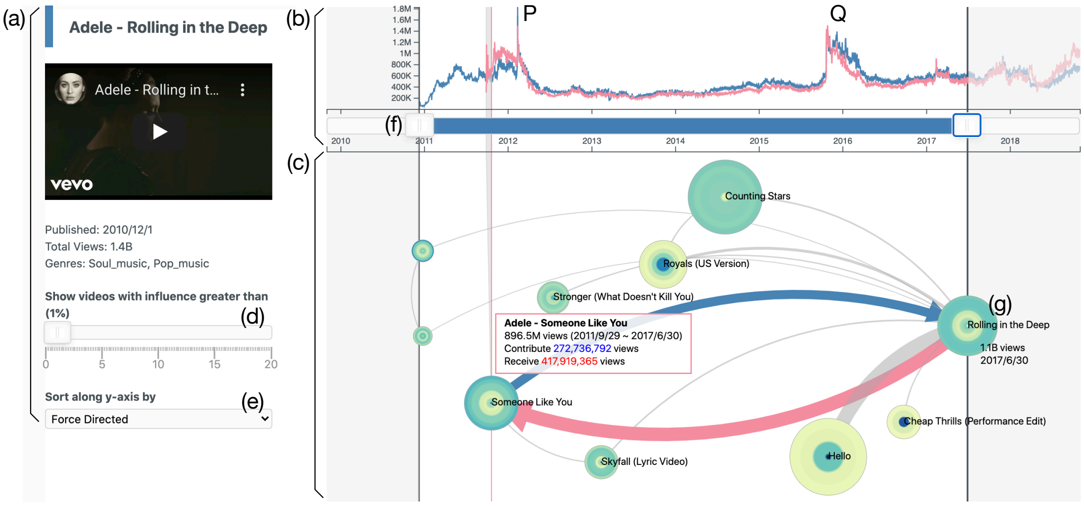
Secondly, we study the question that which artists are affected most *if* the recommender systems were to be turned off? Figure 5.13(b) shows the popularity percentile *change* at the level of artist. We first compute the network-subtracted views, i.e., subtracting the network contribution $\sum_{t=1}^{\mathbf{T}} \sum_{(u,v)\in G} \beta_{u,v} \mathbf{y}_u[t]$ from the observed views $\sum_{t=1}^{\mathbf{T}} \mathbf{y}_v[t]$. We then aggregate and compute the popularity percentiles for both observed views and network-subtracted views at the level of artist. The x-axis plots the artists' popularity percentiles without recommendation network, and y-axis plots the percentile changes when turning on the network. The range of percentiles stays constant between $[0, 100\%]$, reflecting the concept of finite attention — one video gains popularity at the expense of others. The top outliers identify artists who gain much more popularity than their peers with similar popularity due to the recommendation network; whereas the bottom outliers represent artists who lose popularity. There are 2,340 artists having target videos in the persistent network. We observe that 1,378 (58.89%) artists losing a small amount of popularity (less than 5%) while 948 (40.51%) gaining. We notice there is no bottom outlier. On the contrary, the top outliers show that the network can help some artists massively increase their relative popularity (as high as 26%, J-Kwon (American rapper) in 4th bin).

We take a closer look at the outliers by scattering them in Figure 5.13(c). 70 artists gain significant popularity from the recommendation network, implying a better utilization of network effects. We retrieve the artist genres from the music database MusicBrainz, and we notice two notable groups. One is the Indie group by matching genre keywords "indie", "alternative", or "new wave". The top 3 most popular Indie artists are 4 Non Blondes, Hoobastank, and The Police. The other is the Hip hop group by matching genre keywords "hip hop", "rap", "reggae", or "r&b". The top 3 most popular Hip hop artists are Mark Ronson, French Montana, and Pharrell Williams. This finding reveals that the recommender systems can lead users to find niche artists.

5.7 Visualizing attention flow in recommendation network

In this section, we present AttentionFlow, a new system to visualize the attention time series of videos in the YouTube recommendation network and the dynamic influence they have among each other. We choose the ego network as the main interface. An ego network consists of a focal node (the ego) and its direct neighbors (the alters). It emphasizes the changing relations and evolving influence between the ego and alters. More broadly, AttentionFlow can be generalized to visualize web traffic patterns, search trends, or movement of people and goods among urban hubs.

Centred around an ego video, our system progressively reveals its neighbors at



Adele - Rolling in the Deep

Published: 2010/12/1
Total Views: 1.4B
Genres: Soul_music, Pop_music

Show videos with influence greater than (1%)

Sort along y-axis by: Force Directed

Adele - Someone Like You
896.5M views (2011/9/29 ~ 2017/6/30)
Contribute 272,736,792 views
Receive 417,919,365 views

Rolling in the Deep
1.1B views
2017/6/30

Figure 5.14: AttentionFlow visualizes the temporal attention trend of a video and the dynamic attention flowing over its ego network. Focusing on an ego video (g), the metadata view (a) shows its descriptive information. The trend view (b) presents two attention series of the ego and the hovered alter node, while the network view (c) highlights the incoming and outgoing attention between them. Users can filter the alter nodes by choosing influence threshold (d) and select a sorting criterion for the vertical axis (e). The time slider (f) defines an observation window, in which the ego video (g) is always placed at the rightmost position. In this snapshot, we observe two spikes in the attention dynamics of the music video "Rolling in the Deep" by Adele. The first spike (P) is related to the Grammy Awards of that year, while the second (Q) is due to the release of Adele's new song "Hello".

the time of their significant influences, while simultaneously presenting the temporal pattern of each series using two visual encodings: a line chart for comparing with other nodes and a tree ring for summarizing temporal patterns. The ego network shares the same time axis with the line chart, where the position of an alter node represents the time when it starts to have influence on the ego node. The main interaction component is a time slider that allows users to select an observation window. As the window shifts, the ego network structure and influence flows change accordingly. We also provide controllers to hide less influential edges and sort nodes based on different criteria.

The frontend is rendered in D3.js and the backend uses the Neo4j graph database. On each page, users can access the visualization of an entity by searching for its name (e.g. the video name or the artist name), or by clicking on the corresponding node in the influence network of another entity. Figure 5.14 presents the main visualization layout of the video network page, which includes three components: a metadata view, a trend view, and a network view.

The metadata view Figure 5.14(a) shows the detailed attributes of an ego node. For example, video title, embedded snippet, creation time and genres are presented for a given music video. Below the description panel, two controllers can be used to alter the network layout. The influence slider (Figure 5.14d) is used to filter out alter nodes with influence less than a chosen threshold, defaulted at 1%. This threshold also determines the influencing time of the alters, which in turn decides the horizontal positions of the alter nodes in the network view. The drop-down box (Figure 5.14e) provides five criteria for sorting nodes along the vertical axis: force-directed (default), total views, incoming views, outgoing views, or artist names.

The trend view Figure 5.14(b) provides a line chart to visualize the time series of attention. A time slider (Figure 5.14f) is located on the horizontal axis to select an observation window. The left handle changes the start time of the observation, while the right handle changes the position of the ego. The periods outside the selected time range are greyed out.

The network view Figure 5.14(c) visualizes the ego network structure and influence flows. They change dynamically when users interact with the time slider and the influence slider. Alter nodes that have influence with the ego by more than the chosen threshold will appear in the network view. When hovering over an alter node, the edges between the alter and the ego are highlighted, the alter's line chart is revealed in the trend view, and a card containing influence information pops up. For each alter node, the horizontal position is related to the influence threshold and the vertical position is related to the sorting criteria.

5.8 Conclusion

This work presents a large-scale study for online videos on YouTube. We collect a new dataset that consists of 60,740 Vevo music videos, representing some of the most popular music clips and artists. We construct the YouTube recommendation network. We present measurements on the global component structure and temporal persistence of links. A model that leverages the network information for predicting video popularity is proposed, which achieves superior results over other baselines. It also allows us to estimate the amount of attention flow over each recommendation link. We derive a metric — estimated network contribution ratio, and we quantify this ratio at both the entire Vevo network level and individual artist level. We also develop a new system AttentionFlow to visualize a collection of video attention series and the dynamic network influence. To the best of our knowledge, this is the first work that links the video recommendation network structure to the attention consumption for the videos in it.

5.8.1 Discussion

Much progress has been made to algorithmically optimize or increase the attention for individual digital item (from videos to products to connections in social networks), whereas the theory about attention flow among different items is still fairly nascent. Our data includes a series of network snapshots that are constructed by the platform's recommender systems, and visible to both content producers and consumers. We believe that the area of understanding the implications of content recommendation networks has many worthy problems and fruitful applications. However, definitions and properties of a recommendation network that is fair and transparent to the content hosting site, producers and consumers remain as open issues.

5.8.2 Limitations and future work

Our limitations include: interpretations of importance are directly based on regression weights; some observations may not generalize to other digital items other than the most popular music videos; the prediction does not explore all the potential deep learning architecture and parameter tuning. Future work includes modeling attention flow that takes into account item rank on the relevant list; connecting aggregate attention with individual click streams; and improving deep neural network models, specifically, three directions for us to exploit. Firstly, extract additional features, such as audio-visual, artist, and network features. Secondly, measure the relations between estimated link strength and link properties, such as the diversity and/or novelty of the target video relative to the source video [Ziegler et al., 2005]. Lastly, train a shared RNN model on videos with similar dynamics for increasing the volume of training data [Figueiredo et al., 2016].

Chapter 6

Conclusion

In this chapter, we first summarize the main contributions of this thesis. Next, we discuss several possible future research directions.

6.1 Summary

The work in this thesis aims to understand the collective human behaviors in online platforms from the perspectives of social data sampling, user engagement patterns, and network effects of recommender systems. In particular, we focus on item-centric, quantitative analysis instead of user-centric, qualitative approach. Broadly, this thesis makes the following contributions:

- **Quantify sampling effects of online social data.** We present a comprehensive study of the Twitter sampling effects on common measurements in Chapter 3. We show that Twitter rate limit messages can estimate the volume of missing tweets accurately. Tweets sampling rates also vary across different timescales. While the hourly sampling rate is affected by the diurnal rhythm in different time zones, the millisecond level sampling is heavily influenced by the Twitter's implementation choices. We find the Bernoulli process with a uniform rate approximates the empirical entity distribution well. We also propose a new method to infer the true entity distribution and ranking based on sampled observations. In the user-hashtag bipartite graph and user-user retweet network, we observe that the network structures are altered with denser components more likely to be preserved. For the diffusion models, sampling compromises their quality because tweet inter-arrival time is significantly longer in the sampled stream, while user influence is lower.
- **Measuring and predicting engagement in online content.** We present the first large-scale measurement study on how users collectively engage with online content in Chapter 4. We study a set of metrics including time and percentage of videos being watched, and we observe that video duration is an important covari-

ate on watching patterns. To calibrate engagement measures against video length, we construct a new tool called engagement map to convert the watch percentage into a new metric – relative engagement. The relative engagement metric captures the watch percentage rank percentile among videos of similar lengths. We show relative engagement is closely correlated with recognized notions of content quality, stable over time and predictable before videos' upload. We extract features such as video content, topics, and channel reputation to predict watch percentage and relative engagement in a cold-start setup. And we can achieve coefficient of determination R^2 of 0.45 and 0.77, respectively. Lastly, we use a Hawkes process model to forecast the video viewing and watching dynamics. The result suggests the engagement metric (daily watch time) is more predictable than the popularity metric (daily view count).

- **Measuring and modeling content recommendation networks.** We present the first large-scale measurement study on the network effects induced by YouTube recommender systems in Chapter 5. The study is done on a new VEVO MUSIC GRAPH dataset, which contains the content recommendation network for 60,740 music videos. We discover the popularity bias that videos are disproportionately recommended towards more popular videos. This means the recommender system is likely to take a random viewer to more popular videos and keep them there. Furthermore, we use the bow-tie structure to characterize the recommendation network. We find that its core component (23.1% of the videos) occupies most of the attention (82.6% of the views). We also propose a new model, called ARNet, which accounts for the network structure and can predict video popularity 9.7% better than other baselines. More importantly, the ARNet model allows us to quantify the latent influence between videos and artists. To our knowledge, this is the first work that measures YouTube recommendation as a network and links to the attention consumption for the videos in it.

- **Large-scale datasets, open tools, and web demonstrations.** The contributions of our work go beyond quantitative observations. Altogether, we have released 3 new datasets, 2 data collection tools, and 2 new web demos. They include: (1) YOUTUBE ENGAGEMENT '16 dataset: 5.3M videos published and tweeted between July and August, 2016, and 3 quality video datasets. (2) VEVO MUSIC GRAPH dataset: 60K music videos with 63 daily snapshots of the video recommendation network. (3) COMPLETE/SAMPLED RETWEET CASCADES datasets: 2 sets of complete/sampled retweet cascades on the topics of cyberbullying and YouTube video sharing. (4) TWITTER-INTACT-STREAM: a Python package for reconstructing the complete filtered stream on Twitter. (5) YOUTUBE-INSIGHT: a Python package for collecting metadata and historical data for videos on YouTube. (6) HIPie: an interactive web interface for explaining and predicting the popularity of YouTube videos. (7) AttentionFlow: an interactive web interface for visualizing a collection of time series and the dynamic network influence.

6.2 Future work

We envision our long-term goal as developing principles for responsible platforms by measuring and modeling the collective user behaviors.

- **Measuring conversation quality in online platform.** In Chapter 4, we use video watching metrics as a surrogate of collective user engagement. Another important engagement behavior is commenting. Although YouTube becomes a prevailing platform, the conversation (i.e., comments) on YouTube videos is surprisingly understudied. Recently, scholars have used YouTube comments to study hate speech [Mathew et al., 2019], content moderation [Jiang et al., 2019], and political polarization [Ribeiro et al., 2020]. However, we still lack the knowledge of how users comment on YouTube videos in general.

  We can narrow down the scope to the conversation around political videos, so that both the users and videos carry clear ideologies. It is of great importance to understand how users engage with videos and users with the same or opposite ideologies. Specifically, one can ask questions such as how many comments on YouTube are cross-partisan? are the cross-partisan comments more toxic? do cross-partisan comments get more attention? which topics or media are more likely to incur cross-partisan conversation?[1]

  Moving beyond YouTube, we can intersect the videos with the sharing behaviors on other social media, e.g., tweets on Twitter. Hence, an important question is to measure the difference between the comments on YouTube and tweets on Twitter for the same video. Overall, this question can help us understand the engagement patterns across platforms and shed light on advertising online products.

- **Auditing biases in the recommender systems.** In Chapter 5, we discover the popularity bias in the YouTube recommender systems. We can extend the study to other biases, e.g., in politics, education, medical information, or culture trends.

  For the political bias in recommender systems, there have been many anecdotal reports on how YouTube radicalizes the users by recommending more extreme

[1] A preliminary attempt to answer these questions is available in [Wu and Resnick, 2021].

videos [Tufekci, 2018; Lewis, 2018]. Recently, Ribeiro et al. [2020] presented a quantitative analysis and used YouTube comments as a proxy of video consumption. We can bridge the gaps with prior work from two fronts: Firstly, all existing research heavily focus on the right-wing party, but we argue that it is important to understand the radicalization problem from the whole political spectrum. Secondly, to perform a systematic study, we propose to build a simulator for mimicking real users and auditing the recommendations given by the platform. The simulator needs to account for different demographic features (e.g., gender, location, political leaning) and different user behaviors (e.g., random click or always click the first recommendation).

Appendix A

Appendix

A.1 Twitter data in ICWSM papers (2015-2019)

82 (31%) out of 265 ICWSM full papers used Twitter data from 2015 to 2019. Twitter search API has been used 25 times, sampled stream 12 times, filtered stream 18 times, firehose 8 times. 12 papers used multiple Twitter APIs for data collection. 7 papers did not clearly specify their Twitter API choices.

Id	Paper	APIs	Notes
1	Audience analysis for competing memes in social media	search	searched keywords “Russia”, “meteor”, “Fox”, and “Obama”
2	Making use of derived personality: The case of social media ad targeting	filtered	mention at least one term related to NYC and one term related to traveling
3	The many shades of anonymity: Characterizing anonymous social media content	unspecified, possibly sampled	500 random publicly available tweets
4	On analyzing hashtags in Twitter	sampled; search	10M messages crawled in December 2013; 200 tweets for each hashtag in our original dataset
5	WhichStreams: A dynamic approach for focused data capture from large social media	sampled; filtered	5000 users first to use one of the keywords “Obama”, “Romney” or “#USElections”
6	Characterizing silent users in social media communities	filtered	all tweets of 140,851 Singapore-based users and 126,047 Indonesia-based users
7	Predicting user engagement on Twitter with real-world events	firehose	nearly 2.7 billion English tweets during August of 2014
8	Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice	sampled	10% sampled stream
9	Characterizing information diets of social media users	sampled	500 randomly selected tweets from Twitter’s 1% random sample
10	Degeneracy-based real-time sub-event detection in Twitter stream	unspecified, possibly filtered	several football matches that took place during the 2014 FIFA World Cup in Brazil, between June, 12nd and July, 13rd 2014
11	CREDBANK: A large-scale social media corpus with associated credibility annotations	sampled	1% random sample
12	Understanding musical diversity via online social media	search	collected U.S. Twitter users who share their Last.fm accounts, then we collected all publicly available tweets
13	Smelly maps: The digital life of urban smellscapes	unspecified	collected 5.3M tweets during year 2010 and from October 2013 to February 2014
14	Project recommendation using heterogeneous traits in crowdfunding	search	retrieving tweets containing URLs that begin with http://kck.st
15	Don’t let me be #misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes	sampled	approximately 15% of the Twitter stream in six month period
16	SEEFT: Planned social event discovery and attribute extraction by fusing Twitter and web content	unspecified, possibly search	querying Twitter API with 3 event types, namely concerts, conferences, and festivals
17	A bayesian graphical model to discover latent events from Twitter	sampled	1% sampled stream and 10% sampled stream
18	Patterns in interactive tagging networks	sample; search	randomly sampled 1 million seed users from sample streams on December 2014; following network starting from the same 1 million seed users
19	Hierarchical estimation framework of multi-label classifying: A case of tweets classifying into real life aspects	search	collected 2,390,553 tweets posted from April 15, 2012 to August 14, 2012, each of which has “Kyoto” as the Japanese location information
20	The lifecyle of a Youtube video: Phases, content and popularity	filtered	tweets containing keyword “youtube” OR (“youtu” AND “be”)

Table A.1: Year 2015. 20 out of 64 papers used Twitter data. search API: 4; sampled stream: 5; filtered stream: 3; firehose: 1; unspecified: 4; multiple APIs: 3.

Id	Paper	APIs	Notes
1	Are you charlie or ahmed? Cultural pluralism in charlie hebdo response on Twitter	search	#JeSuisCharlie, #JeSuisAhmed, and #CharlieHebdo – from 2015-01-07 to 2015-01-28
2	When a movement becomes a party: Computational assessment of new forms of political organization in social media	filtered	extracted 373,818 retweets of tweets that (1) were created by, (2) were retweeted by, or (3) mentioned a user from the list
3	Journalists and Twitter: A multidimensional quantitative description of usage patterns	search	contained 5,358 accounts of journalists and news organizations, crawled all their 13,140,449 public tweets
4	Social media participation in an activist movement for racial equality	filtered	#ferguson, #BlueLivesMatter, #BlackLivesMatter, #AllLivesMatter, #Baltimore, #BaltimoreRiots, #BaltimoreUprising, and #FreddieGray
5	Understanding communities via hashtag engagement: A clustering based approach	firehose	tweets from all English language Twitter users in U.S. that used certain hashtag from Jan 15 to Feb 15, 2015
6	Investigating the observability of complex contagion in empirical social networks	filtered; search	45 Nigerian cities with at least 100K population within a radius from 25 to 40 miles; collected tweets from the timelines of selected users
7	Dynamic data capture from social media streams: A contextual bandit approach	sampled; filtered	leverage sampled stream to discover unknown users; filtered stream for realtime data of the subset users
8	On unravelling opinions of issue specific-silent users in social media	search	asked some Twitter users to provide their screen names for crawling
9	Distinguishing between topical and non-topical information diffusion mechanisms in social media	search	a dataset that nearly contains all public tweets produced by users until 2009-09 and a snapshot of social graph crawled in 2009-09
10	TweetGrep: Weakly supervised joint retrieval and sentiment analysis of topical tweets	search	the queries are issued to the Twitter Search Web Interface via a proxy that we developed
11	What the language you tweet says about your occupation	search	download users' 3,000 most recent tweets
12	TiDeH: Time-dependent hawkes process for predicting retweet dynamics	firehose	SEISMIC dataset by Zhao et al. 2015
13	Emotions, demographics and sociability in Twitter interactions	search	collect tweets from an area that included Los Angeles, then collect all (timeline) tweets from subset users
14	Analyzing personality through social media profile picture choice	search	we have collected up to 3,200 most recent tweets for each user
15	Cross social media recommendation	unspecified, possibly sampled	corpora were sampled between 2012-09-17 and 2012-09-23
16	Understanding anti-vaccination attitudes in social media	firehose	snowball (firehose) from a selected 1000 Twitter users
17	Twitter's glass ceiling: The effect of perceived gender on online visibility	sampled	10% sampled stream
18	Mining pro-ISIS radicalisation signals from social media users	search	Twitter user timeline of 154K users
19	Predictability of popularity: Gaps between prediction and understanding	sampled	URLs tweeted by 737k users for three weeks of 2010
20	Theme-relevant truth discovery on Twitter: An estimation theoretic approach	search	collected through Twitter search API using query terms and specified geographic regions related to the events
21	#PrayForDad: Learning the semantics behind why social media users disclose health information	filtered	collect tweets in English and published in the contiguous United States during a four-month window in 2014
22	Your age is no secret: Inferring microbloggers' ages via content and interaction analysis	filtered	record all the tweets which contain one of the keywords "happy *y*th birthday" with y ranging from 14 to 70
23	EigenTransitions with hypothesis testing: The anatomy of urban mobility	filtered	collected geo-tagged Tweets generated within the area covering NYC and Pittsburgh from 2013-07-15 to 2014-11-09

Table A.2: Year 2016. 23 out of 52 papers used Twitter data. search API: 10; sampled stream: 2; filtered stream: 5; firehose: 3; unspecified: 1; multiple APIs: 2.

Id	Paper	APIs	Notes
1	Who makes trends? Understanding demographic biases in crowdsourced recommendations	sampled; search	1% random sample; queried search API every 5 minutes and collected all topics which became trending in US
2	#NotOkay: Understanding gender-based violence in social media	sampled; filtered	1% random sample; collect tweets that contain the indicated hashtags from October 26th to November 26th, 2016
3	Online popularity under promotion: Viral potential, forecasting, and the economics of time	filtered	tweets containing keyword "youtube" OR ("youtu" AND "be")
4	Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter	filtered	tracked "shooter, shooting, gunman, gunmen, gunshot, gunshots, shooters, gun shot, gun shots, shootings" between January 1 and October 5, 2016
5	State of the geotags: Motivations and recent changes	filtered	selected all coordinate-geotagged tweets within 0.2 degrees latitude and longitude from Pittsburgh
6	Online human-bot interactions: Detection, estimation, and characterization	search	collected the most recent tweets produced by those accounts
7	Identifying effective signals to predict deleted and suspended accounts on Twitter across languages	sampled; search	1% random sample; batches of 100 unique users were queried against the public Twitter API
8	Adaptive spammer detection with sparse group modeling	search	crawled a Twitter dataset from July 2012 to September 2012 via the Twitter Search API
9	Wearing many (social) hats: How different are your different social network personae?	search	76% of About.me users in our dataset have linked their profiles to their alternate account in Twitter

Table A.3: Year 2017. 9 out of 50 papers used Twitter data. search API: 3; filtered stream: 3; multiple APIs: 3.

Id	Paper	APIs	Notes
1	Peer to peer hate: Hate speech instigators and their targets	sampled; search	1% random sample; we use search API to fetch tweet traces of users
2	Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses	search	obtain the list of individuals who have retweeted each tweet from the disclosers during this period of analysis
3	Facebook versus Twitter: Cross-platform differences in self-disclosure and trait prediction	search	we collected participants' social media posts
4	Can you verifi this? Studying uncertainty and decision-making about misinformation using visual analytics	filtered	collected 103,248 tweets posted by these 178 accounts along with account metadata from May 23, 2017 to June 6, 2017
5	Using longitudinal social media analysis to understand the effects of early college alcohol use	firehose	extract 639k tweets that match these keywords in August-December 2010 in our organization's archive of the Twitter firehose
6	Modeling popularity in asynchronous social media streams with recurrent neural networks	filtered; firehose	tweets containing keyword "youtube" OR ("youtu" AND "be"); SEISMIC dataset by Zhao et al. 2015
7	The effect of extremist violence on hateful speech online	sampled	10% random sample
8	You are your metadata: Identification and obfuscation of social media users using metadata information	sampled	random sample of the tweets posted between October 2015 and January 2016
9	#DebateNight: The role and influence of socialbots on Twitter during the first 2016 U.S. presidential debate	firehose	Twitter discussions that occurred during the 1st 2016 U.S presidential debate between Hillary Clinton and Donald Trump
10	Ecosystem or echo-system? Exploring content sharing across alternative media domains	filtered	tracked various keyword terms related to the Syrian conflict including geographic terms of affected areas
11	COUPLENET: Paying attention to couples with coupled attention for relationship recommendation	filtered	collected tweets with emojis contains the keyword "heart" in its description
12	Beyond views: Measuring and predicting engagement in online videos	filtered	tweets containing keyword "youtube" OR ("youtu" AND "be")
13	Understanding web archiving services and their (mis)use on social media	sampled	1% random sample

Table A.4: Year 2018. 13 out of 48 papers used Twitter data. search API: 2; sampled stream: 3; filtered stream: 4; firehose: 2; multiple APIs: 2.

Id	Paper	APIs	Notes
1	Linguistic cues to deception: Identifying political trolls on social media	firehose	a list of 2,752 Russian troll accounts, then collected all of the trolls' discussions
2	Tweeting MPs: Digital engagement between citizens and members of parliament in the UK	search	we fetched all the users (?4.28 Million) who follow MPs and also the users that MPs followed (869K)
3	View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations	filtered	collecting all posts from a news organization
4	A large-scale study of ISIS social media strategy: Community size, collective influence, and behavioral impact	firehose	a large dataset of 9.3 billion tweets representing all tweets generated in the Arabic language in 2015 through full private access to the Twitter firehose
5	Who should be the captain this week? Leveraging inferred diversity-enhanced crowd wisdom for a fantasy premier league captain prediction	unspecified	collected their soccer related tweets by scraping Twitter user timelines (for a total 4,299,738 tweets)
6	Multimodal social media analysis for gang violence prevention	search	we scraped all obtainable tweets from this list of 200 users in February 2017
7	Hot streaks on social media	search	we obtained all tweets, followers, and retweeters of all tweets using the Twitter REST API
8	Understanding and measuring psychological stress using social media	search	601 active users who completed the survey
9	Studying cultural differences in emoji usage across the east and the west	sampled	10% random sample
10	What Twitter profile and postedImages reveal about depression and anxiety	search	downloaded the 3200 most recent user tweets for each user, leading to a data set of 5,547,510 tweets, out of which 700,630 posts contained images and 1 profile image each across 3498 users
11	Polarized, together: Comparing partisan support for Trump's tweets using survey and platform-based measures	sampled; search	collecting a large sample of Twitter users (approximately 406M) who sent one or more tweets that appeared in the Twitter Decahose from Jan 2014 to Aug 2016; select from this set the approximately 322M accounts that were still active in Mar 2017
12	Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 U.S. cities	sampled	1% sample of Twitter's public stream from January 1st, 2011 to December 31st, 2016
13	A social media study on the effects of psychiatric medication use	sampled; filtered	public English posts mentioning these drugs between January 01, 2015 and December 31, 2016
14	SENPAI: Supporting exploratory text analysis through semantic&syntactic pattern inspection	filtered	gathered a dataset of Twitter messages from 103 professional journalists and bloggers who work in the field of American Politics
15	Empirical analysis of the relation between community structure and cascading retweet diffusion	search	we used the Search API and collected Japanese tweets using the query q=RT, lang=ja
16	Measuring the importance of user-generated content to search engines	unspecified	a row of three cards with one tweet each. Google obtains the tweets either from Twitter's search (a SearchTweetCarousel) or a single user (a UserTweetCarousel)
17	Detecting journalism in the age of social media:Three experiments in classifying journalists on Twitter	filtered	tracking a set of event-related keywords and hashtags

Table A.5: Year 2019. 17 out of 51 papers used Twitter data. search API: 6; sampled stream: 2; filtered stream: 3; firehose: 2; unspecified: 2; multiple APIs: 2.

Bibliography

Abisheva, A.; Garimella, V. R. K.; Garcia, D.; and Weber, I., 2014. Who watches (and shares) what on YouTube? and when?: Using Twitter to understand YouTube viewership. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on page 16)

Ahmed, M.; Spagna, S.; Huici, F.; and Niccolini, S., 2013. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on page 15)

Airoldi, M.; Beraldo, D.; and Gandini, A., 2016. Follow the algorithm: An exploratory investigation of music on YouTube. *Poetics*, (2016). (cited on pages 18, 74, 77, 79, and 98)

alexa.com, 2020. Alexa top 500 global sites. https://www.alexa.com/topsites. [Online; accessed 2020/06/01]. (cited on page 1)

Allen, M. P., 1997. The coefficient of determination in multiple regression. *Understanding Regression Analysis*, (1997). (cited on page 60)

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A., 2013. Tweets are forever: A large-scale quantitative analysis of deleted tweets. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. (cited on page 11)

Alstott, J.; Bullmore, E.; and Plenz, D., 2014. powerlaw: A python package for analysis of heavy-tailed distributions. *PloS one*, (2014). (cited on page 82)

Arakelyan, S.; Morstatter, F.; Martin, M.; Ferrara, E.; and Galstyan, A., 2018. Mining and forecasting career trajectories of music artists. In *Proceedings of the ACM Conference on Hypertext and Social Media*. (cited on page 37)

Arantes, M.; Figueiredo, F.; and Almeida, J. M., 2016. Understanding video-ad consumption on YouTube: A measurement study on user behavior, popularity, and content properties. In *Proceedings of the ACM Conference on Web Science*. (cited on pages 4 and 13)

Arapakis, I.; Lalmas, M.; and Valkanas, G., 2014. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. (cited on pages 13 and 47)

Backstrom, L.; Bakshy, E.; Kleinberg, J. M.; Lento, T. M.; and Rosenn, I., 2011. Center of attention: How facebook users allocate attention across friends. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 9)

Barbieri, N.; Silvestri, F.; and Lalmas, M., 2016. Improving post-click user engagement on native ads via survival analysis. In *Proceedings of the International Conference on World Wide Web*. (cited on page 14)

Bauckhage, C.; Hadiji, F.; and Kersting, K., 2015. How viral are viral videos? In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 15)

Bendersky, M.; Garcia-Pueyo, L.; Harmsen, J.; Josifovski, V.; and Lepikhin, D., 2014. Up next: Retrieval methods for large scale related video suggestion. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (cited on pages 17, 73, and 74)

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E. H.; et al., 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop*. (cited on page 17)

Beutel, A.; Covington, P.; Jain, S.; Xu, C.; Li, J.; Gatto, V.; and Chi, E. H., 2018. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on pages 17, 74, and 87)

Bista, U.; Mathews, A.; Shin, M.; Menon, A. K.; and Xie, L., 2019. Comparative document summarisation via classification. In *Proceedings of the International AAAI Conference*. (cited on page 11)

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J., 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. (cited on page 61)

Bovet, A. and Makse, H. A., 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, (2019). (cited on page 20)

Boyd, D. and Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, (2012). (cited on pages 2 and 12)

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; and Wiener, J., 2000. Graph structure in the web. *Computer networks*, (2000). (cited on pages 5, 40, 75, 81, 83, and 84)

Brodersen, A.; Scellato, S.; and Wattenhofer, M., 2012. YouTube around the world. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 15 and 49)

Buolamwini, J. and Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. (cited on page 20)

Buscher, G.; Cutrell, E.; and Morris, M. R., 2009. What do you see when you're surfing?: Using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (cited on pages 4 and 13)

Carmi, E.; Oestreicher-Singer, G.; Stettner, U.; and Sundararajan, A., 2017. Is Oprah contagious? the depth of diffusion of demand shocks in a product network. *MIS Quarterly*, (2017). (cited on pages 18 and 87)

Celma, Ò. and Cano, P., 2008. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop*. (cited on page 73)

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S., 2007. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the ACM SIGCOMM conference on Internet Measurement*. (cited on page 14)

Chabris, C. F. and Simons, D. J., 2010. *The invisible gorilla: And other ways our intuitions deceive us*. Harmony. (cited on page 8)

Chen, M.; Beutel, A.; Covington, P.; Jain, S.; Belletti, F.; and Chi, E. H., 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings* *of the ACM International Conference on Web Search and Data Mining*. (cited on pages 9, 17, 73, and 74)

Chen, Y.; Zhang, B.; Liu, Y.; and Zhu, W., 2013. Measurement and modeling of video watching time in a large-scale internet video-on-demand system. *IEEE Transactions on Multimedia*, (2013). (cited on page 14)

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J., 2014. Can cascades be predicted? In *Proceedings of the International Conference on World Wide Web*. (cited on pages 15 and 66)

Cheng, J.; Adamic, L. A.; Kleinberg, J. M.; and Leskovec, J., 2016. Do cascades recur? In *Proceedings of the International Conference on World Wide Web*. (cited on page 96)

Cheng, L.; Shu, K.; Wu, S.; Silva, Y. N.; Hall, D. L.; and Liu, H., 2020. Unsupervised cyberbullying detection via time-informed gaussian mixture model. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. (cited on page 22)

Cheng, X.; Dale, C.; and Liu, J., 2008. Statistics and social network of YouTube videos. In *Proceedings of International Workshop on Quality of Service*. (cited on pages 14, 18, 74, and 93)

cisco.com, 2017. Cisco visual networking index: Forecast and methodology, 2016–2021. https://www.reinvention.be/webhdfs/v1/docs/complete-white-paper-c11-481360.pdf. [Online; accessed 2020/08/22]. (cited on page 2)

Clauset, A.; Shalizi, C. R.; and Newman, M. E., 2009. Power-law distributions in empirical data. *SIAM review*, (2009). (cited on page 82)

cnet.com, 2018. Youtube's ai is the puppet master over most of what you watch. https://www.cnet.com/news/youtube-ces-2018-neal-mohan//. [Online; accessed 2020/06/01]. (cited on page 2)

Competition, M., 2018. M4-methods. https://github.com/M4Competition/M4-methods/blob/master/ML_benchmarks.py. (cited on page 92)

Covington, P.; Adams, J.; and Sargin, E., 2016. Deep neural networks for YouTube recommendations. In *Proceedings of the ACM Conference on Recommender Systems*. (cited on pages 13, 17, 47, 61, 65, 73, 74, 80, and 87)

Crane, R. and Sornette, D., 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, (2008). (cited on page 15)

Davidson, J.; Liebald, B.; Liu, J.; Nandy, P.; Van Vleet, T.; Gargi, U.; Gupta, S.; He, Y.; Lambert, M.; Livingston, B.; et al., 2010. The YouTube video recommendation system. In *Proceedings of the ACM conference on Recommender Systems*. (cited on pages 13, 17, 73, 74, and 87)

De Choudhury, M.; Jhaver, S.; Sugar, B.; and Weber, I., 2016. Social media participation in an activist movement for racial equality. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 2, 11, and 20)

De Choudhury, M.; Lin, Y.-R.; Sundaram, H.; Candan, K. S.; Xie, L.; and Kelliher, A., 2010. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 12)

Dhar, V.; Geva, T.; Oestreicher-Singer, G.; and Sundararajan, A., 2014. Prediction in economic networks. *Information Systems Research*, (2014). (cited on pages 18, 73, and 87)

Dobrian, F.; Sekar, V.; Awan, A.; Stoica, I.; Joseph, D.; Ganjam, A.; Zhan, J.; and Zhang, H., 2011. Understanding the impact of video quality on user engagement. *ACM SIGCOMM Computer Communication Review*, (2011). (cited on pages 13, 60, and 61)

Drutsa, A.; Gusev, G.; and Serdyukov, P., 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 14 and 47)

Dunbar, R. I., 1992. Neocortex size as a constraint on group size in primates. *Journal of human evolution*, (1992). (cited on page 9)

Dupret, G. and Lalmas, M., 2013. Absence time and user engagement: evaluating ranking functions. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on pages 13 and 14)

facebook.com, 2017. Updating how we account for video completion rates. https://about.fb.com/news/2017/01/news-feed-fyi-updating-how-we-account-for-video-completion-rates/. [Online; accessed 2020/06/01]. (cited on pages 2 and 13)

FIGUEIREDO, F.; ALMEIDA, J. M.; BENEVENUTO, F.; AND GUMMADI, K. P., 2014. Does content determine information popularity in social media?: A case study of YouTube videos' content and their popularity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (cited on page 13)

FIGUEIREDO, F.; ALMEIDA, J. M.; GONÇALVES, M. A.; AND BENEVENUTO, F., 2016. Trendlearner: Early prediction of popularity trends of user generated content. *Information Sciences*, (2016). (cited on pages 15, 51, 96, and 102)

FIGUEIREDO, F.; BENEVENUTO, F.; AND ALMEIDA, J. M., 2011. The tube over time: Characterizing popularity growth of YouTube videos. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on page 15)

GAFFNEY, D. AND MATIAS, J. N., 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one*, (2018). (cited on page 12)

GHOSH, S.; ZAFAR, M. B.; BHATTACHARYA, P.; SHARMA, N.; GANGULY, N.; AND GUMMADI, K., 2013. On sampling the wisdom of crowds: Random vs. expert sampling of the Twitter stream. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. (cited on page 11)

GILL, P.; ARLITT, M.; LI, Z.; AND MAHANTI, A., 2007. YouTube traffic characterization: A view from the edge. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*. (cited on page 15)

GIRIDHARADAS, A., 2019. *Winners take all: The elite charade of changing the world*. Vintage. (cited on page 9)

GOMEZ-URIBE, C. A. AND HUNT, N., 2016. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, (2016). (cited on pages 17 and 73)

GONÇALVES, B.; PERRA, N.; AND VESPIGNANI, A., 2011. Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PloS one*, (2011). (cited on page 9)

GONZÁLEZ-BAILÓN, S.; WANG, N.; RIVERO, A.; BORGE-HOLTHOEFER, J.; AND MORENO, Y., 2014. Assessing the bias in samples of large online networks. *Social Networks*, (2014). (cited on pages 11, 35, and 40)

GUO, P. J.; KIM, J.; AND RUBIN, R., 2014. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the ACM Conference on Learning@ Scale*. (cited on pages 14 and 55)

GUO, Q. AND AGICHTEIN, E., 2012. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the International Conference on World Wide Web*. (cited on page 14)

GÜRSUN, G.; CROVELLA, M.; AND MATTA, I., 2011. Describing and forecasting video access patterns. In *Proceedings of the IEEE International Conference on Computer Communications*. (cited on page 15)

HAWKES, A. G., 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, (1971). (cited on page 67)

HERLOCKER, J. L.; KONSTAN, J. A.; TERVEEN, L. G.; AND RIEDL, J. T., 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, (2004). (cited on page 80)

HESSEL, J.; LEE, L.; AND MIMNO, D., 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *Proceedings of the International Conference on World Wide Web*. (cited on page 60)

HOFMAN, J. M.; SHARMA, A.; AND WATTS, D. J., 2017. Prediction and explanation in social systems. *Science*, (2017). (cited on pages 15 and 47)

HUANG, L.; DING, B.; WANG, A.; XU, Y.; ZHOU, Y.; AND LI, X., 2018. User behavior analysis and video popularity prediction on a large-scale VoD system. *ACM Transactions on Multimedia Computing, Communications, and Applications*, (2018). (cited on page 93)

IE, E.; JAIN, V.; WANG, J.; NARVEKAR, S.; AGARWAL, R.; WU, R.; CHENG, H.-T.; CHANDRA, T.; AND BOUTILIER, C., 2019. SlateQ: A tractable decomposition for reinforcement learning with recommendation sets. In *Proceedings of the International Joint Conferences on Artificial Intelligence*. (cited on pages 17 and 74)

JAMES, W., 2007. *The principles of psychology*. Cosimo, Inc. (cited on page 8)

JIANG, S.; ROBERTSON, R. E.; AND WILSON, C., 2019. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 105)

JOSEPH, K.; LANDWEHR, P. M.; AND CARLEY, K. M., 2014. Two 1%s don't make a whole: Comparing simultaneous samples from Twitter's streaming API. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. (cited on pages 12 and 35)

KERGL, D.; ROEDLER, R.; AND SEEBER, S., 2014. On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. (cited on page 20)

KHAN, M. L., 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, (2017). (cited on page 2)

KIM, M.; XIE, L.; AND CHRISTEN, P., 2012. Event diffusion patterns in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 84)

KOENKER, R. AND HALLOCK, K. F., 2001. Quantile regression. *Journal of economic perspectives*, (2001). (cited on page 54)

KONG, Q.; RIZOIU, M.-A.; WU, S.; AND XIE, L., 2018. Will this video go viral: Explaining and predicting the popularity of YouTube videos. In *Companion Proceedings of the International Conference on World Wide Web*. (cited on page 5)

KONG, Q.; RIZOIU, M.-A.; AND XIE, L., 2020a. Describing and predicting online items with reshare cascades via dual mixture self-exciting processes. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. (cited on page 16)

KONG, Q.; RIZOIU, M.-A.; AND XIE, L., 2020b. Modeling information cascades with self-exciting processes via generalized epidemic models. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on page 15)

KONSTAN, J. A. AND RIEDL, J., 2012. Recommender systems: From algorithms to user experience. *User modeling and user-adapted interaction*, (2012). (cited on pages 9, 17, and 87)

KOSSINETS, G., 2006. Effects of missing data in social networks. *Social Networks*, (2006). (cited on page 12)

KRUMME, C.; CEBRIAN, M.; PICKARD, G.; AND PENTLAND, S., 2012. Quantifying social influence in an online cultural market. *PloS one*, (2012). (cited on pages 1, 10, and 55)

KUZNETSOV, V. AND MARIET, Z., 2019. Foundations of sequence-to-sequence modeling for time series. In *Proceedings of International Conference on Artificial Intelligence and Statistics*. (cited on pages 16, 95, and 96)

KWAK, H.; LEE, C.; PARK, H.; AND MOON, S., 2010. What is Twitter, a social network or a news media? In *Proceedings of the International Conference on World Wide Web*. (cited on page 82)

LAGUN, D. AND LALMAS, M., 2016. Understanding user attention and engagement in online news reading. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on page 13)

LALMAS, M.; LEHMANN, J.; SHAKED, G.; SILVESTRI, F.; AND TOLOMEI, G., 2015. Promoting positive post-click experience for in-stream Yahoo gemini users. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (cited on pages 13 and 17)

LEHMANN, J.; GONÇALVES, B.; RAMASCO, J. J.; AND CATTUTO, C., 2012. Dynamical classes of collective attention in Twitter. In *Proceedings of the International Conference on World Wide Web*, 251–260. (cited on page 2)

LERMAN, K. AND HOGG, T., 2014. Leveraging position bias to improve peer recommendation. *PloS one*, (2014). (cited on page 10)

LESKOVEC, J. AND FALOUTSOS, C., 2006. Sampling from large graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (cited on pages 12 and 31)

LEWIS, R., 2018. Alternative influence: Broadcasting the reactionary right on youtube. *Data & Society*, (2018). (cited on page 106)

LI, C.; MA, J.; GUO, X.; AND MEI, Q., 2017. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the International Conference on World Wide Web*. (cited on page 16)

LI, S.; ABBASI-YADKORI, Y.; KVETON, B.; MUTHUKRISHNAN, S.; VINAY, V.; AND WEN, Z., 2018. Offline evaluation of ranking policies with click models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (cited on page 17)

MA, C.; YAN, Z.; AND CHEN, C. W., 2017. LARM: A lifetime aware regression model for predicting YouTube video popularity. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. (cited on page 16)

MARTIN, T.; HOFMAN, J. M.; SHARMA, A.; ANDERSON, A.; AND WATTS, D. J., 2016. Exploring limits to prediction in complex social systems. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 15, 41, 47, 60, 61, 62, 66, 90, and 96)

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherjee, A., 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 105)

Matsubara, Y.; Sakurai, Y.; Prakash, B. A.; Li, L.; and Faloutsos, C., 2012. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (cited on pages 57 and 65)

Meusel, R.; Vigna, S.; Lehmberg, O.; and Bizer, C., 2014. Graph structure in the web—revisited: A trick of the heavy tail. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 82 and 84)

Mishra, S.; Rizoiu, M.-A.; and Xie, L., 2016. Feature driven and point process approaches for popularity prediction. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. (cited on pages 11, 16, 41, 42, 44, 61, and 67)

Morstatter, F.; Pfeffer, J.; and Liu, H., 2014. When is it biased?: Assessing the representativeness of Twitter's streaming API. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 11 and 35)

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M., 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 2, 11, 21, 35, and 40)

Mosteller, F., 1949. The pre-election polls of 1948. *Social Science Research Council*, (1949). (cited on page 20)

Nand, P.; Perera, R.; and Kasture, A., 2016. How bullying is this message?: A psychometric thermometer for bullying. In *Proceedings of the International Conference on Computational Linguistics*. (cited on page 22)

Newman, M. E., 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, (2001). (cited on page 37)

Nguyen, T. T.; Hui, P.-M.; Harper, F. M.; Terveen, L.; and Konstan, J. A., 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the International Conference on World Wide Web*. (cited on page 1)

Oestreicher-Singer, G. and Sundararajan, A., 2012. Recommendation networks and the long tail of electronic commerce. *MIS Quarterly*, (2012). (cited on pages 18 and 73)

Olteanu, A.; Castillo, C.; Diaz, F.; and Kiciman, E., 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, (2019). (cited on pages 2, 12, and 20)

Park, M.; Naaman, M.; and Berger, J., 2016. A data-driven study of view duration on YouTube. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 14, 47, 55, and 61)

Pfeffer, J.; Mayer, K.; and Morstatter, F., 2018. Tampering with Twitter's sample API. *EPJ Data Science*, (2018). (cited on pages 11, 20, and 35)

Piketty, T., 2015. About capital in the twenty-first century. *American Economic Review*, (2015). (cited on page 9)

Pinto, H.; Almeida, J. M.; and Gonçalves, M. A., 2013. Using early view patterns to predict the popularity of YouTube videos. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on pages 2, 15, 46, 49, 67, and 96)

Resnick, P.; Garrett, R. K.; Kriplean, T.; Munson, S. A.; and Stroud, N. J., 2013. Bursting your (filter) bubble: Strategies for promoting diverse exposure. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing Companion*. (cited on page 1)

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A.; and Meira Jr, W., 2020. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. (cited on pages 105 and 106)

Rizoiu, M.-A.; Lee, Y.; Mishra, S.; and Xie, L., 2017a. A tutorial on Hawkes processes for events in social media. *Frontiers of Multimedia Research*, (2017). (cited on page 16)

Rizoiu, M.-A.; Mishra, S.; Kong, Q.; Carman, M.; and Xie, L., 2018. SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In *Proceedings of the International Conference on World Wide Web*. (cited on page 90)

Rizoiu, M.-A. and Xie, L., 2017. Online popularity under promotion: Viral potential, forecasting, and the economics of time. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 2, 49, 58, 67, and 96)

Rizoiu, M.-A.; Xie, L.; Sanner, S.; Cebrian, M.; Yu, H.; and Van Hentenryck, P., 2017b. Expecting to be HIP: Hawkes intensity processes for social media popularity. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 2, 16, 22, 46, 48, 49, 65, 66, 68, and 90)

Ruths, D. and Pfeffer, J., 2014. Social media for large studies of behavior. *Science*, (2014). (cited on pages 12 and 20)

Sadikov, E.; Medina, M.; Leskovec, J.; and Garcia-Molina, H., 2011. Correcting for missing data in information cascades. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on pages 12 and 40)

Salganik, M. J.; Dodds, P. S.; and Watts, D. J., 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, (2006). (cited on pages 1, 10, 47, and 55)

Sampson, J.; Morstatter, F.; Maciejewski, R.; and Liu, H., 2015. Surpassing the limit: Keyword clustering to improve Twitter sample coverage. In *Proceedings of the ACM Conference on Hypertext and Social Media*. (cited on pages 12, 21, 25, 26, and 28)

Sedhain, S.; Sanner, S.; Xie, L.; Kidd, R.; Tran, K.-N.; and Christen, P., 2013. Social affinity filtering: Recommendation through fine-grained analysis of user interactions and activities. In *Proceedings of the ACM conference on Online Social Networks*. (cited on page 64)

Sharma, A.; Hofman, J. M.; and Watts, D. J., 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the ACM Conference on Economics and Computation*. (cited on page 18)

Sharma, A. and Yan, B., 2013. Pairwise learning in recommendation: Experiments with community recommendation on Linkedin. In *Proceedings of the ACM Conference on Recommender Systems*. (cited on page 73)

Shen, W.; Hu, Y. J.; and Ulmer, J. R., 2015. Competing for attention: An empirical study of online reviewers' strategic behavior. *Mis Quarterly*, (2015). (cited on page 9)

Shin, M.; Tran, A.; Wu, S.; Mathews, A.; Wang, R.; Lyall, G.; and Xie, L., 2021. Attentionflow: Visualising influence in networks of time series. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on page 5)

Shuyo, N., 2010. Language detection library for Java. http://code.google.com/p/language-detection/. (cited on page 60)

Simon, H. A., 1971. Designing organizations for an information-rich world. In *Computers, communications, and the public interest*. (cited on page 8)

Stella, X. Y. and Shi, J., 2003. Multiclass spectral clustering. In *Proceedings of the International Conference on Computer Vision*. (cited on page 38)

Stoddard, G., 2015. Popularity dynamics and intrinsic quality in Reddit and Hacker News. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 10)

Su, J.; Sharma, A.; and Goel, S., 2016. The effect of recommendations on network structure. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 18 and 73)

Sun, E.; de Oliveira, R.; and Lewandowski, J., 2017. Challenges on the journey to co-watching YouTube. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. (cited on page 13)

Swart, M.; Lopez, Y.; Mathur, A.; and Chetty, M., 2020. Is this an ad?: Automatically disclosing online endorsements on youtube with adintuition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (cited on page 13)

Szabo, G. and Huberman, B. A., 2010. Predicting the popularity of online content. *Communications of the ACM*, (2010). (cited on pages 15, 51, and 67)

Tarjan, R., 1972. Depth-first search and linear graph algorithms. *SIAM journal on computing*, (1972). (cited on page 83)

Tong, L. C.; Acikalin, M. Y.; Genevsky, A.; Shiv, B.; and Knutson, B., 2020. Brain activity forecasts video engagement in an internet attention market. *Proceedings of the National Academy of Sciences*, (2020). (cited on page 14)

Tran, A.; Mathews, A.; Ong, C. S.; and Xie, L., 2021. Radflow: A recurrent, aggregated, and decomposable model for networks of time series. In *Proceedings of the International Conference on World Wide Web*. (cited on page 96)

tubefilter.com, 2019. More than 500 hours of content are now being uploaded to youtube every minute. https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/. [Online; accessed 2020/06/01]. (cited on page 1)

Tufekci, Z., 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 2, 12, and 20)

Tufekci, Z., 2018. Youtube, the great radicalizer. *The New York Times*, (2018). (cited on page 106)

twitter.com, 2013. New tweets per second record, and how! https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html. [Online; accessed 2020/08/22]. (cited on page 1)

twitter.com, 2020a. Filter realtime tweets. https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter. [Online; accessed 2020/06/01]. (cited on pages 11 and 77)

twitter.com, 2020b. Rate limit in Labs streaming endpoint. https://developer.twitter.com/en/docs/labs/filtered-stream/troubleshooting. [Online; accessed 2020/06/01]. (cited on pages 22 and 44)

twitter.com, 2020c. Sample realtime tweets. https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample. [Online; accessed 2020/06/01]. (cited on page 11)

twitter.com, 2020d. Search tweets. https://developer.twitter.com/en/docs/tweets/search/overview. [Online; accessed 2020/06/01]. (cited on page 11)

twitter.com, 2020e. Twitter rate limit notices. https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/streaming-message-types. [Online; accessed 2020/06/01]. (cited on pages 24 and 27)

twitter.com, 2020f. Twitter rate limits. https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits. [Online; accessed 2020/08/22]. (cited on page 2)

Valera, I. and Gomez-Rodriguez, M., 2015. Modeling adoption and usage of competing products. In *Proceedings of IEEE International Conference on Data Mining*. (cited on page 8)

Van Hentenryck, P.; Abeliuk, A.; Berbeglia, F.; Maldonado, F.; and Berbeglia, G., 2016. Aligning popularity and quality in online cultural markets. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 10 and 46)

Wagner, C.; Singer, P.; Karimi, F.; Pfeffer, J.; and Strohmaier, M., 2017. Sampling from social networks with attributes. In *Proceedings of the International Conference on World Wide Web*. (cited on page 12)

Wang, Y.; Callan, J.; and Zheng, B., 2015. Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Transactions on Web*, (2015). (cited on pages 11 and 34)

Weng, L.; Flammini, A.; Vespignani, A.; and Menczer, F., 2012. Competition among memes in a world with limited attention. *Scientific Reports*, (2012). (cited on pages 1 and 8)

wikipedia.com, 2020a. List of most-viewed YouTube videos. https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos. [Online; accessed 2020/06/01]. (cited on page 76)

wikipedia.com, 2020b. Vevo. https://en.wikipedia.org/wiki/Vevo. [Online; accessed 2020/06/01]. (cited on pages 50 and 76)

Wilhelm, M.; Ramanathan, A.; Bonomo, A.; Jain, S.; Chi, E. H.; and Gillenwater, J., 2018. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. (cited on page 17)

Wu, F. and Huberman, B. A., 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, (2007). (cited on pages 8 and 16)

Wu, S. and Resnick, P., 2021. Cross-partisan discussions on YouTube: Conservatives talk to liberals but liberals don't talk to conservatives. *Under review*, (2021). (cited on page 105)

Wu, S.; Rizoiu, M.-A.; and Xie, L., 2018. Beyond views: Measuring and predicting engagement in online videos. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 4, 34, and 77)

Wu, S.; Rizoiu, M.-A.; and Xie, L., 2019. Estimating attention flow in online video networks. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. (cited on pages 5 and 40)

Wu, S.; Rizoiu, M.-A.; and Xie, L., 2020. Variation across scales: Measurement fidelity under Twitter data sampling. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on page 4)

Yang, J. and Leskovec, J., 2011. Patterns of temporal variation in online media. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on page 57)

Yi, X.; Hong, L.; Zhong, E.; Liu, N. N.; and Rajan, S., 2014. Beyond clicks: Dwell time for personalization. In *Proceedings of the ACM Conference on Recommender Systems*. (cited on pages 2, 13, 52, and 74)

Yi, X.; Yang, J.; Hong, L.; Cheng, D. Z.; Heldt, L.; Kumthekar, A.; Zhao, Z.; Wei, L.; and Chi, E., 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the ACM Conference on Recommender Systems*. (cited on page 17)

youbube.com, 2017. Official YouTube blog: You know what's cool? a billion hours. https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html. [Online; accessed 2020/06/01]. (cited on pages 2 and 82)

youtube.com, 2012. YouTube now: why we focus on watch time. https://youtube-creators.googleblog.com/2012/08/youtube-now-why-we-focus-on-watch-time.html. [Online; accessed 2020/06/01]. (cited on pages 2, 13, and 65)

Yu, H.; Xie, L.; and Sanner, S., 2014. Twitter-driven YouTube views: Beyond individual influencers. In *Proceedings of the ACM International Conference on Multimedia*. (cited on pages 15 and 49)

Yu, H.; Xie, L.; and Sanner, S., 2015. The lifecyle of a YouTube video: Phases, content and popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*. (cited on pages 13, 15, 65, and 87)

Yue, Y.; Patel, R.; and Roehrig, H., 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the International Conference on World Wide Web*. (cited on page 10)

Zannettou, S.; Chatzis, S.; Papadamou, K.; and Sirivianos, M., 2018. The good, the bad and the bait: Detecting and characterizing clickbait on YouTube. In *Proceedings of the IEEE International Conference on Security and Privacy Workshops*. (cited on page 18)

Zarezade, A.; Khodadadi, A.; Farajtabar, M.; Rabiee, H. R.; and Zha, H., 2017. Correlated cascades: Compete or cooperate. In *Proceedings of the International AAAI Conference*. (cited on pages 1 and 93)

Zhang, J.; Ackerman, M. S.; and Adamic, L., 2007. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 40, 82, and 84)

Zhang, Y. C.; Séaghdha, D. Ó.; Quercia, D.; and Jambor, T., 2012. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. (cited on pages 17 and 73)

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W., 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*. (cited on page 20)

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J., 2015. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (cited on pages 16, 20, 41, 42, 44, and 67)

Zhou, R.; Khemmarat, S.; and Gao, L., 2010. The impact of YouTube recommendation system on video views. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*. (cited on pages 15, 17, 74, 86, and 98)

Zhu, L. and Laptev, N., 2017. Deep and confident prediction for time series at uber. In *Proceedings of the IEEE International Conference on Data Mining Workshops*. (cited on page 16)

Ziegler, C.-N.; McNee, S. M.; Konstan, J. A.; and Lausen, G., 2005. Improving recommendation lists through topic diversification. In *Proceedings of the International Conference on World Wide Web*. (cited on pages 87 and 102)