Understanding the Popularity Evolution of Online Media: A Case Study on YouTube Videos

Honglin Yu

August 2015

A thesis submitted for the degree of

Doctor of Philosophy

The Australian National University

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

Honglin Yu

August 2015

To my parents and parent-in-law who unreservedly give us their love

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor Dr. Lexing Xie who brings me to such a fruitful field and continuously supports my research greatly with motivation and patience. Her broad knowledge, sharp insights, strong ability to organize resources and dedication to research set an excellent example for me to follow. My sincere thanks also goes to my supervisor Dr. Scott Sanner for his precious advice on both of my research and thesis. His broad knowledge and extensive experience are very important in making my projects succeed. I also want to thank my supervisor Prof. Henry Gardner who carefully read my thesis and provide many constructive advice. I also feel grateful to his guidance and kind encouragement all along.

I am also very grateful to Dr. Xinhua Zhang who kindly teaches me many technical knowledge and shares his academic experience with me. I have also greatly benefited from discussions with other researchers such as Zichuan Xu, Changyou Chen, Dr. Marian Andrei Rizoiu, Suvash Sedhain, Trung Nguyen among others.

I also want to thank the CECS and NICTA's administration&IT staff, especially Elspeth Davies, Steve Marlor, Paul Semczuk, Peter Shevchenko and others for creating an efficient and friendly working environment for us.

I would like to thank Andrew Bell who has carefully proofread my thesis and helped correct the typos, grammar mistakes and make some sentences more fluent.

I dedicate this thesis to my little daughter Yueyang Yu who "forces" me to enjoy life and helps me avoid being "overfitted" on work. Last but not the least, I want to express my deep gratitude to my wife who encourages and helps me all the time. During the PhD, every time I feel upset and trapped, it is she who makes me confident and smile again.

Abstract

Understanding the popularity evolution of online media has become an important research topic. There are a number of key questions which have high scientific significance and wide practical relevance. In particular, what are the statistical characteristics of online user behaviors? What are the main factors that affect online collective attention? How can one predict the popularity of online content? Recently, researchers have tried to understand the way popularity evolves from both a theoretical and empirical perspective. A number of important insights have been gained: e.g., most videos obtain the majority of their viewcounts at the early stage after uploading; for videos having identical content, there is a strong "first-mover" advantage, so that early uploads have the most views; YouTube video viewcount dynamics strongly correlate with video quality. Building upon these insights, the main contributions of the thesis are: we proposed two new representations of viewcount dynamics. One is popularity scale where we represent each video's popularity by their relative viewcount ranks in a large scale dataset. The other is the popularity phase which models the rise and fall of video's daily viewcount overtime; We also proposed four computational tools. The first is an efficient viewcount phase detection algorithm which not only automatically determines the number of phases each video has, but also finds the phase parameters and boundaries. The second is a phase-aware viewcount prediction method which utilizes phase information to significantly improve the existing state-of-the-art method. The third is a phase-aware viewcount clustering method which can better capture "pulse patterns" in viewcount data. The fourth is a novel method of predicting viewcounts using external information from the Twitter network. Finally, this thesis sets out results from large-scale, longitudinal measurement study of YouTube video viewcount history, e.g. we find

videos with different popularity and categories have distinctive phase histories. And we also observed a non-trivial number of concave phases. Dynamics like this can not be explained in terms of existing models, and the terminology and tools introduced here have the potential to spark fresh analysis efforts and further research. In all, the methods and insights developed in the thesis improve our understanding of online collective attention. They also have considerable potential usage in online marketing, recommendation and information dissemination e.g., in emergency & natural disasters.

Publications

- Honglin Yu, Lexing Xie and Scott Sanner, 2015. The Lifecycle of a Youtube Video: Phases, Content and Popularity. *The International AAAI Conference on Web and Social Media*. (Full Paper, Acceptance Rate: 19%)
  (Chapter 4 and 5 and Section 6.1)
- Honglin Yu, Lexing Xie, Scott Sanner, 2014. Twitter-driven Youtube Views: Beyond Individual Influencers. *ACM Conference on Multimedia 2014*. (Acceptance Rate: 29.8%)
  (Chapter 7)
- New materials.
  (Chapter 3 and Section 6.2)
- Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck, 2015. Can this video be promoted? – Linking endogenous and exogenous processes to forecast popularity on Youtube. *In submission*.

Software Released

- YTCrawl: A fast YouTube video viewcount history crawler.
  URL: *https://github.com/yuhonglin/YTCrawl*
  (Chapter 3, 5 and 6)
- SegFit: A fast viewcount phase detection program written in C++.
  URL: *https://github.com/yuhonglin/segfit*
  (Chapter 4)
- ShotDetect: A powerful video shot detection program written in C++.
  URL: *https://github.com/yuhonglin/shotdetect* (1st year project)
- Lbfgsb.jl: A wrapper of the L-BFGS-B fortran routine in Julia language.
  URL: *https://github.com/yuhonglin/Lbfgsb.jl*
  (Chapter 4)

Dataset Released

– YouTube viewcount history dataset.

URL: https://github.com/yuhonglin/ytphasedata

(Chapter 5)

– Tweeting time stamps of a large set of YouTube videos.

ULR: http://bit.ly/1KWXR1C

(Chapter 3)

Media Coverage

– **ANU Reporter** *How the viral video star is born*. August 2015.

http://www.anu.edu.au/news/all-news/how-the-viral-video-star-is-born

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "The scarce, and therefore valuable, resource is now attention"
>
> — B. A. Huberman

This thesis studies the evolution of online popularity over time. In particular, we measure, describe and predict popularity that is centered around YouTube and the related online social networks.

In this chapter, I outline the reasons of why popularity, especially of online content, is an important research topic nowadays. Then I explain why YouTube is an ideal target for such research. Finally, I formally set out the research goals.

1.1 Why study popularity?

Popularity, which can be manifested in the forms like viewcount of online videos or click rate of hyperlinks, is a direct measure of people's aggregate attention on a given social media item. It is an important quantity for understanding many practical problems related to online media. For example, it is well known that, for web companies, popularity is closely related to revenue. For marketers, viewership is often believed to correlate with sales. For singers and movie makers, popularity of their video clips/trailers often implies success and income (Asur and Huberman [2010]). For internet service providers, understanding popularity can improve service by smart caching (Gummadi et al. [2003]; Cha et al. [2009]; Wang et al. [2012b]);

Even for individuals, people are often eager to make their post on the web popular. But making online items popular is not an easy task due to the "intense competition" nowadays. Take YouTube, for example: on average, 100 hours of videos are uploaded every minute (YouTube.com [2015]). This is how the communities of Web 2.0 work, where enormous content is active online and rapidly growing. Comparatively, people's attention has become increasingly scarce and therefore valuable (Huberman [2013]). This raises the question, how is this limited attention distributed among the multiple online content, in particular, YouTube videos? What factors can affect their popularity? How can one become successful in this "attraction competition"? These are the sorts of the important questions that this thesis tries to answer.

Popularity is also an important attribute for answering scientific questions about collective human behavior. By examining popularity data, we can understand how collective attention evolves and how an individual's behavior leads to collective attention. For example, we want to ask which item will become viral and will a celebrity's tweeting make something viral or not? Which characteristics of online users mostly reflect social influence? Exploring the social insights like these is also a major objective of our work.

Finally, there are significant computational challenges for understanding popularity. First, measuring large scale popularity dataset over time is difficult. Second, discovering suitable representations for popularity which are robust and computationally tractable is still a very active area of research. Third, the algorithms for extracting such representations and predicting future popularity are of great importance but there has not been a satisfactory solution.

1.2 Why YouTube?

Why do we focus on YouTube videos? First, YouTube is one of the largest *user-generated content* (UGC) websites. After its founding in 2005, it has accumulated a huge number of videos, metadata, and user interaction (e.g., comment, subscrip-

tion), and the data keeps growing faster and faster. It has been reported that in 2009, YouTube could already serve more than 100 million users per month and that 10 hours of videos are uploaded per minutes (Figueiredo et al. [2011]). In 2014, the number had increased to "1 billion unique user visits each month" and "100 hours of video are uploaded every minute". YouTube often rates among the top 3 most popular websites (Alexa.com [2015]). Besides the huge size, YouTube's data is also very diverse. It includes various kinds of videos (e.g., News, Comedy, Music, Animals, Travel) in 61 languages from 75 countries (YouTube.com [2015]). Being the nexus of large amounts of both user-generated content and people's attention, YouTube is an ideal place for popularity analysis.

A second reason we chose YouTube is that the data is available and open. Researchers have been using data from YouTube to study popularity (Figueiredo [2013]; Pinto et al. [2013]) and social networks (Cheng et al. [2008]; Wattenhofer et al. [2012]). There is an open API for researchers to efficiently download the meta-data of millions of videos. However, given the volume and variety of data, it remains a challenge to get the right data to answer the right scientific computational questions about popularity.

Third, YouTube has played an important role in different aspects of our society. Popularity on YouTube has great social impact. For example, pop stars like Justin Bieber and Katy Perry became extremely popular with its help. The 2008 US presidential election used YouTube as a platform to promote the candidates and interact with users. Their videos easily received millions of views[1], etc[2]. So studies on YouTube such as ours are well posed to actually make a difference in real social processes.

Lastly, studies of YouTube can potentially be generalized to other online social networks (OSNs) as well. This is because users of many other important OSNs, such as Facebook and Twitter, frequently discuss YouTube videos, so their effects on a

[1] http://scholarworks.umass.edu/jitpc2009/

[2] We refer readers to the wikipage http://en.wikipedia.org/wiki/YouTube for more about YouTube's social impact

video's popularity can be significant. For example, a celebrity's recommendation on Twitter can cause a YouTube video's viewcount to suddenly increase and even become viral (Allocca [2011]). The viewcount dynamics of certain videos can be successfully explained by models built based on word-of-mouth propagation in a social network (Crane et al. [2008]); Yu et al. [2014] also showed that the early viewcount of a YouTube video can be better predicted using its related Twitter feed. Studies like these have improved people's knowledge of both YouTube and other OSNs. This means that the significance of research on YouTube goes beyond understanding a single website.

1.3 Main research goals

This thesis tries to address the scientific and computational challenges of understanding popularity from four perspectives.

- Gathering the right data and obtaining insights from large-scale measurements. This includes the measurements of both YouTube video viewcount history itself and related interactions between YouTube and other OSNs.

- Constructing appropriate representations for how popularity evolves over time and developing robust algorithms for extracting such representations. These algorithms must be efficient and can handle viewcount data from a large video collection.

- Devising algorithms to predict the future popularity of a video given its history, temporal evolution, and related user activities from other OSNs.

- With these representations and computational tools, we would like to understand how popularity evolves. For example, what does a typical videos' lifecycle look like and how does popularity relate to user behaviors within YouTube and other OSNs?

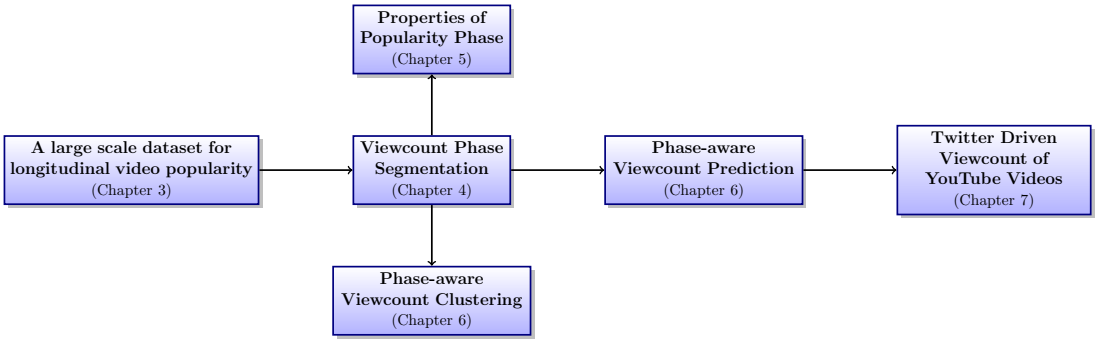


Figure 1.1: The road map of this thesis.

Figure 1.1 presents an overview of the high level structure of this thesis. First, we collect a large-scale dataset on longitudinal popularity of 880,000 videos and carry out a measurement study (Chapter 3). Then to handle the temporal complexity of video lifecycles, we propose a new viewcount phase representation and an efficient algorithm to detect those phases (Chapter 4). By another large scale measurement study of phases (Chapter 5), we obtained a number of new observations and insights about the evolution of video popularity. Then using phases as features, we propose a new viewcount prediction method and a new viewcount clustering method (Chapter 6). Finally, we set ourselves the task of predicting, based on Twitter feeds, sudden increases in viewcounts and a video's early popularity (Chapter 7).

1.4 Summary of contributions

The main contributions of this thesis which address the research goals mentioned above are now summarized.

1.4.1 Measurements study of the evolution of YouTube video popularity

We have collected and analysed a large and unique dataset of YouTube video viewcount for over 4 years and observed novel characteristics. Notably, we find that a video's popularity can be well represented by its popularity percentiles, and that

many videos exhibit multiple rising-and-falling popularity stages. Moreover, the variance of a viewcount series is not constant over time. On Twitter, old "Music" and "Comedy" videos are more likely to be discussed than old "News" and "Games" videos. Many videos' viewcounts have clear weekly or yearly periodicity. Finally, it has been observed that there are strong correlations between increases in a video's viewcount and external events or Twitter users' discussion. More details on each of these findings are given in Chapter 3.

Main contributions:

1. This thesis documents many new observations of popularity evolution based on a unique YouTube video viewcount datasets, such as the clear weekly or yearly periodicity of many viewcount series and the strong correlations between increases in video viewcounts and corresponding tweets.

1.4.2 Modeling viewcount dynamics and phase detection

How does the popularity of online content evolve over time? This has been a long-standing question in the area of online popularity analysis. Previous studies have tried to address this problem from many perspectives, such as statistical profiling (Cha et al. [2007,?]; Figueiredo et al. [2011]), clustering (Yang and Leskovec [2011]; Figueiredo et al. [2014]) and mathematical modelling (Sornette and Helmstetter [2003]; Crane and Sornette [2008]). But all those studies underestimate the temporal complexity of popularity dynamics: they all assume that the popularity of each video only goes through one peak or follows a single model at all times. But as mentioned above, in reality many videos (especially popular ones) go through many rising-and-falling popularity stages over time. So based on previous research, we propose a new way to represent popularity phase and an efficient viewcount phase detection algorithm. The algorithm automatically determines the number of phases a video goes through, and also calculates the phase parameters and boundaries. Details of the representation and algorithm are found in Chapter 4.

Main contributions:

1. This thesis contains a new viewcount phase representation which is more suitable to describe the rising-and-falling stages of online popularity evolution.
2. We have also proposed an efficient algorithm to detect the phases.

1.4.3 Observations of popularity evolution based on phases

Through another large-scale measurement study based on popularity phases, we have discovered a number of new features of how YouTube video popularity evolves. For example, the phase statistics strongly correlates strongly with a video's popularity scale and category; the phase changes may imply strong endogenous fluctuations in the social system or exogenous interventions. Moreover, phases can also help us understand phenomena like "viewcount revival"; with the help of phases, we can now more closely observe how videos go viral. In general, viewcount phase appears to be an important concept and a very effective tool. It enables us to understand the popularity evolution of online content from a new perspective. Details are given in Chapter 5.

Main contributions:

1. We have found that videos' phase properties correlate strongly with their popularity and user-assigned categories.
2. With the help of phases, new observations relating to important research questions about online popularity research emerge. How do video viewcounts evolve over time? What are the differences among videos' lifecycles with respect to different popularity and categories? And How do videos go viral?

1.4.4 Viewcount prediction and clustering

This thesis also includes two new viewcount prediction methods and one new viewcount clustering method. The first new prediction algorithm is based phase infor-

mation. Its performance significantly outperforms the baseline method (for details, see Chapter 6). The second prediction method leverages Twitter information to successfully predict sudden increases in videos' viewcount and videos' early popularity near upload (see Chapter 7). By comparing the predicting power of different Twitter features, we have found new observations – notably that, it takes a diverse set of Twitter users, and not just the most influential users, to cause wide-spread activity in social media.

Since the viewcount phases are good tools to describe and summarize popularity dynamics, we have devised a phase-aware viewcount clustering method. It is efficient at capturing pulse-patterns in viewcount data (rather than just increasing or decreasing trend, which the previous methods are limited to). Details on the clustering method are given in Chapter 6.

Main contributions:

1. We have proposed a phase-aware viewcount prediction algorithm which significantly outperforms the baseline method.

2. We have successfully utilized Twitter information to predict YouTube video viewcounts.

3. We construct a new time-series clustering method based on phases. Compared with existing methods, it is much better at capturing pulse-patterns in popularity data rather than just increasing or decreasing trends.

Chapter 2

Related Work

This thesis relates to several active research areas. In this chapter, the prior work is grouped into 3 main sections. First, measurement studies on online media popularity are reviewed in Section 2.1. In Section 2.2, related research in developing popularity representations is discussed. Lastly, previous studies on online popularity prediction are reviewed in Section 2.3.

2.1 Measurement study of online media popularity

The first type of prior research of this thesis relates to large-scale measurement studies on popularity. The main challenges facing such studies are first how to collect the right dataset and then how to apply descriptive statistics and data visualization techniques to reveal meaningful insights. Concerning YouTube, an important and pioneering study was first done by Cha et al. [2007]. They analyzed a large number of videos from VoD websites e.g., YouTube and Daum. They found that the shape of the distribution of video popularity depended on the category of videos. All the distributions followed "power-law" in the middle, but demonstrated cutoffs at the heads and tails. Based on users' "fetch-at-most-once" behavior first proposed by Gummadi et al. [2003], the authors gave an explanation of the cutoff in the heads. They also discussed the possible reasons for the cutoff in the tails, e.g., sampling bias or recommendation engines. The authors went on to analyze the evolution of video popularity over time. They observed an "ephemeral popularity of young videos", and noted "if we exclude the very new videos, user's preference seems relatively

insensitive to video's age". This is consistent with our observation of possible phase transitions over time.

In the early years, the reason some researchers analyzed YouTube viewing patterns was to help relieve broadband pressure, but their works is also a good source for video popularity analysis. Gill et al. [2007] monitored video traffic in the network of the University of Calgary as well as viewcount traces of top videos on YouTube. By comparing the local and global usage patterns, they found that YouTube traffic varies significantly by time-of-day and day-of-week. In another study, Cheng et al. [2008] assembled data on more than 3 millions videos and derived considerable knowledge about YouTube network traffic. (Here I only list their insights related to video popularity.) They found that a video's viewcount is better fitted by power-laws rather than linear functions. According to their fitting result, 70% of the videos have growth trend factors less than 1, which means the viewcount increasing rate of most videos slowly declines with time. They also found that most videos are only frequently watched in a short "active life span". This is similar to the "passing fad" observation of Cha et al. [2007].

Chatzopoulou et al. [2010] collected the meta data on 37 millions YouTube videos (estimated to be 25% of all the YouTube videos at that time) and analyzed the correlation between a video's viewcount and other metadata (#comments, ratings, #favorites). They found strong linear correlations between a video's viewcount, #comments and #favorites, and the correlation became stronger when popular videos were considered. More interestingly, "ratings" barely correlated with other measures[1]. The authors also examined the video uploading frequencies over time and found the daily uploading peaks usually occurred at 1 PM, while the weekly peaks usually occurred on Sunday for most video categories.

Figueiredo et al. [2011] have compared the growth pattern of YouTube video popularity on three video datasets (deleted videos, top videos and random queries). By

[1] I also observed this in our experiment. Extreme examples are Justin Bieber's music videos which are extremely popular but very lowly rated.

comparing the patterns in popularity growth of the three datasets, the authors found that the deleted videos often received most of their views early on. Videos from the top list of YouTube often have large viewcount peaks; in comparison, viewcounts of videos in the other two sets often have multiple smaller peaks. By analyzing different referrers to videos, they found YouTube's search/recommendation mechanisms have a great effect on a video's popularity — specifically, a search engine is most influential for the random dataset whereas YouTube's recommendation engine has more effect on removed and top videos. This work has shown that the popularity profile of different types of videos is heterogeneous, and suggests that there is probably a complex taxonomy underlying YouTube videos.

Borghol et al. [2011] assembled a large dataset of videos that had been newly uploaded, an approach which is generally considered the most unbiased way of sampling. They found a strong non-stationary characteristic in viewcount. This is primarily because, after uploading, the time which a video takes to reach a viewcount peak differs very much from video to video. Another reason is that there are large fluctuations in popularity as a video evolves. Surprisingly, they also found that the viewcount at the peak is independent of the time it takes to reach the peak. They also constructed a model to generate synthetic viewcounts which had the same statistical properties as the real data.

YouTube is not only a video sharing website but also a social network where users can interact with each other through subscription or commenting. Besides video contents, the research described in the following also takes user characteristics into consideration. Wattenhofer et al. [2012] analyzed YouTube user-level statistics and investigated their relationship with content popularity. They looked at the subscription graph, which represents "social activities", and the comment graph, which represents "content activities". They found, as a typical "content-driven" OSN, the subscription graph and the comment graph had similar scales (in nodes and edges) but barely overlapped, demonstrating there is a dichotomy of user interaction and subscription behaviors on YouTube. They also found the subscription graph had low

homophily/reciprocity, which is different from traditional OSN but similar to other content-driven networks like Twitter. Another feature of this research is that it was based on the entire set of YouTube data. This means that if one does research based on a sub-sample of YouTube, it is possible to compare the corresponding statistical plots and see whether the data is strongly biased or not. In a related paper, Broxton et al. [2013] also studied how social factors affected the popularity of YouTube videos. They also considered many other social networks and weblogs, and found that the patterns of viewcount evolution differed greatly between social videos and non-social videos (classified by whether or not they are shared in some online social networks). For example, the rise-and-fall of non-social videos was much slower than of social videos. They also ranked 25 websites by their ability to propagate viral YouTube videos.

Besides factors such as video content and social networks, *geographic* features also play important roles in video popularity evolution. An earlier work by Zink et al. [2008] analyzed the YouTube traffic in a university campus network. Among their many observations, they found that there was only a small correlation between the popularity of global and local videos. Many users watch the same video more than once. Neither trace duration nor user population seem to have an influence on local popularity distributions. And videos with local interest usually have high local popularity. Brodersen et al. [2012] analyzed more than 20 million videos uploaded in one year from different regions. They found that more than half of YouTube videos received more than 70% of their views from a single country. However, social sharing can widen a video's geographic reach. Even more interesting, the reach of a video often goes through an "expansion-contraction" process — first the focus is on one region with occasional views from other regions, then the focus shifts back to the main region. Correspondingly, after uploading, the viewcount often immediately increases and then gradually fades out. This research demonstrates the importance of considering geographic factors in analysing video popularity or designing recommendation engines.

Besides YouTube, Mitra et al. [2011] analyzed large-scale data from four video sharing web services: Dailymotion, Yahoo! video, Veoh and Metacafe. The main contribution of this work was to find seven common characteristics among the different services: 1) The social behaviours of users are less frequent than watching videos; 2) The number of each user's uploads follows Pareto's rule and the number of uploaders is one order of magnitude smaller than the number of uploaded videos; 3) Most videos are of short duration; 4) The distribution of videos' views follows Pareto rule: 20% of the most popular videos accounting for $\geq$ 80% views; 5) The popularity distributions are heavy-tailed and can be modelled by a power-law with cut-off; 6) The distribution of viewing rate popularity also follows a power-law; 7) When considering total views, "one-timer" videos are less than traditional media server workloads, whereas when it is defined by fixed period, they are comparable. The authors also found some difference among the services: for example, they found the users of Veoh are more likely to upload multiple times than users of Yahoo!. This research is very comprehensive and makes a good point of reference for any YouTube video study.

In a recent work, Abisheva et al. [2014] have done a pioneering and comprehensive research on how to utilize Twitter information to understand YouTube viewership. Based on a high quality dataset, they have obtained many insights into the watching and sharing patterns of YouTube videos on Twitter. For example, they designed a method to identify promotional Twitter accounts by looking at video sharing speed. They also applied a set of heuristics to predict user demographic profiles. Among all video categories, "News & Politics" videos involved the most social and sharing behaviors. They also proposed a simple regression model for predicting a video's final viewcount based on early Twitter sharing information. They found that retweet rates are much more predicative than the number of followers a user has. This research somewhat overlaps with the study in Chapter 7, although it must be said that both pieces of research were done independently. Moreover, the study in Chapter 7 is much more comprehensive in terms of adding predictive ability, and

our models/features are much more complex. It is perhaps the nature of research, but the insights we report are distinctly different to those of Abisheva. In general, researchers have already tried for some time to analyze the interactions between different OSNs and use them for popularity prediction, and thesis thesis is part of an evolving branch of research.

There are also many studies (Benevenuto et al. [2009], Chatzopoulou et al. [2010] etc.) on the effect of "video responses" (which used to be a prominent "social feature" of YouTube) on popularity. But since this feature was in fact very little used by users, it was removed in early 2014[2], so this research is not reviewed in detail here.

Summary In this section, we have reviewed the previous measurement studies of online popularity. In comparison with the prior works, the novelties of this thesis are as follows. First of all, I have collected a unique dataset of longitudinal YouTube video popularity history; second, I have analysed user behaviors, both of individuals and in aggregate, of two major OSN platforms (Twitter and YouTube). Further more, we have found that a video's popularity can be well described by popularity scales.

2.2 Online popularity representation and modeling

Although the measurement studies mentioned above have been very successful in improving our knowledge of online popularity, laying the foundations for further research, important questions such as how popularity evolves over time and how videos go viral are still left largely unanswered. This is mainly because the prior studies have mostly used only *scalars* (like the total views of a video or the number of hashtag adoptions at some time) to measure popularity. In this thesis, one of our main contributions is that we propose a new popularity representation called popularity phases which can describe popularity *dynamics* over some time range. The algorithm we propose to detect the phases is basically a kind of *time series segmentation* algorithm. And the way we describe each popularity phase originates from the

[2]See this announcement: http://bit.ly/1viWotn

prior research on popularity dynamics modeling. In this section, previous studies in these two areas (in Section 2.2.1 and 2.2.2) are reviewed.

In addition, I also proposed, in this thesis, a new time series clustering algorithm suitable for summarizing viewcount data which provides a separate representation for each cluster. Accordingly, previous studies on the clustering online popularity data are briefly discussed in Section 2.2.3.

2.2.1 Time series segmentation

In the data-mining community, time series segmentation is an old but still active research field. To the writer's best of knowledge, the problem of approximating a smooth nonlinear functions by a predefined number of linear segments was first proposed by Stone [1961] and was solved by Bellman [1961] with dynamic programming. After this, time series segmentation was mostly treated as an alternative way (other than Fourier Transform or wavelet) of representing and compressing time series data (Lin et al. [2003]) and has been used as a preprocessing procedure in time series data-mining. To fulfill the requirements of different undertakings (in e.g., computer vision research, financial data mining), many new segmentation algorithms were proposed (Fu [2011]).

In general, time series segmentation methods can be classified from a number of perspectives, e.g.,

- Whether continuity at the break-point is guaranteed
- How to describe each segment, e.g., a constant function, linear function (which can be again classified into "interpolation" or "regression"), or a polynomial etc.
- How to determine the boundaries, e.g., balancing residual, minimizing sum of error norms etc.
- How to determine the number of segments, e.g., predefined or using some heuristics

- How to optimize the loss function – different algorithms can be used, e.g., greedy algorithms, dynamic programming and discrete optimization etc.
- Online methods or batch methods

Making a full review of all the types of segmentation algorithms is beyond the scope of this thesis. For the purposes of this research, which is to consider the dynamics of YouTube viewcount data, we only need to use summarize the existing methods that does not require the continuity at the break points. This rules out many classical methods.

In the years, Pavlidis [1973] proposed a fast algorithm based on a discrete optimization method. L_∞ norm was used to measure the fit of a uniform linear function to each segment where the boundaries of between segments were determined by "balancing the residuals" in fitting each segment. The authors justified their method by eyeballing some examples. Although the algorithm is very fast, its limitations, as an early piece of pioneering research, are that, 1) the number of segments must be set before hand; 2) a uniform linear function is too simple to capture the shape of phases; 3) most importantly, "balancing residual" is not suitable for time series which do not have identical variance [3]. These limitations make it unsuitable for YouTube viewcount analysis.

Later, researchers in this field gradually discarded "strictly balancing residuals" and instead, used a "softer" heuristic like penalizing the standard deviation of residuals (Keogh [1997])). The algorithms used have mainly converged to three types: "sliding window", "bottom-up" and "top-down" (Keogh et al. [2001, 2004],Fu [2011]).

Although the goal of the paper by Keogh [1997] was to do fast time series similarity searching, the authors proposed a new time series segmentation method in the preprocessing stage, which plays an important role in their whole method. Their

[3] Quotation from Keogh [1997] on this point: "... given two time series, A and B, where B is simply A plus noise, it will produce two segmentations, similar in shape, but with the segmentation representing B containing far too many segments. Ideally in this situation we would like the algorithm to produce identical segmentations."

algorithm first segments the series into as many segments as possible. Then it gradually merges neighboring segments by minimizing the standard deviation (*std*) of the fitting errors among segments. They also show, by example, that the *std* with respective to the total number of segments has one global minimum point. The novelty of this method is that it can determine the number of segments automatically and minimizes the *std* of residuals. However the authors still use linear functions in their approach and their method is still greedy in nature (it can not guarantee a global optimum), which makes it unsuitable for our purposes.

In the papers of Keogh et al. [2001, 2004], the authors first make a good survey of existing time series segmentation approaches and summarize them into three categories, sliding window, bottom-up and top-down. By their definitions, these are three types of greedy algorithms with three different schemes. The "sliding window" method begins fitting a phase ranging from the first (earliest) data point one by one until the fitting error exceeds some threshold. It is an online algorithm. The "top down" method gradually "cuts" the sequence until the fitting error of each segment is below some threshold. The position of every cut is chosen to mostly reduce the total fitting error. The "bottom up" method first splits the sequence into a series of short segments and then applies "lowest cost merging" on adjacent ones until some stopping criterion is met. In all these methods, linear regression is used to fit the segments. The authors compared the three strategies on ten datasets with various parameters and found that, in average, "bottom up" method performed the best. Then the authors proposed a new online algorithm combining "sliding window" and "bottom up". Further experiment showed that their method can easily performed as well as "bottom up" method. Although this paper can be seen as a milestone in time series segmentation research, the algorithms it uses are, due to their greedy nature, not suitable for video viewcount research since we want the segmentation result to be stable and the boundaries between segments as accurate as possible.

Terzi and Tsaparas [2006] proposed an "divide and segment" algorithm to approximate the results of dynamic programming. Rather than apply dynamic pro-

gramming directly to the whole sequence, the most basic form of their algorithm was to first divide the sequence into a few disjoint intervals. They then used dynamic programming on each interval. By assuming that each segment can be treated as one data point, the segments found on all the intervals form a new sequence weighted by the length of corresponding segment. Dynamic programming is then applied on the new sequence, which will output the final result. It can be proved that the fitting error of their algorithm is no more than 3 times that of the optimal segmentation. The author also investigated some variants of the basic algorithm and demonstrated a trade-off between an algorithm's speed and performance. By experiments with synthetic data, the authors showed that their method outperforms classic bottom up algorithms and randomized algorithms proposed by Himberg et al. [2001]. They also found that the real fitting error was far below the theoretical bounds. This work is a pioneering work on using the idea of "divide-and-conquer" to obtain speedup. But their method applies only to the cases that the number of segments is predetermined which makes it not suitable for our purposes.

Some other research has explored the usage of segmentation in time series data-mining. Das et al. [1998] proposed methods, based on segments, to discover rules in time series data based on segments. Although their method is too simple to be called a time series segmentation algorithm, it demonstrates an important way of using segmentation in time series data mining. In another work, based on segmentation, Keogh and Pazzani [1998] proposed a time series representation method which help applies classic machine learning techniques to time series data. The authors point out that, the main difficulties of time series data mining is that 1) the data is often of high dimensionality; and 2) it is hard to define a similarity measure that captures human's preferences. In this context, segmentation can be seen as a way of compressing time series data and generating high-level features (segments). The authors then proposed a segmentation method that used linear functions to fit each phase using a certain weight. They also define a new similarity measure based on segments. Their experiments showed that, in a classic clustering task, the training speed can be greatly

improved over working directly on the original data, and the results are similar. This work demonstrated the power of segmentation in time series data mining.

Based on time series segmentation, Lin et al. [2003] proposed a method of transforming real-valued time series data into symbols. The procedure is not complex, requiring only that the series first be normalized, and then *piecewise linear/constant segmentation* is applied to the data. The series is turned into symbols by discretizing the segments. But what is important is that they also prove that the distance measures defined on their symbolized series provides lower bounds on the distances computed from the original data. This provides a theoretical guarantee that for most time series data-mining tasks, one can work directly on low-dimensional symbolized segments to get identical results but with a huge speedup. The authors justify their methods with experiments of time series classification and clustering on many datasets. Although the authors still used simple linear functions to fit each segment, this work can be seen as a milestone in integrating time series segmentation and machine learning techniques. This paper also includes a good review of time series representations.

Many other research efforts based on time series segmentation have been done. (As pointed out in Keogh et al. [2001], researchers in many different fields invent their own segmentation algorithms for different purposes). To keep it concise, we refer readers to two good review papers (Fu [2011]; Esling and Agon [2012]) that have been published recently on time series data-mining.

The last thing that must be mentioned in this context is that, empirical *evaluation* and *comparison* of time series algorithms have always been difficult (Keogh and Kasetty [2003]). This is partially due to the wide diversity of researchers' purposes and datasets. For example, as said above, early researchers mostly justified their methods by eyeballing; Later, some researchers working on data compression looked to segmentation to help speeding up post-processing (such as clustering and classification) while keeping the performance at a level that does not fall much below that from working on the original data. In our case, we want the segmentation to suc-

cessfully capture the phases of video viewcount evolution. Because our purpose is different from other research, we developed our own algorithm and used "systematic eyeballing" for parameter tuning and performance evaluation. By using a number of examples, we will also show that our methods outperforms existing approaches in modeling viewcount dynamics.

2.2.2 Popularity dynamics modeling

Beside large efforts on measuring video popularity dynamics, previous research also tried to answer one important question: "what is the underlying mechanism generating the dynamics of video popularity?" D. Sornette *et al.*have published a series of papers addressing this problem. Their work put the problem into a more general perspective: modelling the "social system". They assumed that the popularity or sales of online content like videos on YouTube or books on Amazon are the output of the social systems. More specifically, Sornette and Helmstetter [2003] assumed the social system as a linear system to be with a power-law memory kernel function (an *impulse response function*). With an extra assumption that the input (or "fluctuations") of the system is a Gaussian process, they deduced that the expectation of viewcount peaks can have two kinds of shapes and named them as endogenous and exogenous peaks. They also claimed that a social system can also be modelled by *branching processes*. Branching processes are classical tools used to model phenomena like species reproduction or information propagation. In an earlier work on modelling earthquakes, Sornette and Sornette [1999] proved that the asymptotic responses of a branching process with power-law kernel functions are themselves power-laws. In this way, they deduced, from two different sets of theoretical assumptions, the same form of the expectation of popularity dynamics around peaks. The authors also showed some simulated results of their model. Afterwards, a number of studies have been done to provide empirical evidence for their theories.

Crane and Sornette [2008] investigated about 5 million YouTube videos. Based on the theorem above, they proposed four categories of phase shapes, namely "ex-

ogenous sub-critical", "exogenous critical", "endogenous critical" and "endogenous sub-critical". Except "endogenous sub-critical" which is roughly random, all the other three types can be described by power-laws. Among their data, they found that 90% of videos' viewcounts are random noise, which leaves 10% of the total videos to be classified into the other three types and fitted with power-laws. This pioneering work has provided evidence that "collective human dynamics can be robustly classified by epidemic models". But the limitation is, for simplicity, the authors only assumed that there was at most one peak in each video, which is far not true in reality. Based on this work, we propose a new time series segmentation algorithm to deal with cases in which there may be multiple peaks. In another paper, Crane et al. [2008] gives evidence showing that the "shape" of peaks has power to distinct viral, quality and junk videos. They also discussed the possible reasons and the underlying propagation processes.

Concerning popularity spike patterns, existing theoretical studies distinguish four types (Crane and Sornette [2008]) while empirical research finds six types (Yang and Leskovec [2011]). Matsubara et al. [2012] proposed a unifying model to explain all of them. Their model, in the basic form, can be seen as a combination of the classic "susceptible-infected" (SI) model by Bass [1969] and the "self-excited Hawkes process" proposed by Crane and Sornette [2008]. The full model, SpikeM, further takes periodicity into consideration. By experiments, they showed that their model can 1) accurately model the clusters found by the K-Spectral Centroid (KSC) algorithm (Yang and Leskovec [2011]); 2) predict the falling phase better than the auto regression model; 3) estimate the susceptibility of items and the size of the potential user space. Their work is a significant improvement on popularity peak modeling. However, the authors still assume there is only one peak in each popularity series.

2.2.3 Viewcount clustering

Yang and Leskovec [2011] have analyzed the popularity evolution of hashtags and short text phrases in 589 million tweets and 170 million blog posts/news media arti-

cles. They proposed a new time series clustering algorithm called K-spectral centroid (KSC) whose clustering result is invariant to time series scaling and shifting. The authors showed that KSC outperforms the classic K-means clustering method in finding distinct shapes for popularity time series. They also found that the evolution of online content attention followed 6 main shapes. Their research is a pioneering work on empirically discovering the shapes of popularity time series of online content, shapes which may be directly related to underlying processes of human interactions (Sornette and Helmstetter [2003]). The KSC algorithm has been influential and used in many ensuing works (e.g., Figueiredo [2013]). But since KSC still minimizes a variant of L_2 loss among time series, it tends to produce smooth clusters and does not capture short-term shocks.

Instead of predicting actual popularity, Figueiredo [2013] tried to predict long term trends of YouTube videos. Using the KSC clustering algorithm, he found 4 types of trends (Yang and Leskovec [2011]) and used rich features, including video categories, link features, and popularity features, to predict which trend cluster a video will belong to. His work seems to be still ongoing, and it has been reported he wants to work out what kind of predictions are needed by UGC companies to increase revenue, not just the number of webpage hits.

Summary In this section, prior studies on time series segmentation has been reviewed. In this context, the novelty of this thesis is that we devise a global optimal segmentation and fit algorithm for non-linear monotonic phases. Previous studies on popularity dynamics modeling have also been discussed. Here, the main improvement of this thesis is that we recognize that a video can have multiple phases in its lifecycle; Lastly, I have mentioned previous studies on online popularity clustering. In this thesis, a new clustering algorithm is proposed that not just favors smooth temporal trends.

2.3 Online popularity prediction

YouTube video viewcount prediction is an active research topics recently. It has great practical values, such as estimating YouTube revenue, advertisement deployment and video pre-caching etc. The pioneering study on this topic was done by Szabo and Huberman [2010]. They observed that there is a strong linear correlation between a YouTube video's viewcount in the first week and the first month. Based on this, they successfully built a simple linear viewcount prediction model using the former to predict the latter.

Pinto et al. [2013] improved the above method. They proposed two new methods. One is multiple linear regression (MLR) with daily viewcount history as features. In the other, they randomly picked some videos' viewcounts from the training set as *centers* and used the *distance* of other videos' viewcounts to the centers as features. By comparing their methods and the baseline on two datasets (one consisting of popular videos and the other randomly sampled), they showed their methods both performed better than that of in Szabo and Huberman [2010].

In the same year, Ahmed et al. [2013] analyzed large-scale of data from not only YouTube but also Digg and VIMEO. They found that content can be classified by its popularity evolution patterns. They have also proposed a novel popularity prediction method and showed that it outperformed baseline method (i.e., K-means regression).

Intuitively, a video's content should be very important in determining its popularity. But in nearly all the existing research, video content is treated as noise and not considered. Borghol et al. [2012] propose a clever way to circumvent specific content and more rigorously analyze content-agnostic factors – they only considered groups of the near-duplicate videos (clone videos). By analyzing the correlation between features and using PCA, two sets of correlated features are found, namely features related to past popularity and features related to uploader properties. They then applied a classical linear regression model to the clone video sets and found that in each set, total former viewcounts and videos' ages are most predictive of view-

counts in the next week. More importantly, by comparing prediction results on each clone set and that over all videos (encoding clone set ID as content features), they found it was indeed important to consider a video's content in predicting its popularity (although former viewcount was still the most predictive feature). Besides viewcount prediction, the authors also demonstrated using clone sets that there was a strong *rich-get-richer* effect in viewcount evolution. They also found that the predictive power of features may change with time, e.g., uploader characteristics is usually the strongest predictor of a video's early life but can become superseded by former viewcount at the half-year point. In all, this work is comprehensive and informative; although it does not provide an answer of how to make use of video content, it does make people aware of its importance.

In the following, I will review some closely related studies predicting the popularity of online content other than YouTube. Li et al. [2013] analyzed a large number of YouKu (the largest video sharing website in China) videos shared in other OSNs. They found classical models like ARIMA, MLR or kNN were ineffective in modeling viewcount dynamics. Then a dynamic model based on video propagation was proposed and the prediction result based on it significantly outperformed classical methods.

Yang et al. [2012] carried out a large-scale empirical research, based on hashtags, on Twitter user behaviors. By correlation analysis, regression analysis, and prediction analysis, they revealed the dual functionality of a hashtag: it serves as both a tag of content and a symbol of membership of an online community. More importantly, the authors proposed many effective features in predicting hashtag adoption. These ideas greatly have inspired our research on using Twitter information to predict YouTube viewcounts.

In a recent work, Cheng et al. [2014] examined the problem of predicting growth of cascades in social networks. They redefined the cascade prediction problem and applied their method to a complete photo-resharing dataset from Facebook. Their results were encouraging in that cascade growth can be effectively predicted and the

performance of different classes of features (time, graph structure and individual) were similar. They also found that cascades with fast early reshares are more likely to grow significantly and that propagation breadth on the social network is more predicative than depth. However, the most significant features may change if the size of the cascades becomes large (specifically, the features of content and original author then become less important). The authors also tried to predict the eventual structure of cascades and their method significantly outperformed the baseline. Thus, this work is very comprehensive; the way they formulate the problem and design features greatly assists popularity research on YouTube videos.

Summary In this section, previous studies on popularity prediction have been reviewed. This thesis documents two new viewcount prediction methods. One is to utilize phase information to better predict future video viewcounts. The other is to predict a video's early viewcount and sudden viewcount increases using Twitter feed.

Chapter 3

A Large Scale Dataset for Longitudinal Video Popularity

In this chapter, we describe our tweeted video dataset. It is a unique resource containing 880,000 diverse YouTube videos with their complete daily viewcount history over 4 years. We first describe how this dataset was collected, and then present a series of observations on how the popularity is distributed and how it changes over time. In particular, in our dataset we observed that, except for the most popular and unpopular videos, the popularity of videos is distributed exponentially over their relative rankings; some of the old videos, especially "Muisc" and "Comedy", are more likely to be tweeted than old "News" and "Games" videos; on average, most videos received most of their viewcount over the first 10 days; the noise level of viewcount series is not homogeneous over time; some of the videos' viewcounts clearly represent weekly periodicity or seasonality; and certain external events can have strong interventions on viewcount dynamics and cause sudden increases. In general, the observations in this chapter not only give the reader a feel for the diverse and complex characteristics of viewcount dynamics, but also provide a worthwhile expansion of existing measurement studies on this topic.

3.1 YouTube viewcount history datasets

Before mid-2013, one could only retrieve 100 evenly distributed data points of a video's viewcount history, no matter how long the video had been uploaded. After that, YouTube started to publish complete daily viewcount traces of most videos on their website (see Figure 3.1). Given a video's ID, I wrote a powerful crawler to efficiently download the data[1] when uploader makes it public.

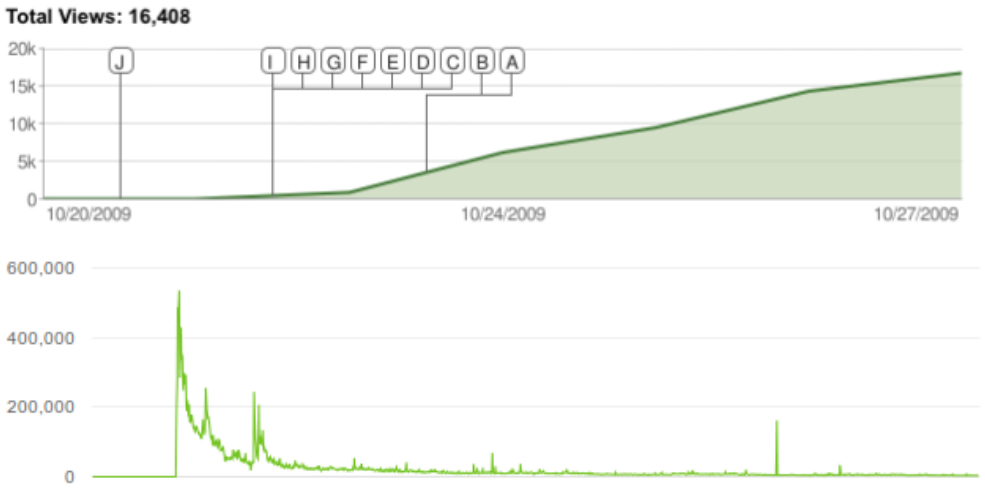


Figure 3.1: Top: The YouTube video viewcount history before mid-2013. It is based on 100 data points; Below: The new version which has daily viewcounts available from the time of upload.

We extracted video IDs from a large Twitter feed (Yang and Leskovec [2011]) of 467 million tweets from June 1st to December 31st in 2009, roughly 20−30% of total tweets in this period. We extracted URLs from all tweets and resolved the shortened URLs, retaining those referring to YouTube videos. This yielded 2.4 million unique YouTube videos, among which 1.5 million are still publicly online. We removed videos that had less than 50 views in their first 2 years (that is, not enough views to meaningfully extract phases). Our final dataset includes 0.88 million videos with fully available metadata. Figure 3.2 shows how many video IDs etc. we found for

[1] See https://github.com/yuhonglin/YTCrawl. This crawler actually disguises as a "browser" and downloads the AJAX response of video statistics on the website.

each day.

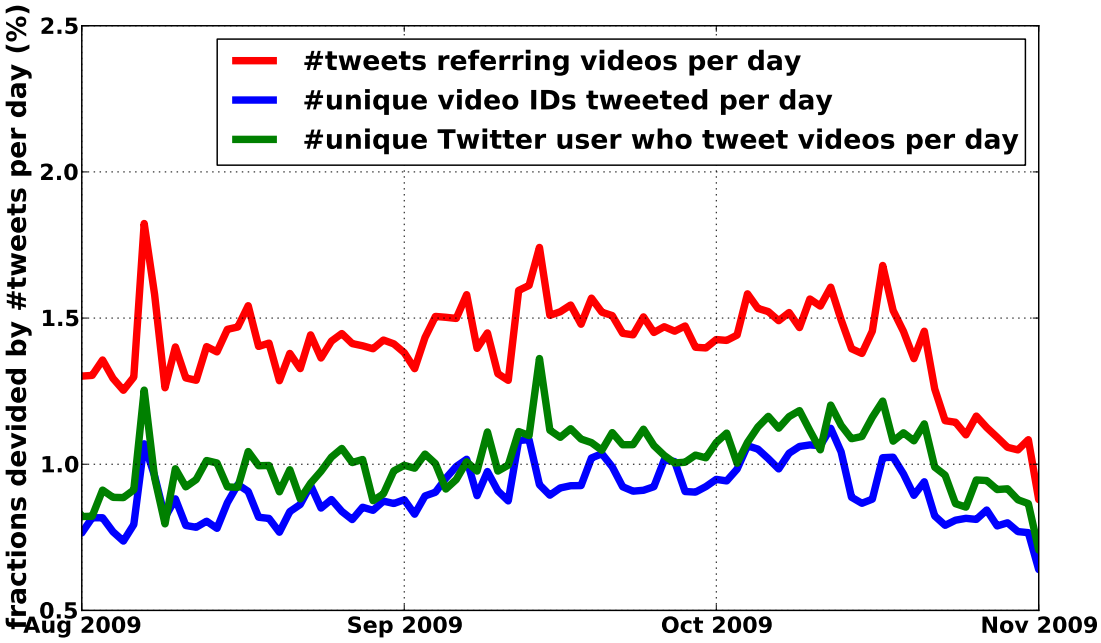


Figure 3.2: Red curve: The number of tweets containing YouTube video URLs on each day; Blue curve: The number of unique videos found on each day. Green curve: The number of Twitter users who tweeted YouTube video URLs on each day. All the three values are normalized by the total number of tweets on each day. It can be seen that the three ratios are stable during the 6 months. There are slightly more users tweeting videos than videos tweeted, and on average, about one video is tweeted twice per day.

Other recent YouTube datasets have been constructed using standard feeds ("most recent", "most popular", "deleted") (Figueiredo et al. [2011]; Pinto et al. [2013]), as well as via within-category searches (Cha et al. [2007]), text searches (Xie et al. [2011]), or sampling random video IDs (Pinto et al. [2013]). However constructing a Twitter-driven YouTube dataset in this way will not be biased to the most popular videos, nor will it be biased towards a small list of topics or keywords. These can be justified by Figure 3.3 and Table 3.1 which show the dataset covers videos with a wide range of popularity and categories. Moreover, this approach will mostly return videos that received more than a minimum amount of attention (assuming people who tweet a video likely watch it). Studying videos that are at least a few years old

provides a long-enough history to observe different phases of popularity. Choosing videos that are mentioned in (a random sample of) tweets will yield a set of videos covering diverse topics. Furthermore, discussions that happened on Twitter naturally engenders both endogenous and exogenous evolution of popularity.

Category	**#videos**	**Category**	**#videos**
Music	305450	Games	23384
Entertainment	128319	Howto	22112
People	87161	Travel	22011
Comedy	64607	Nonprofit	18299
News	50862	Animals	13313
Film	41554	Autos	12532
Sports	35748	Shows	3030
Education	30216	Trailers	158
Tech	25566	Movies	54
Total number: 884376			

Table 3.1: The number of videos broken down by user-assigned categories. We can see that Music videos are the most-tweeted (305,450 unique videos), over twice as many as Entertainment (128,319) and over 5 times as many as News (50,862). 15 distinct categories (from Music to Animals) have more than 0.88 millions, or 99% of all videos.

Table 3.1 summarizes the number of unique videos per user-assigned category in this dataset. We can see that Music videos are the most tweeted, 7 categories (until Sports) have more than 700K (or 80%) unique videos, and 15 categories (until Animals) have more than 880K (or 99%) unique videos. The categories Movies and Trailers are at least an order of magnitude less frequent than other categories, likely resulting from a change in YouTube category taxonomy – these 212 videos are excluded from the statistics across categories in later discussions.

Unless otherwise noted, the data explorations in this chapter are based on the 0.88 million videos with complete information.

3.2 The popularity scales of YouTube videos

How is people's attention distributed over the YouTube videos, e.g. is it even or is it highly skewed towards popular ones? Simple observation tells us that the video

viewcounts can be very different. For example, in our dataset, many videos have less than 10 views over more than 5 years. But there is also video whose viewcount is beyond 2 billions[2]. Despite the extreme cases, what is the viewcount distribution of the millions of "intermediate" videos? To examine this, we ranked all videos by the total viewcounts they receive at age t-days, i.e. $sum([x_v(1), ..., x_v(t)])$, where $x_v(t)$ is the daily viewcount of video v on day t after its upload. The rank for each video is converted to a percentile scale, i.e. video v at 1% will be less popular than exactly 1%, or ∼8800 other videos in the collection. We quantize this percentile into bins, each of which contains 5%, or ∼45K videos.

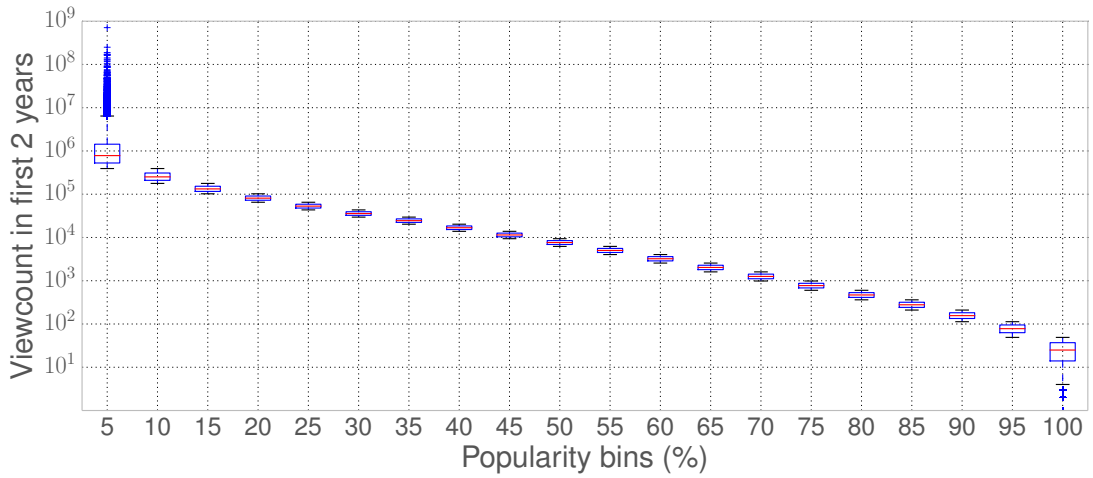


Figure 3.3: Boxplot of viewcount in each popBin. Each bin contains about 45k videos. It can be seen that except for the extreme cases (popBin 5 and 100), the variance of viewcount in each bin is very small and the medians are well fitted by a straight line.

Figure 3.3 shows a boxplot of video viewcounts in each bin after $T = 735$ days. We can see that viewcounts of the 5% most popular (leftmost bin) and least popular (rightmost bin) videos span more than three orders of magnitude. But for the rest of the collection, the popularity distribution is linear on a log scale, implying that the viewcounts of the videos are exponentially distributed over popularity rankings.

[2]See the video named "Gangnam Style" at https://www.youtube.com/watch?v=9bZkp7q19f0

Denoting viewcount as v and viewcount rank as r, then we have,

$$v = ae^{br} \tag{3.1}$$

where the a and b are constants. This kind of pattern for the way users access online content has also been observed and modelled by Guo et al. [2008] using a *stretched exponential distribution*. It is clearly different from a power-law/Zipf-law ($v = ar^b$) which best describes the YouTube access pattern in a local campus network (Gill et al. [2007]) or when restricted to certain categories (Cha et al. [2007]). It is also different from user popularity in an online social network like Twitter (Newman and Park [2003]; Kwak et al. [2010]). All these findings mean that although the distribution of YouTube videos' popularity clearly has its own pattern, it can not be easily explained by classic mechanisms such as *preferential attachment*. Understanding why the pattern has this shape calls for further research.

3.3 Age distribution of tweeted videos

We examine the age of videos in our dataset. This allows us to examine what kind of videos tend to get tweeted a long time after they were uploaded. Analysis of our dataset shows that most tweeted videos were uploaded close to the time of the tweets. To see this, we use T to denote the time range of the Twitter data (i.e., Jun. 1st to Dec. 31st of 2009) and define,

$$\gamma = \frac{\#video\ uploaded\ after\ T}{\#all\ video} \tag{3.2}$$

which roughly measures how likely it was that Twitter users tweeted older videos. Figure 3.4 shows the age distributions of four representative categories of videos in our dataset with γ in plot and Twitter data time-window marked as T. Videos with larger x-values (ages) are uploaded earlier. We can see that, for categories such as "Music" and "Comedy" (videos which have more persistent values), old videos

are more likely to be tweeted ($\gamma > 60\%$). But for "News" and "Games" videos, most of the tweeted videos are newly uploaded. These plots reflect the temporal distribution of how people's attention to different types of YouTube videos occurs due to discussion on Twitter.

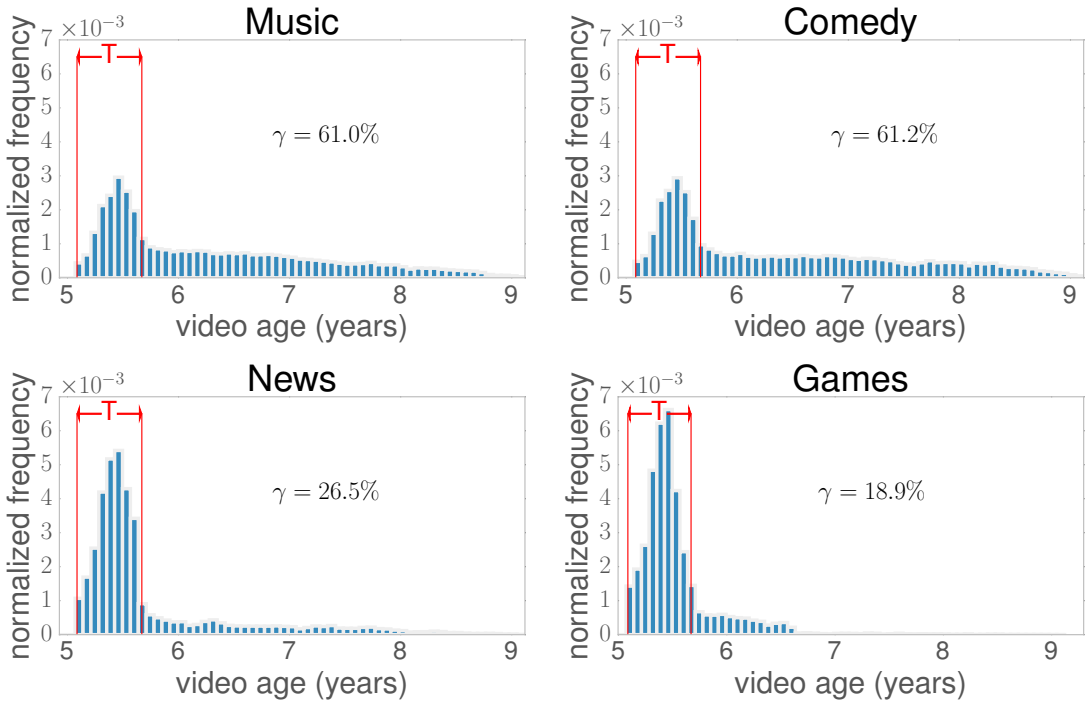


Figure 3.4: Distribution of video ages in our dataset w.r.t. four different video categories (assumed "current" at Feb. 1st 2015). The red time range T denotes the time of tweets from which the videos were sampled from. We can see that, in general, people tend to tweet about newly uploaded videos. However the distributions are different in each category.

3.4 Viewcount over time

Before beginning to model video viewcounts, a natural question to ask is, what is the average viewcount of different types of videos over time? Since videos are uploaded on different days, we explore this topic from two perspectives.

3.4.1 Viewcount over video age

First, we examine people's attention to YouTube videos in terms of age of the videos. Figure 3.5 shows the daily average viewcount over the first 2 months for the categories "Entertainment", "Music", "Comedy", "Tech", "News" and "Games". We can see that, on average, all categories have many more viewcounts in the first week (Wu and Huberman [2007, 2008]): the average viewcount first surges to a peak and then gradually relaxes following a power-law-like shape. This forms the base of many existing research which only considered a single peak in analyzing viewcount dynamics (e.g., Chatzopoulou et al. [2010]; Crane et al. [2008]). But we will see in the following chapters that many videos, especially the popular ones have more than a single peak in their lifecycle.

In particular, we find the behavior of "Shows" videos to be surprising (Figure 3.5:below). They receive many more views per video than other categories. This is probably because "Shows" videos are mostly professionally produced videos of high quality.

3.4.2 Viewcount over date

Figure 3.6:top gives the smoothed average viewcount of six categories by date. The volume of views all increases before T simply because new videos keep being uploaded. The salient feature is the viewcount after T: "Entertainment", "Music" and "Comedy" videos preserve similar viewing level for a long time ($>$ 1 year). But the daily views of "Tech", "News" and "Games" videos drop immediately after T. The difference depends on whether those videos have persistent value.

As described above and shown in Figure 3.6:below, the number of views received by "Shows" videos received is at least one order of magnitude greater than other video types. We can also see that, since the time range in Figure 3.6 is quite long ($>$6 years), old videos still receive a significant amount of views/attention.

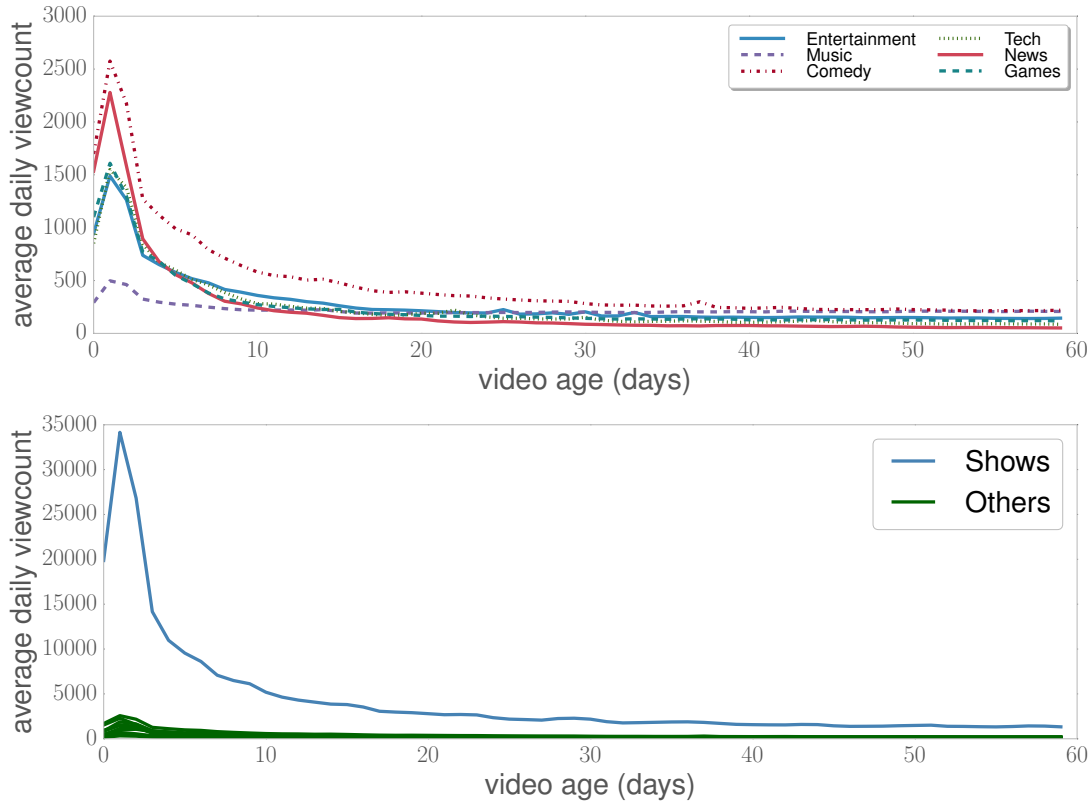


Figure 3.5: Top: Average daily viewcount (since upload) of 6 different categories of videos; Below: Comparison of the average daily viewcount of "Shows" and other types of videos. In this plot, we have omitted the "near-horizontal" long tails from 60 days to 2 years.

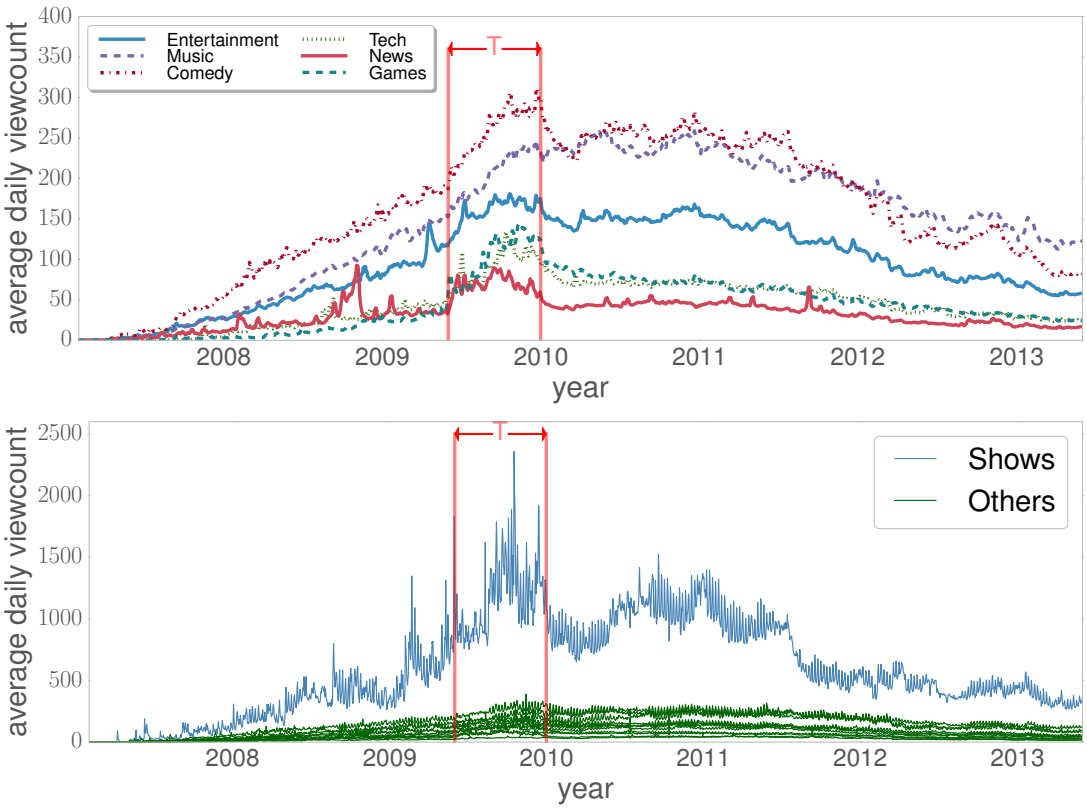


Figure 3.6: Top: Average viewcounts per day of 6 different types of videos; Below: Comparison of average viewcounts per day of "Shows" with other types of videos. The labeled time range T is the when the Twitter data are being downloaded.

3.5 Viewcount temporal correlation

In the pioneer research on predicting the popularity of online content, Szabo and Huberman [2010] pointed out that there is strong linear correlation between videos viewcount over the first week and the first month on a log-log scale. This work has formed the basis of subsequent viewcount prediction research (Pinto et al. [2013] etc.). In this section, we make a more comprehensive analysis by examining the correlation between between every pair of viewcount growth in different months after upload (See Figure 3.7).

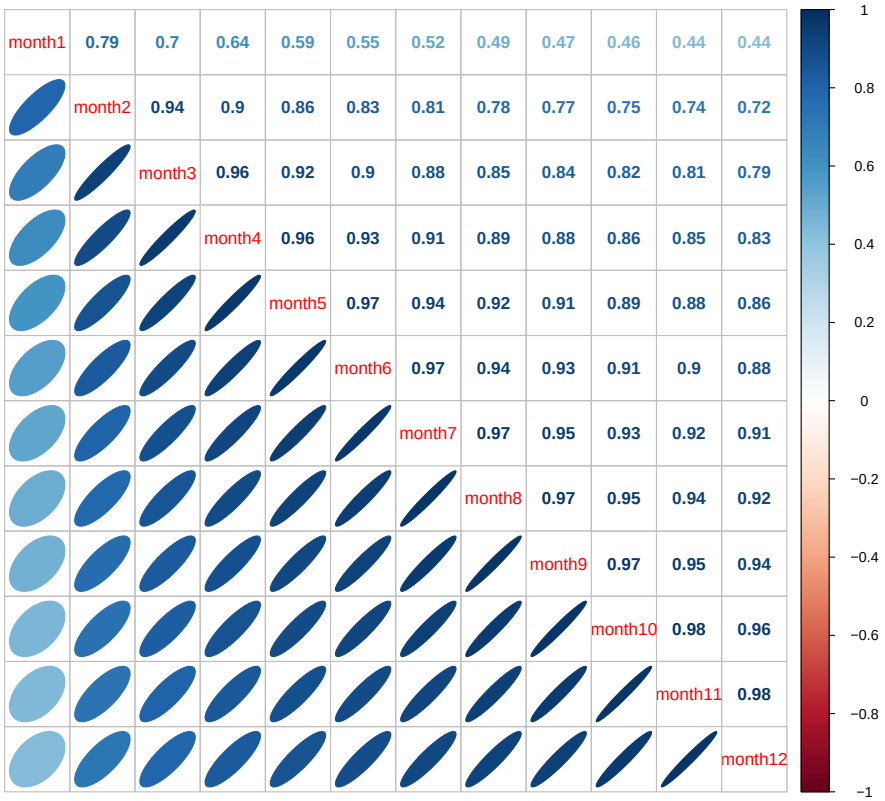


Figure 3.7: The correlation coefficient among viewcount increase of different months after upload on (a log scale). Notice that there are strong linear correlations, especially between adjacent months. And the first month (month1) is special: it has weak correlation with all the other months.

Figure 3.7 shows that there is a strong linear correlation between consecutive months. This means that short-term prediction is feasible. Interestingly however, the first month has the weakest correlation with the others. This again implies that the

viewcount dynamics in the early stage are unique and complex. We will use external information to help predict early viewcount in Chapter 7.

3.6 Heteroscedasticity

Another complexity of viewcount dynamics is that the variance or noise level over time is far not constant (*heteroscedasticity*). Figure 3.8 shows a typical example. The detection of change of variances can not easily be done by classic *change-point detection* methods because the mean also varies over time. Developing new time series analysis method to detect the noise patterns and the implications of the viewcount variance are left for future research.

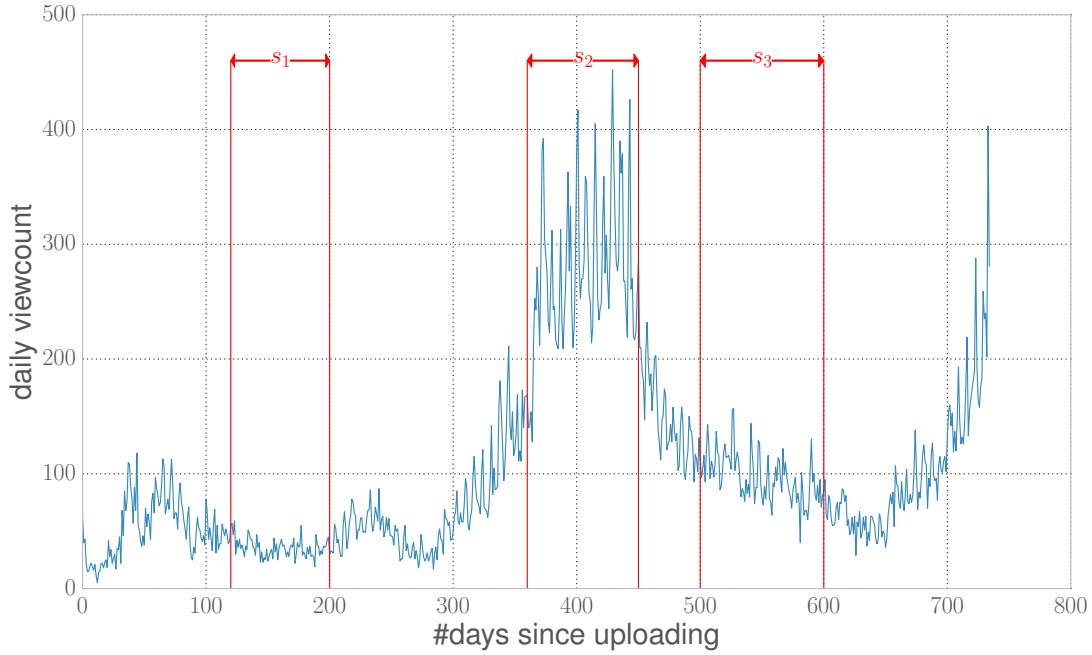


Figure 3.8: Daily viewcount of video "oslCBENbkGw". It can be seen that the strengths of noise magnitude is quite different over time ranges s_1,s_2 and s_3. But part of s_2 and s_3 should belong to the same declining phase after the peak.

3.7 Periodicity

This section focuses on the periodicity of video viewcounts. We first examine the weekly and then yearly periodicity.

3.7.1 Weekly periodicity

Since watching a video is often more time-consuming than reading a short text message such as a tweet, it is intuitive that people tend to watch more videos during weekends. This is proven by our dataset. Let us first look at one typical example. Figure 3.9 shows a video about transformer toys, which may have some specific community interest. Its daily views stabilizes at about 5000 views for more than 2 years. We can clearly see its viewcount reaches a local maximum on most of the Saturdays and that the amplitude of the variance changes over time (although the mean does not change).

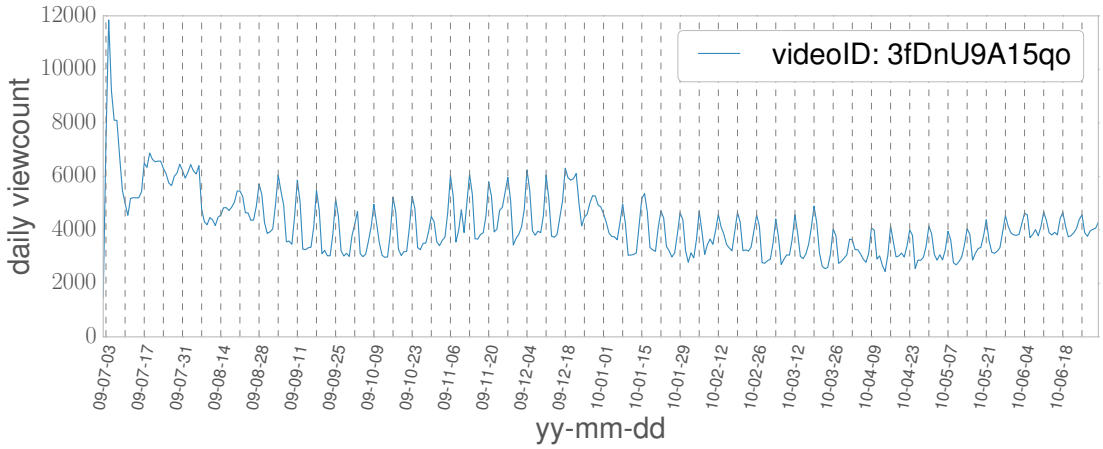


Figure 3.9: A video whose daily viewcount clearly exhibits weekly periodicity. Every vertical dashed line corresponds to a Saturday in each week.

Weekly periodicity also depends on video types. Figure 3.9 gives the distribution of aggregate views over 5 different video categories. Most categories, e.g., "Music", "Games", "Shows" and "Travel" are viewed during weekends. But for "Tech" videos there is no strong difference over different weekdays. In our dataset, we did not find

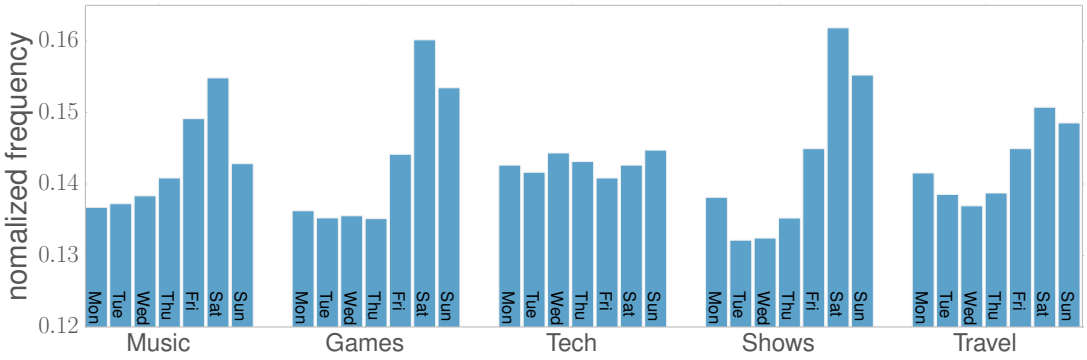


Figure 3.10: Distribution of number of views over weekdays for different categories of videos. All times are transformed to UTC-6 time zone in which most of YouTube users lived in 2009.

any type of videos which had most views on working days.

To make a normalized comparison, Figure 3.11 gives the upload distribution by day of the week. We can see that, video upload is more variable than video viewing. Some categories, e.g., "Tech" and "Shows" are uploaded mostly during working days — this implies again that "Shows" videos may well be uploaded mostly by official accounts, whereas videos like "Travel" are likely to be uploaded at weekends (note that there are also more uploads on Monday, which implies "Travel" videos may be uploaded globally). Finally, for some videos, like "Music" and "Games", we can not, on average, observe any weekly periodicity of people's upload behavior.

3.7.2 Yearly periodicity (seasonality)

Beside weekly periodicity, seasons can also affect the number of views a video recieves, and this causes a yearly periodicity in viewcount data. Figure 3.12 shows viewcount data for a video teaching people how to swim. It reaches a peak around every August, which is one of the hottest months in the Northern Hemisphere. Similarly, there are other examples such as videos on "global warming", which tend to also have viewcount peaks every summer and reach a trough every Christmas.

To summarize, from the two kinds of periodicity, we can see that, 1. Viewcount data for YouTube videos is very diverse and reflects many aspects of the way society

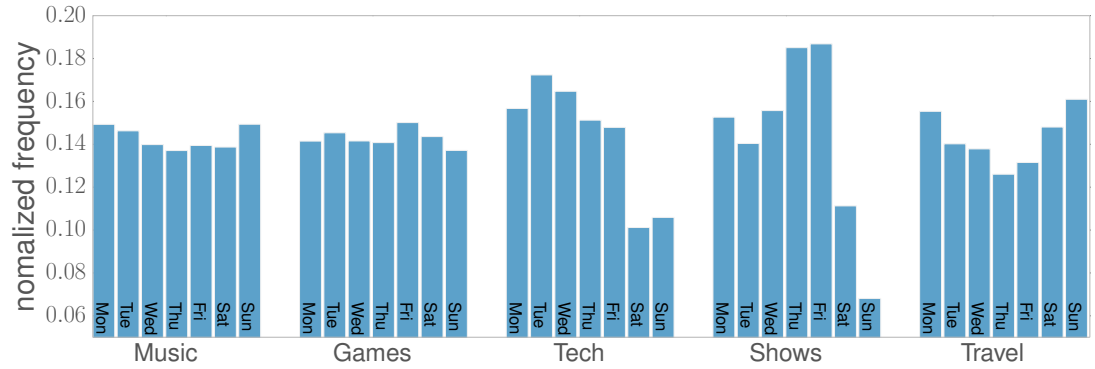


Figure 3.11: The distribution of video upload over weekdays. We can see that, some categories of videos like *Music* and *Games* are uploaded almost equally over a week. Whereas some videos like *Tech* and *Shows* are mostly uploaded during working days. Others like *Travel* are slightly more likely to be uploaded during weekends.

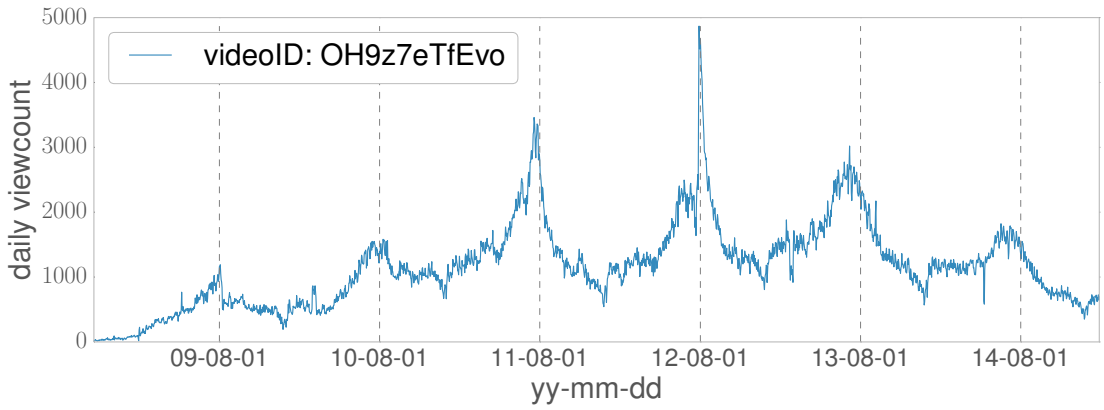


Figure 3.12: A viewcount series showing yearly periodicity. The video teaches how to swim. Its viewcount reach peaks every summer. The vertical dot lines correspond to the date "Aug. 1st" in each year.

functions; 2. For many videos, the viewcount series can be far more complex than the average curve shown in Figure 3.6.

In the following two sections of this chapter, we explore correlations between YouTube video viewcounts and external information.

3.8 Viewcount and external interventions

Since our videos are sampled from Twitter data, a natural and important question is: is there correlation between viewcount changes and video tweets? The answer is yes. We will examine this in more detail in Chapter 7. Here, we only provide observations revealing the correlation.

3.8.1 An example of viewcounts with external intervention

As the world largest user-generate-content (UGC) website, YouTube frequently interacts with main-stream medias and online social networks. Sometimes, an external effect can be huge — making an obscure video become viral. In a TED talk (Allocca [2011]), YouTube trends manager Kevin Allocca has given a famous example. A viral video called "double rainbow"[3] was hardly watched in the first 6 months after its upload, until Jimmy Kimmel[4] who has millions of followers on Twitter tweeted and recommended this video to all his followers. The viewcount suddenly jumps afterwards and surged up to millions. This example reveals that some celebrities are very influential in guiding people's attention.

Figure 3.13 gives another example of how social events affect YouTube viewcounts. In June 25th, 2009, the *King of Pop* Michael Jackson suddenly passed away. We looked up keywords ("RIP", "MJ", etc.) on video tweets and found 300 videos about him. It can be seen that, the viewcount of these videos clearly jumps around June 25 and then gradually drops off until another major event — "Public Memorial

[3] http://bit.ly/1hky3gS
[4] http://bit.ly/1soOuim

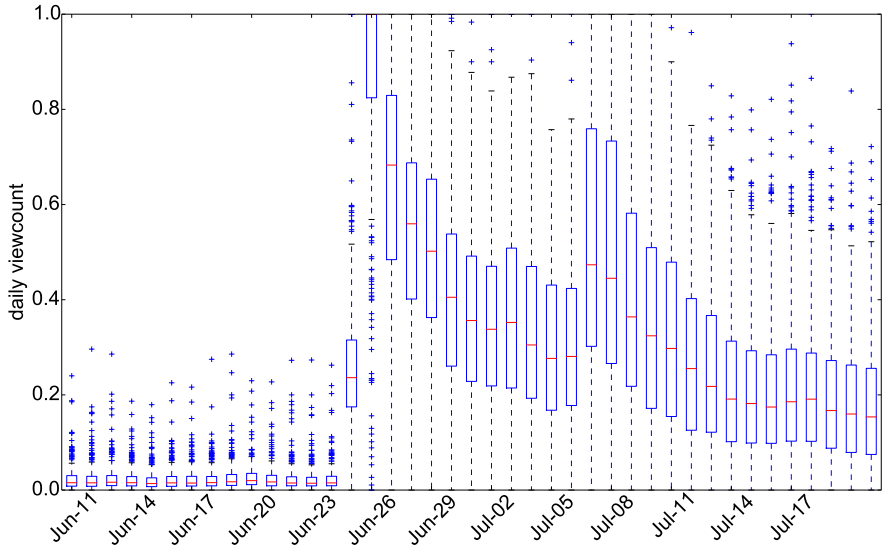


Figure 3.13: Normalized viewcount of 300 videos concerning Michael Jackson. We can clearly see two sudden jumps of the medians round June 25th (when he suddenly passed away) and July 7th, 2009 (when the memorial service was hold).

Service for Michael Jackson"[5] around July 7th, 2009. By examining these viewcounts, we can, from a quantitative point of view, see that the death of Michael Jackson has a great impact and which are the major events associated with it.

Overall, as an aggregate measure of people's attention, YouTube viewcounts depend on many external factors. Figuring out their relationships not only helps us better understand the way in which the popularity of a YouTube video evolves, but also helps us better understand what is influential on our society.

3.8.2 Video Tweets and Viewcount Increases

In this section, we will show the correlation between video tweets and viewcount increase in general. Figure 3.14 is a typical example of this correlation. This video has nearly no views before Sept. 8th 2009, but suddenly attract around 1000 views per day afterwards. Correspondingly, there was also a tweet peak at this time. If we denote the viewcount increase in three time intervals (T_{before}, T_{around} and T_{after}) surrounding the tweet peak as Δv_{before}, Δv_{around} and Δv_{after}, in this example, we have

[5] http://bit.ly/1CKNBTt

$\Delta v_{around} >> \Delta v_{after} > \Delta v_{before}$.

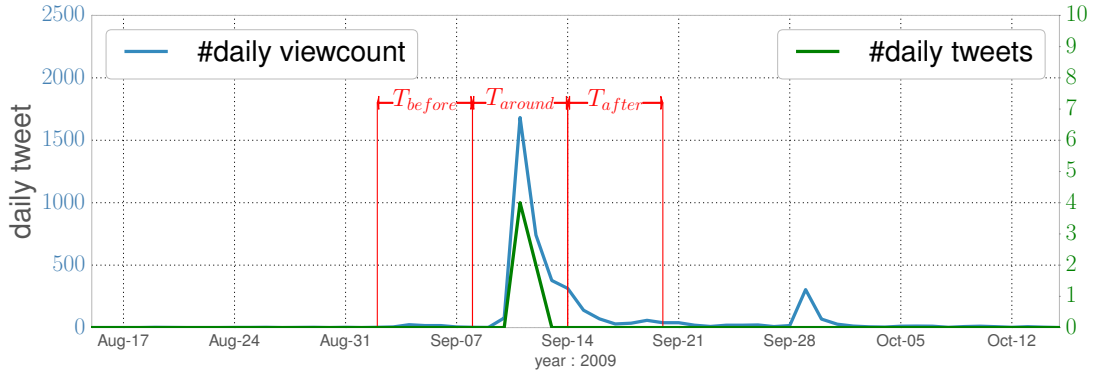


Figure 3.14: An example (videoID:0S00oo-xgdE) of correlation between viewcount and tweets.

One may ask is the correlation above coincidental? Figure 3.15 provides the scatter plot between Δv_{around} and Δv_{after}, Δv_{around} and Δv_{before} for all videos with at least five tweets. From it we can see that, on average most videos will have (at any given time) fastest viewcount increases around tweet peaks.

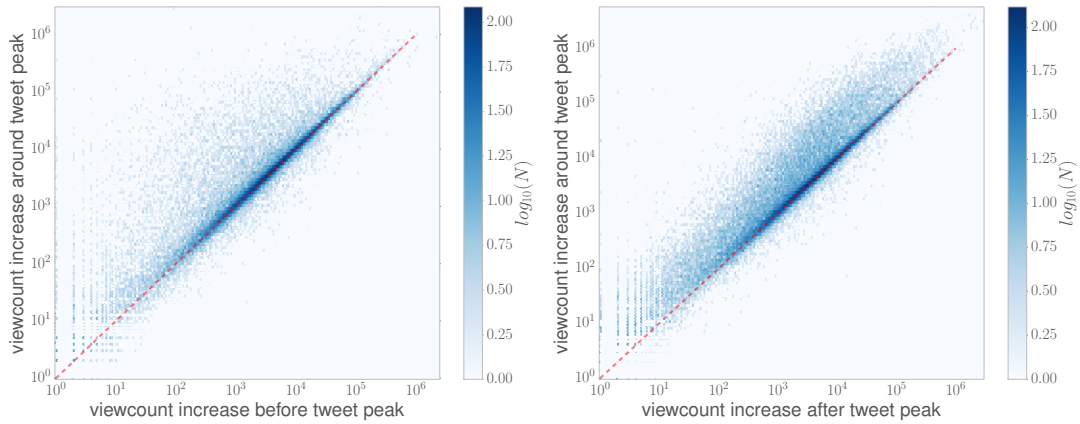


Figure 3.15: Comparison of viewcount increases around tweet peak (for videos with at least 5 tweets). Left: Viewcount increase before and around the peaks of 39,455 videos; Right: Viewcount increases around and after the peaks of 103,470. Diagonal lines mark $y = x$. All the videos here had at least 5 tweets in their tweeting peaks. We can see that most videos lie above the line, meaning that, on average, viewcounts around tweet peaks increase faster than that before or after the tweet peaks. ($t_{before} = t_{around} = t_{after} = 7days$)

The clear intervention patterns in Figure 3.15 imply that we can utilize Twitter

information to help analyze and predict viewcount "jumps", an approach which will be discussed in Chapter 7.

3.9 Summary

The following points summarize this chapter.

1. A large and diverse dataset for gauging video popularity over time has been collected. These videos have been extracted from a large scale Twitter dataset and are videos that received more than a minimum amount of attention.

2. We have defined the popularity scales (popBins) of YouTube videos and found that video views are distributed exponentially in terms of popularity ranks.

3. By exploring videos' popularity over time, it is found that most videos receive most of their viewcount early, just after they have been uploaded.

4. By examining the ages of videos, it is clear that old videos of some categories such as "Music" and "Comedy" are more likely to be tweeted than "News" and "Games".

5. It was found that patterns of viewcount evolution are very diverse and complex (they have periodicity, multiple peaks, etc.).

6. Under major external events, the viewcounts of certain groups of videos change in coordinated ways. By using aggregated plots, it becomes clear that there are correlations between viewcount increases and (sufficiently strong) tweet peaks.

It should be stressed again that, because of many internal factors (quality, category etc.) and external factors, viewcount dynamics can be very complex. This complexity has been underestimated by much previous research. In the next chapter, we introduce a novel time series segmentation method to help us analyze the temporal complexity of viewcount evolution.

Chapter 4

Viewcount Phase Segmentation

In this chapter, we introduce a new method for breaking time series into distinct segments. Based on the large-scale measurement study in Chapter 3, it was found that a lot of videos exhibited complex dynamics in their popularity over time in a way which seemed to show multiple phases. This has not been described by any previous research literature. To analyse these phases, we propose a generalized power-law to describe each rising or falling phase of popularity. Furthermore, we propose a novel algorithm to simultaneously segment and estimate the phase representations of popularity history. This efficient algorithm could be useful not only for the analysis of YouTube video viewcounts, but also for any research on longitudinal data related to popularity.

4.1 Motivation

How does a video become viral? This is a well-known open research question in studies of social media and collective online behavior. An online information network is known to have bursts of activities responding to endogenous word-of-mouth effects or sudden exogenous perturbations (Crane and Sornette [2008]). A number of studies have shown that a video's long-term popularity is often determined by and can be predicted from its early views (Cheng et al. [2008]; Szabo and Huberman [2010]; Pinto et al. [2013]), and that an early-mover has an advantage in competing for attention (Borghol et al. [2012]). Recently several groups of researchers have studied the relationship between content popularity and a variety of other factors, including

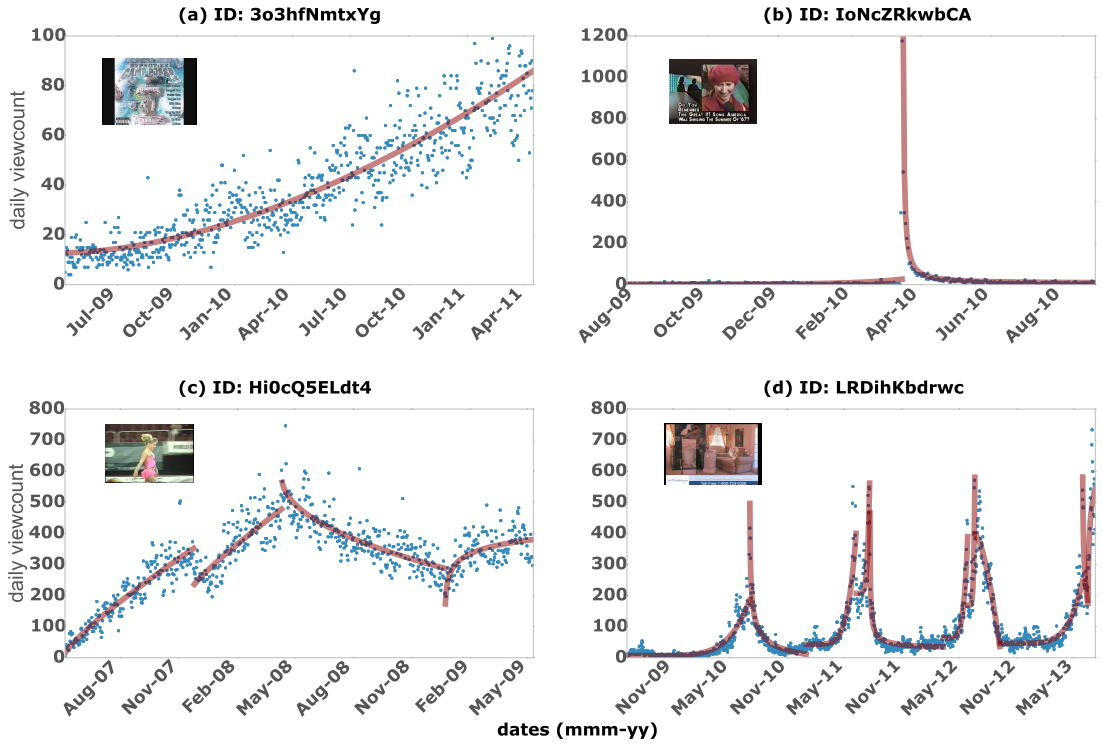


Figure 4.1: The complexity of viewcount dynamics: the lifecycles of four example videos. Blue dots: daily viewcounts; red curves: phase segments found by our algorithm. (a) A video with a single power-law growth trend. (b) A video with a single power-law decay. (c) A video with many phases, including both convex and concave shapes (this video contains a Gymnastic performance). (d) A video with what seems like an annual growth and decay (this video demonstrates how to vent a portable air-conditioner, and reaches viewcount peaks during each summer). Viewcount shapes such as (a) and (b) are explained by the model of Crane and Sornette, but (c) and (d), and many others like them, are not.

network actor properties (Cheng et al. [2014]), content features (Cheng et al. [2014]; Bakshy et al. [2011]), and effects of complex contagion (Romero et al. [2011]). However, some questions remain: what does a video's lifecycle look like? Is there a single perturbation, or multiple endogenous or exogenous shocks?

One well-known model of social media popularity was proposed by Crane and Sornette [2008]; the model suggests that popularity over time consists of power-law precursory growth or power-law relaxations. Such rising and falling power-law curves are indeed often observed – Figure 4.1(a) and (b) show respective examples

of each type. Note, however, that real popularity cycles are often more complex – a video can go through multiple phases of rise and fall, as shown in Figure 4.1(c) and (d).

To address this limitation and adequately describe the rich patterns seen in the lifecycle of many videos, we propose a novel representation that uses "popularity phases". We propose a method to divide popularity histories into distinct phases and at the same time find the optimal parameters to describe the shape of each phase. Later, in Chapter 5), it will be shown that the phases enable us to discern a rich and multi-faceted view of popularity dynamics. There are successive rising and falling phases which are closely related to content type and popularity.

Besides being of benefit to measurement studies, the phases found here can also be used to represent the original time series. In Chapter 6, popularity phases will be utilized to help developing a new viewcount clustering method and significantly improve viewcount prediction. Although the focus of this thesis is on YouTube videos (one of the few sources where the popularity history is publicly available), the method for extracting phases and analysing about viral content could potentially be applied to other media content of a similar type.

The outlines of following sections are as follows, we will first briefly highlight the main related work in Section 4.2 (more can be found in Section 2.2.1). Then the phase finding problem is formulated at the beginning of Section 4.3. Section 4.3.1 and 4.3.2 are dedicated to discussing how to implement the new algorithm. The notations used in these two sections follow the classic way of describing dynamic programming algorithms (Rabiner [1989]). After that, in Section 4.4, we will compare our method with the classical ones. Finally, in Section 4.5 we will introduce the systematic evaluation of our algorithm and how to tune the parameters.

4.2 Related work

Although this work relates to many areas of active research, we will structure our discussion along four lines: (1) an empirical profile of YouTube statistics, (2) models for describing popularity dynamics, (3) predicting social media popularity, and (4) time series segmentation.

The first category is large-scale empirical analysis of YouTube videos. In two pioneering papers, Cha et al. [2007] measured video metadata statistics from nearly 2 million videos, and Gill et al. [2007] analysed video usage and file properties from network traces. Subsequent metadata analysis has concentrated on the relationship between video popularity and other observable metrics. Cheng et al. [2008] found that popularity of most videos was determined in its early stage. Chatzopoulou et al. [2010], after examining 37 million videos, observed that while popularity and user activity metrics were strongly correlated, a video's average rating did not correlate with them. Figueiredo et al. [2011] compared viewcount dynamics of "top", "deleted" and "random" videos, and found that bursts of popularity tended to be caused by external search traffic and referrals. Borghol et al. [2012] took a unique approach by examining duplicate videos, and found that there was a distinct "early-mover advantage". Our finer-grained representation of viewcount evolution derived in this chapter, which involves distinct phases, is inspired by these examples from a rich literature of measurement studies.

Among models to describe social media popularity, our method is closely related to Crane and Sornette's model on endogenous growth and exogenous shocks Crane and Sornette [2008], measured on thousands of videos. The same authors (Crane et al. [2008]) also found that the shapes of popularity dynamics were related to inherent interest – quality videos often relaxed slower than junk videos. Our proposal for popularity phases extends the notion of a shock from just one to multiple times throughout a video's lifetime, and our generalized power-law model captures a wider class of shapes than shapes in the original model. Recently, Yang et al. [2014]

proposed a *progression stage* model for event series that attempts to capture the complex evolution of a response to external stimuli. But this algorithm still requires that the user determines the number of segments beforehand.

Predicting social media popularity is another area of active investigation. Szabo and Huberman [2010] found strong linear correlations between (the log of) view-counts in the first week and those after the first month. Pinto et al. [2013] built on this insight and used multi-linear regression on the shape of the popularity in the first few days to further improve medium-term predictions. Cheng et al. [2014] redefined the cascade prediction problem and successfully designed a method to predict photo-sharing behavior on Facebook. One of the major contributions was that the authors not only considered the total number (popularity) of sharing, but also the structure of the sharing network.

The problem of approximating a smooth nonlinear function by a predefined number of linear segments was formulated by Stone [1961] in 1961, and solved with a dynamic programming algorithm by Bellman [1961]. Subsequent work has used polynomial segments to approximate a sequence; a good review is presented by Pavlidis [1973]. Heuristic approaches can be used to determine the number of segments, and these include balancing the residuals (Pavlidis [1973]) or minimizing the standard error of the residuals (Keogh [1997]). In the last 15 years or so, the time series data-mining community have developed efficient approaches for longer sequences, local approximations to segment description, and selection of models that trade fitting error against the number of segments (as summarized by Keogh et al. [2004]). We refer readers to comprehensive reviews of more recent approaches (Esling and Agon [2012]; Fu [2011]). Compared to the existing approaches, our phase-finding algorithm is one that is specifically tailored to find globally-accurate burst phases in the popularity profiles of social media: The joint segmentation and description of power-law shapes is new, including the heuristic for robust and fast fitting with variable projection.

4.3 The PHASE-FINDING problem

We define a *phase* as one continuous time period in which a video's popularity has a salient rising or falling trend. In this section we present a model to describe such phases in a time series, and propose efficient algorithms to simultaneously find both phase segments and their shape parameters.

Given the daily viewcount for video v ($\mathbf{x}_v = x_v[1:T]$), the goal is to segment this time series as a set of successive *phases* $\boldsymbol{\rho}_v$ in which each phase $\rho_{v,i}$ is uniquely determined by its starting time $t^s_{v,i}$ (with $1 = t^s_{v,1} < t^s_{v,2} < \ldots < t^s_{v,n} < T$). In the rest of this section, we omit subscript v without loss of generality, i.e., $\mathbf{x} = x[1:T]$, ρ_i. We also derive and include the ending time for phase ρ_i as t^e_i. It is one day before the starting time of the next phase, $t^e_i = t^s_{i+1} - 1$, if $i < n$; or equal to the maximum time index T for the last phase, $t^e_n \equiv T$. Phase ρ_i is described by these two timestamps:

$$\rho_i = \{t^s_i, t^e_i\}.$$

We use a generalized power-law curve to describe viewcount evolution in each phase:

$$x[t] = at^b + c \tag{4.1}$$

with a power-law exponent b, scale a and shift c. The power-law shapes are suitable for describing general popularity evolutions for the following reasons:

1. They are the result of an epidemic branching process (Sornette and Sornette [1999]; Sornette and Helmstetter [2003]) with power-law waiting times (Crane and Sornette [2008]).

2. Such a generalized power-law shape is sufficiently expressive for describing a wide range of monotonic curves that are either accelerating or decelerating in their rise (or fall). A change in rising/falling or acceleration/deceleration indicates either an external event or a changed information diffusion condition, and hence they are identified as separate phases.

3. The optimal fit is efficiently computable, as described in Section 4.3.1.

Note that the proposed power-law shape generalizes the curve shapes modeled by Crane and Sornette [2008] – there $a = 1$, $c = 0$, and b is in the range of $[-1.4, -0.2]$. Compared to Crane's model for isolated viewcount peaks, there are three data modeling needs to construct such a generalized power-law model. The first is to account for multiple peaks in the same video's lifetime, potentially generated by a number of exogenous or endogenous events of different strengths – hence varying a. The second is to account for different background random processes that are super-imposed onto the power-law behavior – hence varying c. The third is to empirically determine b – Crane's model presented the empirical mean of the power-law exponent b , while the exponent b recovered from data spans a wider range. Our model relies on the phase-finding algorithm to determine a, b and c from observations.

In order to capture all monotonically accelerating or decelerating power-law shapes, we allow two temporal directions in Equation (4.1), i.e.,

$$\begin{aligned} x[\tau] = a\tau^b + c, \text{ with either} \\ \tau = t, \text{ denoted as } \leftarrow, \text{ or} \\ \tau = \bar{T} - t, \text{ denoted as } \rightarrow . \end{aligned} \tag{4.2}$$

Denote the parameter set of the generalized power-law as:

$$\theta = [a, b, c, \tau]^T$$

Equation(4.3) describes a phase-fitting problem: to find the optimal θ_i^* for a given starting and ending time of a sequence t_i^s, t_i^e, minimizing a loss function $E_i\{\}$ between the observed and fitted volumes.

$$\text{given } t_i^s, t_i^e, \text{ find } \theta_i^* = \arg\min_{\theta} E_i\{x[t_i^s : t_i^e], \theta_i\} \tag{4.3}$$

Given the daily viewcount series $\mathbf{x}_{1:T}$, the PHASE-FINDING problem can be expressed as simultaneously determining the parameter set $\mathcal{S}$ a phase segmentation $\{\rho_i, 1 \leq i \leq n\}$ with n being the number of phases; and the optimal phase parameters $\{\theta_i, 1 \leq i \leq n\}$.

$$\begin{aligned} \text{Find} \quad & \mathcal{S} = \{n;\ t_i^s, \theta_i,\ i = 1, \ldots, n\} \\ \text{minimize} \quad & E\{\mathbf{x}_{1:T}, \rho_{1:n}, \theta_{1:n}\} \\ = \quad & \sum_{i=1}^{n} E_i\{x[t_i^s : t_i^e], \theta_i\} \end{aligned} \tag{4.4}$$

This formulation implies two main subproblems: (1) How to fit the generalized power-law curve; (2) How to efficiently solve the joint segmentation and estimation problem. The solutions to these problems are described in the following two subsections, respectively.

4.3.1 Estimating a generalized power-law phase

In this work, we use a sum-of-squares loss function in problem (4.3). Denote relative time and duration in this phase as time elapsed since the end of the previous phase $\bar{t} = t - t^s + 1$ and $\bar{T} = t^e - t^s + 1$.

$$\begin{aligned} \min_{\theta} \quad & E\{x[t^s : t^e], \theta\} \\ & = \frac{1}{2} \sum_{\bar{t}=1}^{\bar{T}} (a\bar{t}^{\,b} + c - x[\bar{t}])^2 \end{aligned} \tag{4.5}$$

Notice that this loss function is differentiable everywhere, but non-convex in θ – it can be optimized with a general unconstrained optimization technique such as Newton's method, but it will be prone to local minima and slow to converge. We adopt a technique called *variable projection* (Golub and Pereyra [2003]) to address this problem. The basic idea is to separate the nonlinear parameter b and the linear

parameter a, c, by re-writing the loss function as follows.

$$E = \frac{1}{2}\sum_{\bar{t}=1}^{T}(a\bar{t}_i^b + c - x[\bar{t}])^2 = \frac{1}{2}||\Phi(\bar{t}) \cdot \beta - \mathbf{x}||^2 \tag{4.6}$$

where

$$\Phi(\bar{t}) = \begin{bmatrix} 1^b, & 1 \\ 2^b, & 1 \\ & \dots \\ \bar{T}^b, & 1 \end{bmatrix}, \beta = \begin{bmatrix} a \\ \\ c \end{bmatrix} \tag{4.7}$$

$\Phi(\bar{t})$ includes the nonlinear parameter b, and β includes the linear parameters a and c. Given b, there is a unique minimum for the quadratic equation (4.6), with a, c given by the following closed form via the Moore–Penrose pseudoinverse:

$$\begin{bmatrix} a \\ \\ c \end{bmatrix} = (\Phi(\bar{t})^T\Phi(\bar{t}))^{-1}\Phi(\bar{t})^T\mathbf{x} \tag{4.8}$$

Equation (4.8) is a *necessary* condition for the optimal solution of Equation (4.6), and since the dimension of matrix $\Phi(\bar{t})^T\Phi(\bar{t})$ is 2×2, it is very easy to invert. Substituting a and c by it, the loss function becomes:

$$E = \frac{1}{2}||\Phi(\bar{t})(\Phi(\bar{t})^T\Phi(\bar{t}))^{-1}\Phi(\bar{t})^T\mathbf{x} - \mathbf{x}||^2 \tag{4.9}$$

Now, we have reduced the parameter space from $\theta \in \mathcal{R}^3$ to $b \in \mathcal{R}$, and the optimal solutions of Equation 4.9 are the same as those of Equation (4.6).

Implementation and solution quality. We use the *L-BFGS-B* algorithm (Zhu et al. [1997]) to find a solution of this non-linear objective. We observed significant improvement in speed and solution quality with the variable projection technique, consistent with the original proposal Golub and Pereyra [2003]. We also normalize $\mathbf{x}_{1:T}$ into $[0, 100]$ before running the phase-finding algorithm so as to make it the

same order of magnitude with time stamp $\bar{t}$ – thus avoiding numerical issues in power-law fitting across daily viewcounts from ≤ 10 to the order of 10^6. In particular, we employ the following initialization technique to start from a good "guess" of b — we use $\bar{t} = 1$ with each observation $\bar{t} = 2, \ldots, \bar{T}$ to solve for a value b by assuming each pair exactly follows the power-law "$x = at^b$" (without c), and then take the mean of all these estimates as the initial point[1]. As an initial validation for the solution quality of this curve-fitting problem, we generate 500 synthetic power-law curves (length:200) with uniformly random parameters $a \in [-100, 100]$, $b \in [-2, 2]$, and $c \in [-500, 500]$, with a and b bounded away from zero to avoid degenerate cases ($|a| > 3$, $|b| > 0.1$). We optimize Equation 4.9, and observe the relative fitting error in each coefficient ($E_a = |a^* - a|/|a|$, with its confidence interval) as: $E_a = (1.8 \pm 0.3) \times 10^{-3}$, $E_b = (1.1 \pm 0.6) \times 10^{-5}$ $E_c = (1.8 \pm 0.3) \times 10^{-3}$.

4.3.2 Simultaneous fitting and segmentation

A bruce-force enumeration approach to the joint segmentation and curve-fitting problem (4.4) will have a complexity exponential in T, the sequence length. Fortunately, problem (4.4) is in a form suitable for induction with dynamic programming. We describe the algorithm in three stages, similar to (but extending) the description of the well-known Viterbi decoding algorithm Rabiner [1989] with embedded curve-fitting.

As in problem (4.4), denote $1 \leq t' \leq T$ as the current position in the recursion, n' as the number of optimal segments up to position t', and a shorthand $E^*(t')$ for the lowest segmentation and fitting error (under any segmentation) for the subsequence $\mathbf{x}_{1:t'}$

$$E^*(t') = \min E\{\mathbf{x}_{1:t'}, \rho_{1:n'}, \theta_{1:n'}\} \tag{4.10}$$

where the minimization is done over $\{n', t^s[1:n'], \theta_{1:n'}\}$.

[1] This initialization heuristic is documented in the `power2start()` function of the Matlab curve-fitting toolbox.

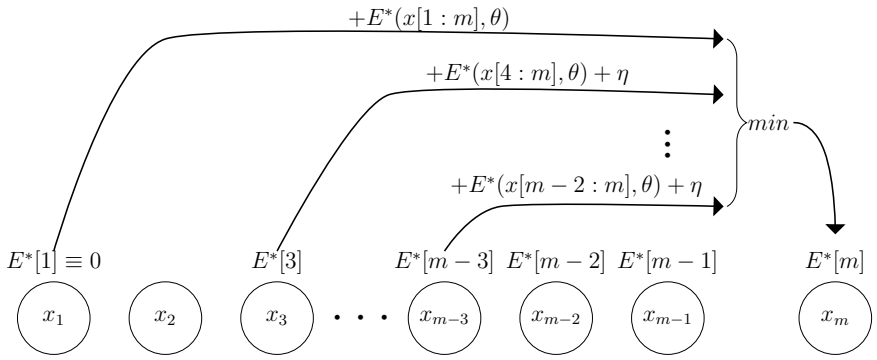


Figure 4.2: How to compute the lowest fitting error $E^*[m]$ (equation 4.10) from $E^*[t],\ t = 1, 2, ..., m-3$. The circles $x_1, x_2 ... x_m$ are the viewcount series. For $t > 1$, we add η to the loss objective $E^*[m]$ to penalize over-segmentation (see Equation 4.16).

In order to retrieve an optimal segmentation, we need to keep track of arguments that minimize Equation (4.10) for each t'. This is done via a pointer for each t' containing the starting position of the last phase and its parameters.

$$\delta(t') = \{t^{s*}_{n'}, \theta^*_{n'}\} \tag{4.11}$$

The complete procedure for finding the best segmentation and their power-law fits is as follows:

Stage 1 Initialization:

$$\text{for } t = 1, 2,\ E^*(t) = 0 \tag{4.12}$$
$$\delta(t) = \emptyset$$

The reason we initialize a cost of zero and an empty parameter set for the first two positions (instead of only for $t = 1$ as in the Viterbi algorithm) is that the generalized power-law curve has three free parameters, and hence takes at least three observations to fit.

Stage 2 Recursion:

$$E^*(t') = \min_{t^s_{n'}, \theta_{n'}} \{E^*(t^e_{n'-1}) + E(x[t^s_{n'} : t'], \theta_{n'})\} \tag{4.13}$$

The step above computes the cumulative minimum error, for $t' = 3, 4, \ldots, T$. This is done by searching for an optimal starting point $t^s_{n'} = 1, 2, \ldots, t'$, for the *current* phase that ends at t', and obtaining optimal parameter $\theta^*_{n'}$ that minimizes fitting error on subsequence $x[t^s_{n'} : t']$ for each $t^s_{n'}$, using the algorithm in Section 4.3.1. We also populate the backtracking pointers:

$$\delta(t') = \arg\min_{t^s_{n'}, \theta_{n'}} \{E^*(t^e_{n'-1}) + E(x[t^s_{n'} : t'], \theta_{n'})\} \tag{4.14}$$

Stage 3 Backtracking: The set of segmentation parameters $\mathcal{S}^* = \{n^*, t^{s*}_{1:n}, \theta^*_{1:n}\}$ is obtained via a mini-recursion:

- Initialize $\mathcal{S}^* \leftarrow \delta(T),\ t' \leftarrow t^s_{n'},\ n^* \leftarrow 1$;
- Recurse $\mathcal{S}^* \leftarrow \mathcal{S}^* \cup \delta(t'),\ t' \leftarrow t^s_{n'},\ n^* \leftarrow n^* + 1$;
- Terminate $\mathcal{S}^* \leftarrow \mathcal{S}^* \cup n^*$.

How to avoid over-fitting Every three observations will provide a unique solution for the curve-fitting problem (4.3) with a set of a, b, c – this can easily lead to over-fitting by over-segmentation. We introduce a segment regularizer by adding a penalty constant η to every new segment introduced by the algorithm. That is, the objective for problem (4.4) is modified as:

$$\tilde{E}\{\mathbf{x}_{1:T}, \rho_{1:n}, \theta_{1:n}\} = \sum_{i=1}^{n} E_i\{x[t^s_i : t^e_i], \theta_i\} + (n-1)T\eta \tag{4.15}$$

where T in the regularizer make η not affected by the total length of the sequence. Minimizing the objective is still done with dynamic programming by simply adding

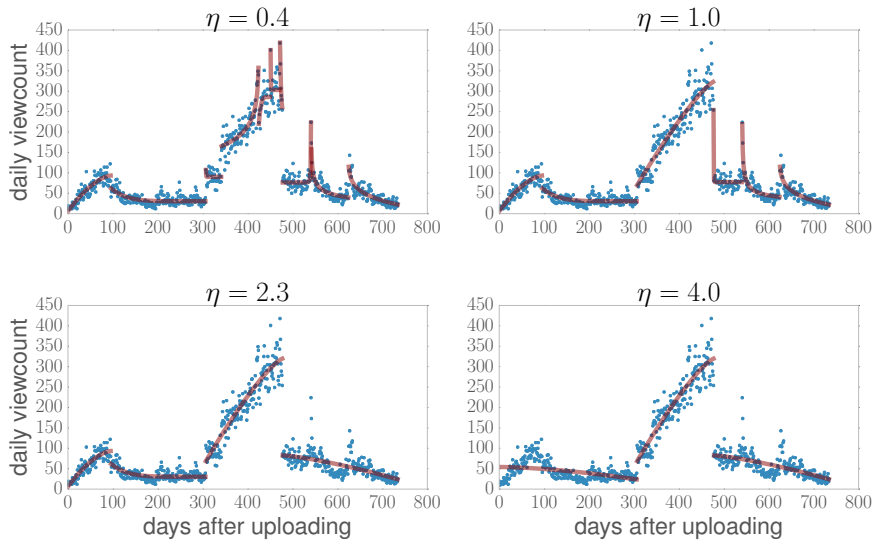


Figure 4.3: Segmentation and curve-fitting result for YouTube video `_3enGWVdgJo` with different ηs. This shows the effect of η in controlling the trade-off between fitting error and the number of phases.

η to the iteration step (4.13),

$$\tilde{E}^*(t') = \min_{t^s_{n'}, \theta_{n'}} \tilde{E}^*(t^e_{n'-1}) + E\{x[t^s_{n'} : t'], \theta_{n'}\} + \eta \tag{4.16}$$

and also modify step (4.14) accordingly.

The effect of η can be seen in Figure 4.3, where the same video has been segmented into ten phases with $\eta = 0.4$ (over-segmentation); only three phases with $\eta = 4.0$ (under-segmentaion), while $\eta = 2.3$ produces four phases, which seems to follow the long-term trend. We acknowledge that the notion of *phases* is inherently subjective, and each video may have multiple "good" representations — when $\eta = 1.0$, the example video is segmented into six phases, this also seems a plausible segmentation with more details after $t = 500$. We will choose η systematically in Section 4.5.

The run time for step (4.13) above is $O(T\Gamma(T))$, where $O(T)$ is the time for searching over $t^s_{n'}$ and $\Gamma(T)$ is the T-dependent time complexity of power-law curve fitting and finding $\theta^*_{n'}$. The complexity of this entire dynamic programming algorithm is

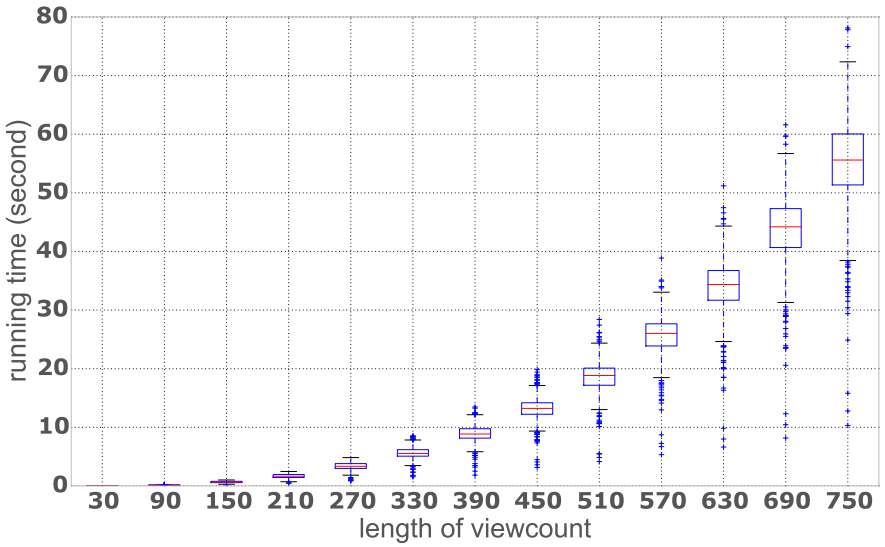


Figure 4.4: Running time of the PHASE-FINDING algorithm on viewcount series of different lengths. Each boxplot contains the running time of segmenting 300 randomly sampled viewcount histories. The empirical run time of the algorithm is approximately $O(T^3)$. Curvefitting shows that when the viewcount length T is large enough, the median of running time is about $1.4 \times 10^{-7} T^3$.

hence $O(T^2\Gamma(T))$. This algorithm was implemented in C++[2]. To evaluate its speed, 300 viewcount histories were randomly sampled and the run time of segmenting them with various lengths was measured. The results are shown in Figure 4.4. We can see the empirical run time of this algorithm is about $O(T^3)$. The throughput for finding phases in one-year long viewcount sequences is about 400 per CPU per hour.

4.4 Comparison with classical methods

Keogh et al. [2004] have summarized most of the classic time series segmentation methods into three types: "sliding windows", "top-down", and "bottom-up", where the curves used to model each phase are straight lines. In this section, we compare our method with each of these three to show why our method is more suitable for video viewcount analysis.

[2]The source code is published at https://github.com/yuhonglin/segfit

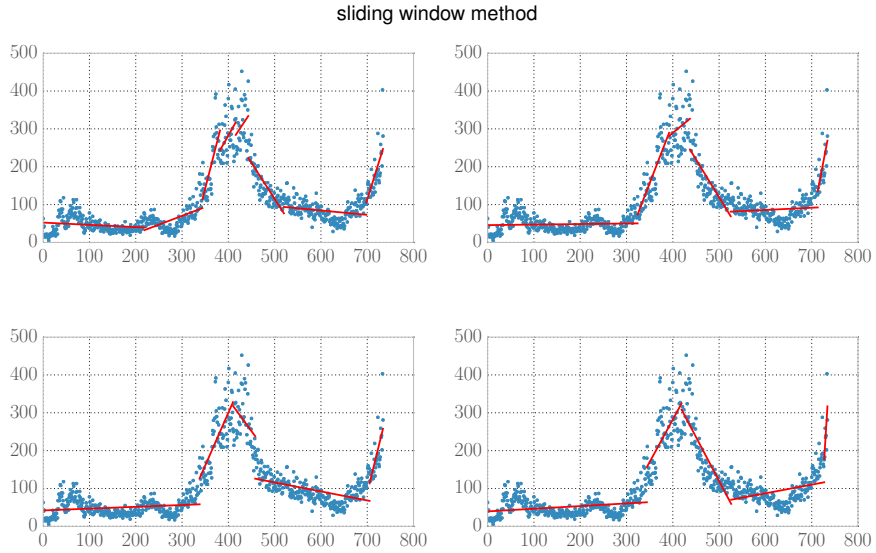


Figure 4.5: Phase-fitting of the classic "sliding window" method using various thresholds. Blue dots: the daily viewcount of video `_3enGWVdgJo` (x-axis is the video's age in days and the y-axis is the daily viewcount). Red lines: the phases found. The declining phase after the peak at about $x = 400$ is clearly nonlinear and this algorithm fails to capture it.

4.4.1 Performance of sliding window method

"Sliding window" algorithms are online time series segmentation methods which follow a segment from left to right and fit a curve at each step. A new phase boundary is found when the fitting error exceeds some error bound. The same procedure is repeated on the rest of the data to find following phases. We have implemented the sliding window method and applied it on the same video as in Figure 4.3 using various parameters. From the results (Figure 4.5), we can see that the biggest flaw of the sliding window method is that the boundaries can be different even when the number of phases are the same. (see the plots in Figure 4.5). This makes the results very sensitive to the hyper-parameters (the fitting error bound) and the algorithms is not robust and very hard to tune.

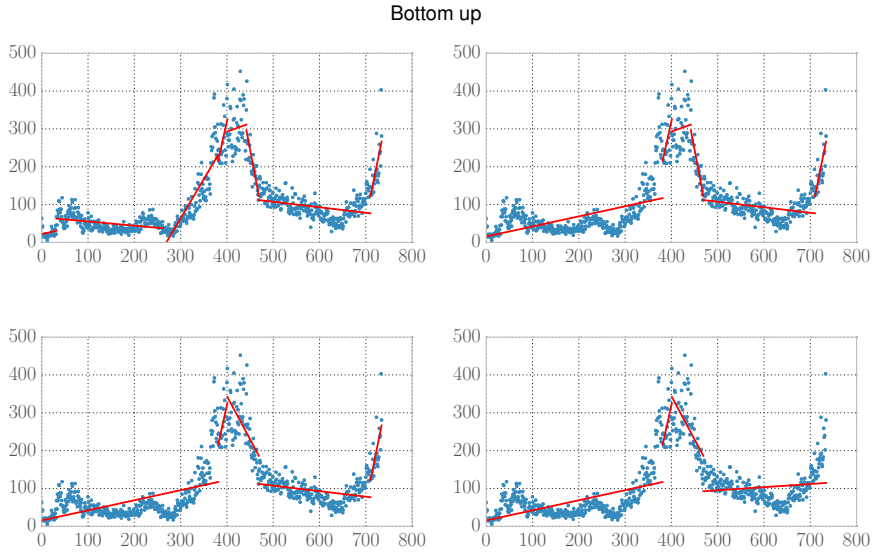


Figure 4.6: Phase-fitting of the classic "bottom-up" method using various thresholds. Blue dots: daily viewcount of video `_3enGWVdgJo` (x-axis is age of video in days and y-axis is daily viewcount); Red lines: phases found. Note that the method fails to capture the nonlinear phase from about $x = 400$ to $x = 600$.

4.4.2 Performance of "bottom-up" and "top-down" methods

According to Keogh et al. [2004], the "bottom-up" algorithm first splits the time series into a series of shortest segments and then merges them until some criterion is met, often that the total fitting error is beyond some bound. In comparison, the "Top-down" algorithms follow the opposite direction in that they start from assuming there are no segments (or one segment as the whole series) and then recursively partition it until some stopping criterion is met. These two algorithms can be very fast and have many applications in diverse fields. We also implemented these two methods and the fitting results can be seen in Figures 4.6 and 4.7.

From the figures above, we can see that, whatever algorithms and hyper-parameters we use, straight lines can not capture the non-linear rising/falling patterns. We have investigated many other examples and the conclusions are the same.

Not only are straight lines not flexible enough to capture the non-linearity of phases, due to their "greedy" nature, bottom-up and top-down algorithms also can not always find the optimum boundaries as illustrated in Figure 4.8.

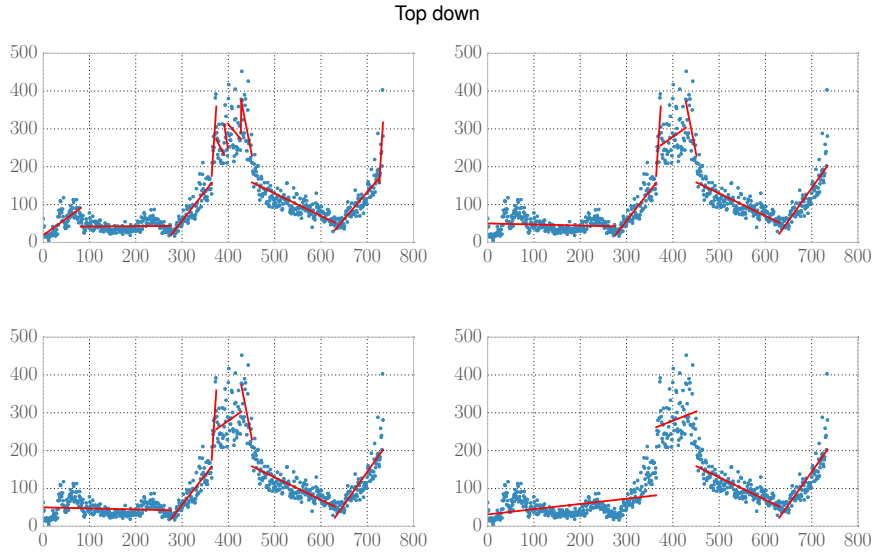


Figure 4.7: Phase-fitting of the classic "top-down" method with various thresholds. Blue Daily viewcount of video `_3enGWVdgJo` since uploading. Blue dots: the daily viewcount of video `_3enGWVdgJo` (x-axis is video's age in days and y-axis is daily viewcount); Red lines: the phases found. Same as "sliding-window" and "bottom-up" methods, it can not model the phase from about $x = 400$ to $x = 600$.

In summary, the classic algorithms are not suitable for viewcount phase detection mostly because,

- Straight lines cannot capture the non-linearity of viewcount phases.

- Greedy algorithms cannot always find the optimum boundaries.

The reasons are that, the classic algorithms are mostly used on very long time series, where the goal is usually compression or removing noise. However, in modelling viewcount dynamics, we need to be as accurate as possible in fitting each phase and determining the boundaries. Moreover, viewcount series are often not very long (< 1000). Therefore, to solve this problem, we have proposed "power-law curves" + "dynamic programming".

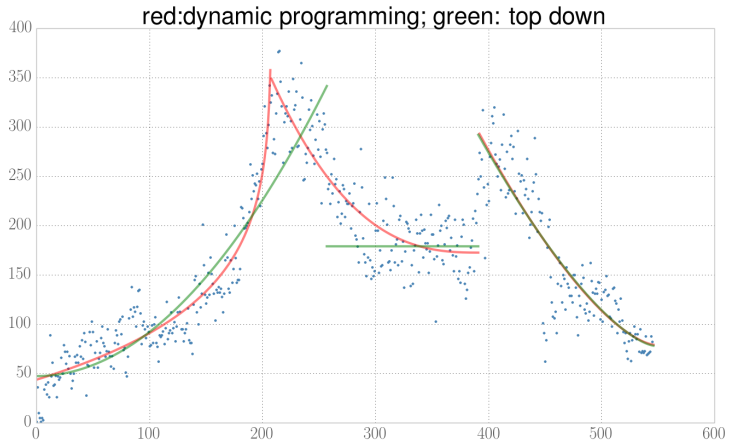


Figure 4.8: The blue points: daily viewcount of video `XQr3Oiu3C1M`. Red curves: phases found by dynamic programming; Green curves: phases found by top-down algorithm. We can see that the green curve is apparently worse. This is because due to its greedy nature, the top-down algorithm first "cut" the series around $x = 260$ which is not a global optimal cut.

4.5 Evaluation and parameters tuning

As seen in Figure 4.3, hyper-parameter η controls the trade-off between fitting each phase well, and having a reasonable number of phases. To systematically evaluate the algorithm and select an η, I have built a website and asked 6 people to label the segments of 210 videos randomly sampled from our dataset. I assigned the videos to these "labelers" in a way that each video was labeled by two researchers. Although the phases of some viewcount series are very clear, in general, many viewcount series had phases which were hard to decide upon, even by humans. Moreover, unlike other classification problems commonly met in image and natural language processing, the goal of detecting phases is not perfectly imitating humans. Labellers' judgements can be considered to be "guidance" rather than "golden rules". Starting from these ideas, I defined two sets of boundaries from the ground truth to help meaningfully evaluate the algorithm,

- $B_\cap$ The set of boundaries both of the labelers agree on ("intersection")
- $B_\cup$ The set of boundaries either of the labelers agree on ("union")

ϵ	Agreement
0	60.2%
1	69.3%
2	70.9%
3	72.0%

Table 4.1: Agreement between labelers on the boundaries in terms of different ϵ

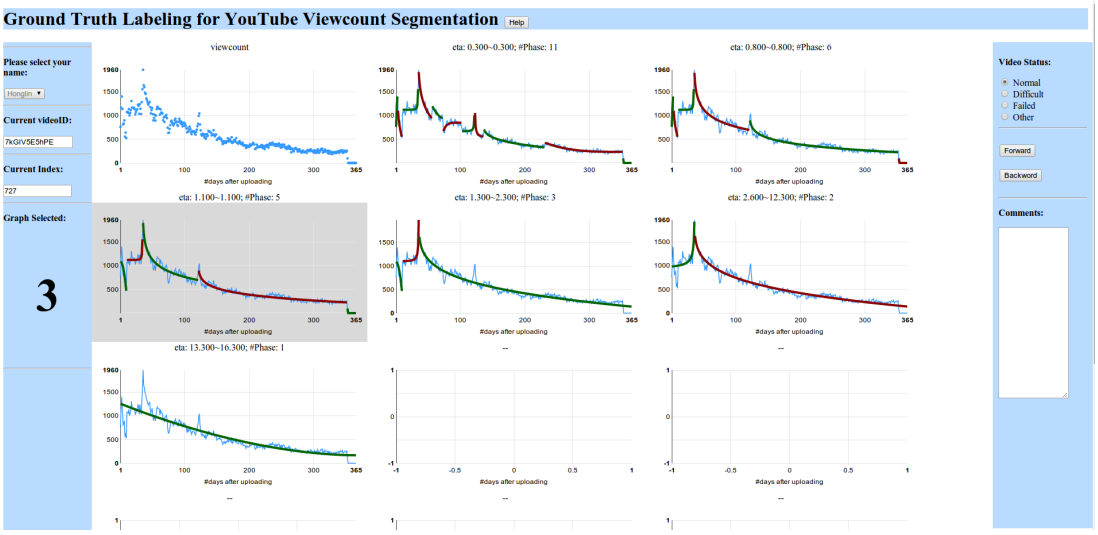


Figure 4.9: The screen shot from the website we have built to label the ground truth. We used the segmentation results with various η as candidates to make the labelers' decision easier. Every user had her own account and their selections were recorded on the server. In this graph, the user has selected the third segmentation whose plot is shaded and whose index is shown in the left column.

I use a "date radius" ϵ to help define the equivalence of two boundaries d_1 and d_2: $d_1 = d_2$ iff $|d_1 - d_2| \leq \epsilon$. The degree of agreement between the 6 labelers can be found in Table 4.1.

To be safe, we only use boundaries in $B_\cap$ as positive instance in ground truth, and define precision and recall as follows,

$$\begin{aligned} recall &= \frac{|B(\eta) \cap B_\cap|}{|B_\cap|} \\ precision &= \frac{|B(\eta) \cap B_\cap|}{|B(\eta)|} \end{aligned} \tag{4.17}$$

Then the precision-recall curve in terms of different η is as follow,

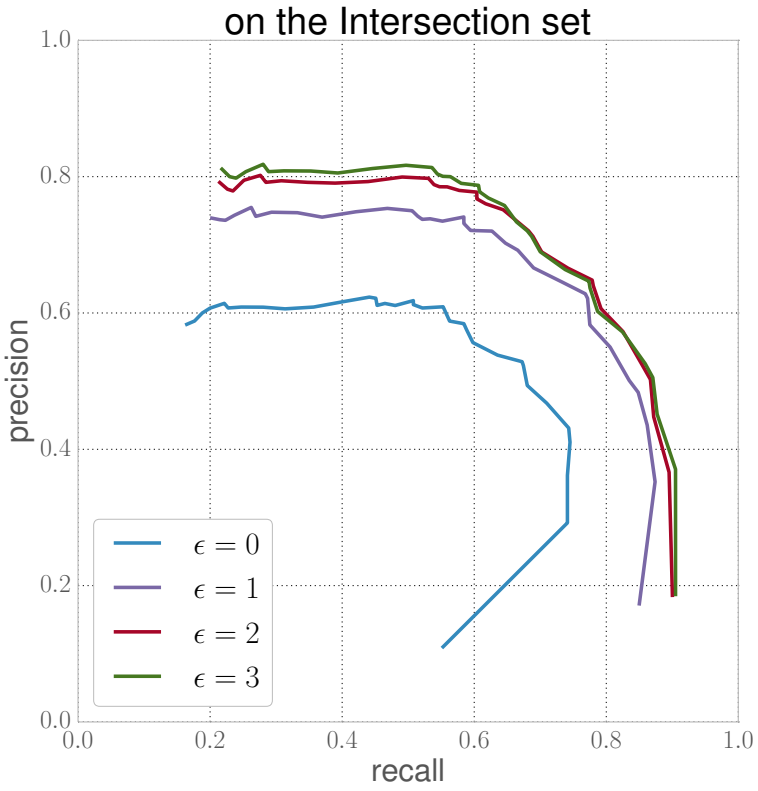


Figure 4.10: The precision-recall (defined in Equation 4.17) curves in terms of different η. It can be seen that the recall scores decreases along the x-axis, which is unusual. This is because the variable here (η) is not a threshold on some scores of predictors, in which cases the recall is always non-decreasing when threshold decreases. In our case, if η decreases, we will find more phases, but the boundaries previously correctly found may change, causing recall to decrease. (We can also see that recall decreases more obviously when ϵ is smaller because then the correctly found boundaries are easier to rule out.)

Based on the Figure 4.10 and Table 4.2, we can see that, by line searching, the segFit algorithm can achieve F_1 scores larger than 0.7 when $\epsilon = 2$ and $\eta = 2.3$. Unless otherwise mentioned, all the following results are based on $\eta = 2.3$.

4.6 Summary

In this chapter, we have proposed phases as a new description for the burst-like popularity lifecycle of a video, and have presented a method to extract phases from popularity history. In other words, we have found a way to simultaneously segment

ϵ	Recall	Precision	F_1
0	0.673	0.529	0.591
1	0.768	0.628	0.691
2	0.779	0.648	0.707
3	0.774	0.647	0.704

Table 4.2: Recall and precision scores of PHASE-FINDING algorithm when corresponding F_1 score is maximized (based on dataset $B_\cap$).

and recover power-law shapes without needing to determine the number of phases beforehand[3]. We have also evaluated the method's performance against the ground truth labelled by 6 persons and the results show that the algorithm performs well. In the next chapter, we will look more deeply into viewcount dynamics based on their phases.

[3] *SegFit*, a fast implementation of our algorithm in C++: https://github.com/yuhonglin/segfit

Chapter 5

Properties of Popularity Phases

"How many *phases* must a *video* walk
down, before you can call *it viral*..."

— Modified from lyrics of
Blowin' in the Wind, by *Bob Dylan*

In this chapter, we present statistical descriptions for 172K+ videos over 2 years, measuring their phases, content types, and popularity evolution. The data show that the number of phases is strongly correlated to a video's popularity. Moreover, videos of different categories exhibit very different phase profiles. For example, in our dataset, nearly 3/4 of videos in the top 5% of popularity have 3 or more phases, whereas only 1/5 of the same number of least popular videos do; More than 60% of news videos are dominated by one long power-law decay, whereas only 20% of music videos do. By introducing the notions of popularity phases, this work exposes the rich rising and falling patterns of popularity dynamics and their close relationship to videos' popularity and content types.

5.1 The 2-year popularity dataset

This chapter is based on a large and diverse dataset of YouTube videos created from Twitter feeds. Video links were extracted from a large Twitter dataset (Yang and Leskovec [2011]) of 184 million tweets from June 1st to July 31st in 2009, about 20-30% of the total tweets in this period. After extracting URLs from all the tweets, the

Table 5.1: The number of videos tweeted in June and July 2009 broken down by user-assigned categories.

Category	**#videos**	**Category**	**#videos**
Music	64096	Howto	4357
Entertainment	26602	Travel	3379
Comedy	14616	Games	3299
People	12759	Nonprofit	2672
News	10422	Autos	2398
Film	8356	Animals	2375
Sports	7872	Shows	407
Tech	4626	Movies	15
Education	4577	Trailers	13
Total number: 172841			

shortened URLs were resolved, retaining those referring to YouTube videos. This yielded 402,740 unique YouTube videos, among which 261,391 videos are still online and have their meta-data publicly available. Videos that had less than 50 views in their first two years were removed (since they did not have enough views to meaningfully extract phases), giving a final dataset of 172,841 videos.

For each video v, we obtained from YouTube API[1] its metadata such as category, duration and uploader as well as its daily viewcount series. We analysed the videos for up to two years after posting, i.e., $T = 735$ days. Compared to related recent work, this dataset is notable in two respects. First, in terms of data resolution, most prior work has used a 100-point interpolated cumulative viewcount series over the lifetime of a video (Figueiredo et al. [2011]; Ahmed et al. [2013]; Borghol et al. [2012]). In contrast, this dataset is one of the first to contain fine-grained history of daily views. Secondly, in terms of the time-span, recent work has examined popularity history either over a video's first month (Szabo and Huberman [2010]; Pinto et al. [2013]; Abisheva et al. [2014]) or over 1 year (Crane et al. [2008]). This dataset is the first to allow longitudinal analysis over multiple years.

To derive popularity phases, two main co-variates were used in our measurement study: video popularity percentile and content category. Table 5.1 summarizes the number of unique videos per user-assigned category in this dataset. The video dis-

[1] https://developers.google.com/youtube/

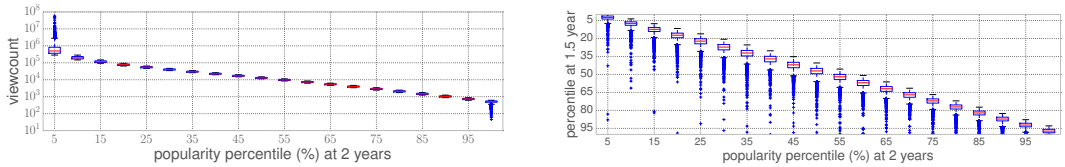


Figure 5.1: Left: Boxplots of video viewcounts at $T = 735$ days, for *popularity percentiles* quantized at 5% (8000+ videos). Viewcounts of the 5% most- and least- popular videos span more than three orders of magnitude, whereas videos in the middle bins (from 10 to 95 percentile) have viewcounts within 30% views of each other. Right: The change in popularity percentile (y-axis 0% to 100%) from 1.5 years to 2 years (x-axis, in 5% bins). While most videos retain a similar rank, video of almost any popularity at 18 months of age could *jump* to the top 5% popularity bin before it was 24 months old (left-most boxplot).

tributions over categories are very similar to Table 3.1. Also similar to Figure 3.3, all videos were ranked by the total viewcounts they had received by age 735-days (Figure 5.1:Left). The rank for each video was converted to a percentile scale, i.e. video v at 1% will be less popular than exactly 1% ($\sim$1720) of other videos in the collection. We can see that although the videos here are extracted only from the tweets in June and July 2009, they are strongly representative in that their distributions with respect to content categories and popularity percentiles are almost the same as the whole dataset explored in Chapter 3.

Figure 5.1:Right shows the change of popularity from 1.5 years (y-axis) to 2 years (x-axis). While most videos retain a similar rank, a video from any bin can *jump* to the top popularity bucket in 6 months (as seen in the left-most boxplot). One can asks: how did these videos go viral? We will present some observations in Section 5.7.

5.2 Total number of phases

First, let us examine how many phases a video can have. Figure 5.2 gives the frequencies of videos having different number of phases. We can see that, except when #phase=1, the distribution is roughly linear (when viewcount is plotted logarithmically). More than half ($\approx$ 55.4%) of the videos have $\geq$ 3 phases.

Figure 5.3(left) breaks down videos in each popularity bin by the number of

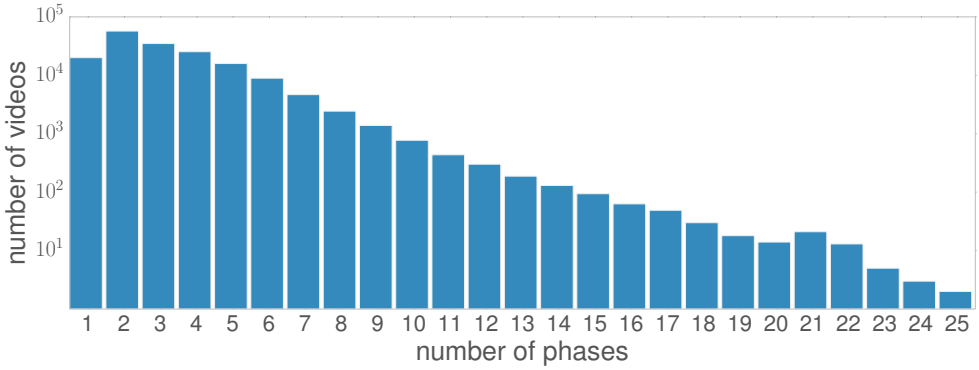


Figure 5.2: Distribution of videos with different number of phases.

Category	avg #phase	popbin	avg #phase
Education	3.53	1	3.52
Music	3.53	3	3.46
Howto	3.37	5	3.43
Comedy	3.32	7	3.39
Entertainment	3.14	9	3.35
Sports	2.99	11	3.31
Tech	2.84	13	3.21
Nonprofit	2.76	15	3.14
Shows	2.5	17	3.0
News	2.25	19	2.76

Table 5.2: Average number of phases detected according to different category and popbin. Top rows contain videos that are most likely to have persistent values (e.g. "Education", "Music" etc.). Popular videos are more likely to have more phases than the unpopular ones.

phases they contain, and Figure 5.3(right) does the same for each content category. We can see that among the top 5% most popular videos, more than 95% have more than one phase, and about 45% have four or more phases. As a general trend, more popular videos have more phases (and hence a complex lifecycle). In terms of different content categories, over 70% of news videos have only one or two phases, whereas videos related to art and entertainment (*music, comedy, animal, film, entertainment*) have the most complex lifecycles. We see that the need to view a *news* item decreases drastically after a few days, whereas arts and entertainment content retains

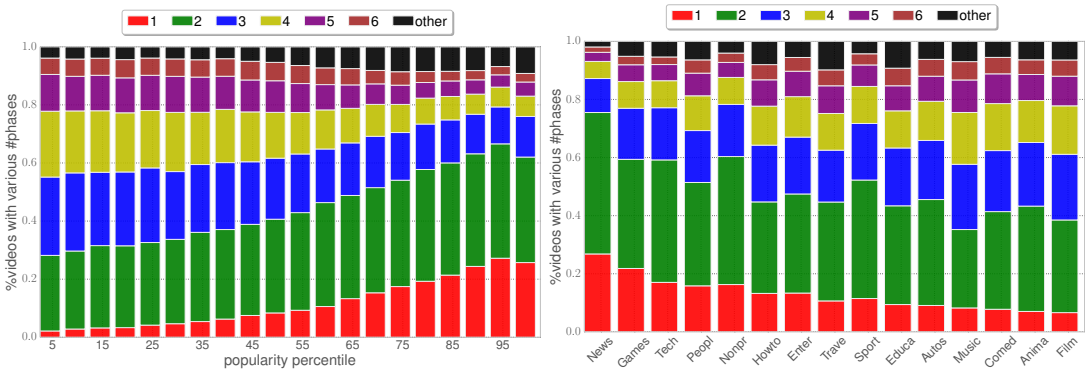


Figure 5.3: Percentage of videos broken down by the number of phases they have, over Left: popularity percentiles and right: video categories. A general trend is that popular videos and entertainment content (e.g. *music* videos) have more phases overtime, and more than half of *news* videos and the least popular videos have one dominant decreasing phase.

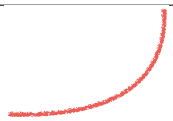
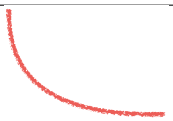
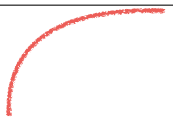
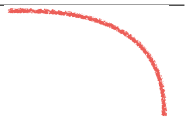
interest over time and is suitable for re-consumption.

5.3 Phase types

We intuitively categorize power-law phases into four types, according to whether the trend over time is increasing or decreasing, and the rate of change is accelerating or decelerating. These four types correspond to convex and concave curves (see the shape sketches in Table 5.3) when the trend is either increasing or decreasing. Furthermore, each type can be uniquely identified by three parameter combinations: the power-law scaling factor $a > 0$ or < 0 (short-handed as $+/-$); exponent $b > 1$ or < 0 or within $[0, 1]$; and the temporal direction of τ as in Equation (4.2) (abbreviated as $\leftarrow$ or $\rightarrow$).

There are 563,624 phases in total, with an average of 3.3 phases per video. Table 5.3 presents a profile of these shapes. We can see that roughly half the segments are convex-decreasing – these phases span more than 60% of the duration and account for less than half the viewcounts. Convex-increasing is the second-most common shape, accounting for another 30% of segments, while concave-decreasing is the least common.

Table 5.3: Four types of phase shapes and their basic statistics.

Phase-type	Convex increasing	Convex decreasing	Concave increasing	Concave decreasing
Shorthand	**vex.inc**	**vex.dec**	**cav.inc**	**cav.dec**
Sketch				
Parameter $(a;b;\tau)$	$+;\ >1;\rightarrow$ $+;\ <0;\leftarrow$ $-;$ [0,1]; $\leftarrow$	$+;\ <0;\rightarrow$ $-;$ [0,1]; $\rightarrow$ $+;\ >1;\leftarrow$	$+;$ [0,1]; $\rightarrow$ $-;\ <0;\rightarrow$ $-;\ >1;\leftarrow$	$-;\ >1;\rightarrow$ $+;$ [0,1]; $\leftarrow$ $-;\ <0;\leftarrow$
Phase count	172,329 (30.6%)	286,070 (50.8%)	67,862 (12.0%)	37,363 (6.6%)
Length (days)	3.0×10^7 (23.7%)	8.2×10^7 (64.6%)	1.0×10^7 (8.0%)	4.6×10^6 (3.7%)
Views	3.5×10^9 (28.2%)	5.8×10^9 (46.8%)	2.2×10^9 (17.4%)	9.6×10^8 (7.6%)

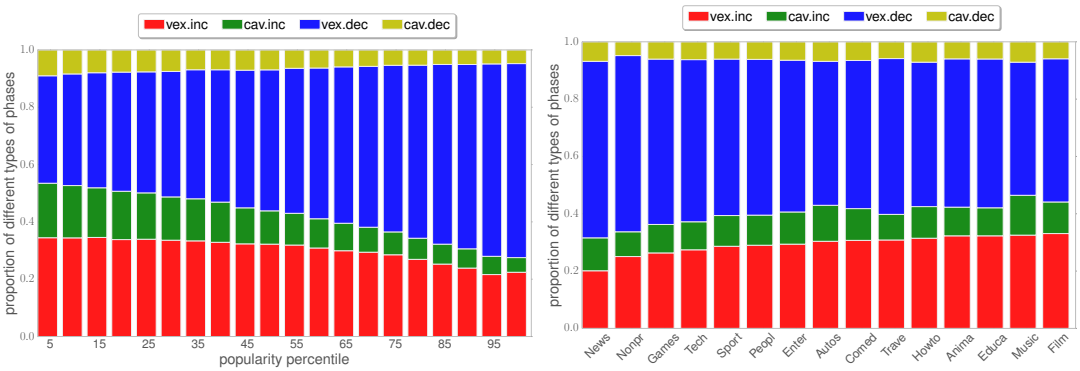


Figure 5.4: Percentage of the four phase types, broken down by popularity bins (left) and content categories (right).

Figure 5.4 reports the fraction of each of the four types of phases classified by popularity bin and category. Overall, popular videos have a greater number of increasing phases (both convex and concave, 53.5%, see Figure 5.4(left)), with this faction decreasing to 27.5% for the least popular videos. Across different content categories, *News* has the least number of increasing phases, while entertainment and instructional videos (such as *Music*, *HowTo* and *Autos*) have the greatest fraction of increasing categories ($\geq$ 42%). This is also explained by the long-lasting value of Entertainment and HowTo videos (e.g., see the viewcount periodicity of the air-conditioner venting video in Figure 4.1(d)).

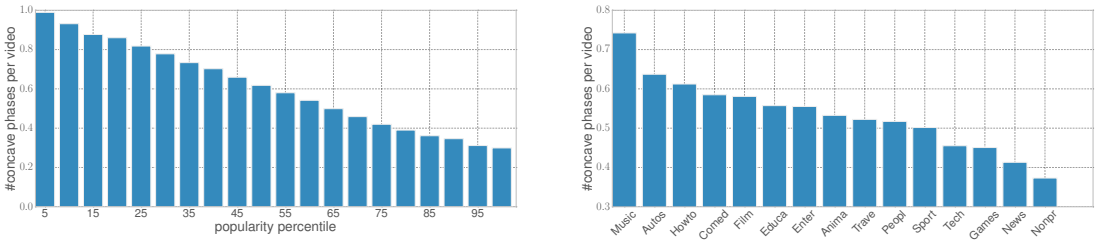


Figure 5.5: Number of concave phases per video with respect to popularity percentiles (left) and video categories (right).

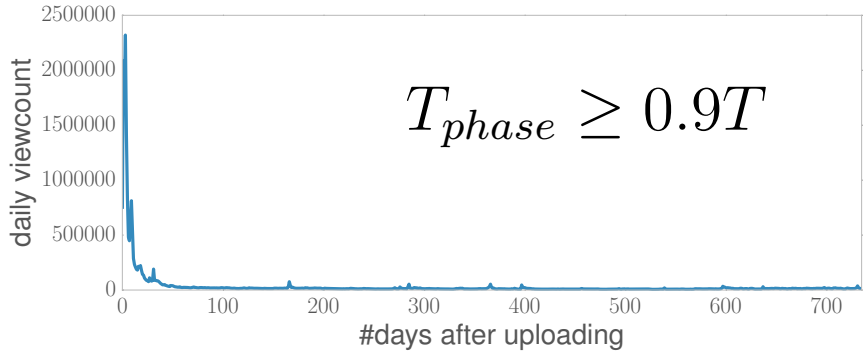


Figure 5.6: The viewcount history of one video with a dominant convex-decreasing phase.

5.4 New phase shapes

From Table 5.3, we can see that, although the phases are predominantly convex, there are still a non-trivial number of concave phases. These concave shapes can not be explained by Crane and Sornette's model and call for further research.

What is particularly interesting about Figure 5.5 is that there is a clear trend for popular videos to have more concave phases. Moreover, "Music" videos also have more concave phases than "News" videos. These observations deserve further investigation as they probably reflect a real difference between how people behave collectively in their use of different types of online media.

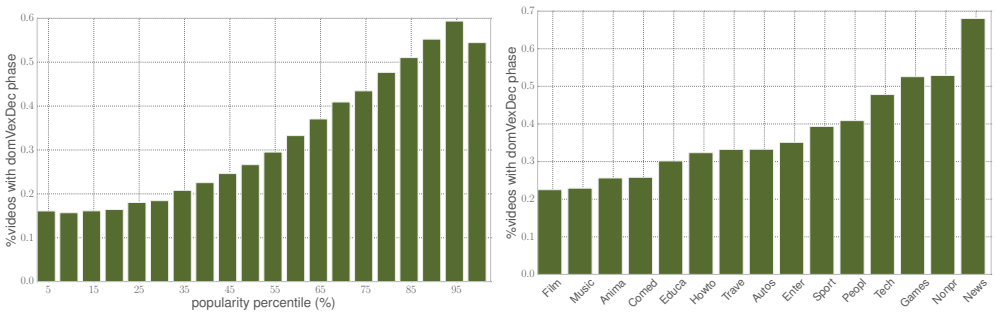


Figure 5.7: Percentage of videos with dominant power-law decreasing phases, broken down by popularity bins (left) and content categories (right).

5.5 Dominant decreasing phases

We now investigate videos that have *dominant convex decreasing phases* (as shown in Figure 5.6) whose timescales are longer or equal to 90% of their entire history, i.e., $t^e - t^s \geq 0.90T$. These videos typically receive a burst of attention from an exogenous shock (e.g. News), and then cease to attract further attention, such as in Figure 4.1(b). Figure 5.7 plots, by popularity bin and content category, the fraction of videos which have a dominant decreasing phase. We can see that more than 60% of *News* videos have a dominant decreasing phase. In other words, more than half of News videos do not start a new phase after having had a main attention-getting shock. On the other hand, only ∼20% of *film* and *music* contain a dominant decreasing phase, with the remaining 80% enjoying "revivals" of attention over their life-cycles. Perhaps not surprisingly, having a single dominant decreasing phase is at odds with being high on the popularity scale – over 50% of the least popular videos have such a phase, while only about 15% of the most popular videos do so. However this data also points to the inherent unpredictability of popularity: despite having just one long decreasing phase, 0.75% of all videos, or ∼1275 of them, still made it to the top 5% in the popularity chart.

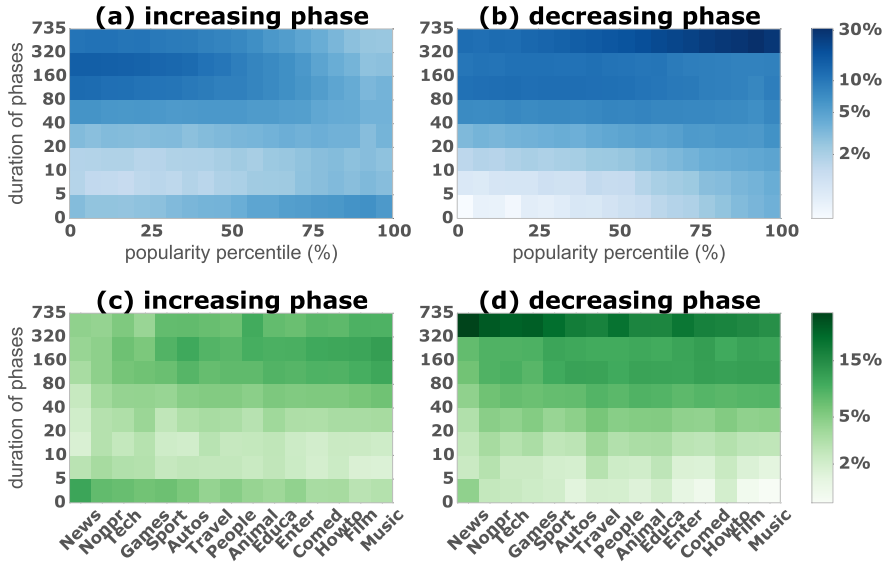


Figure 5.8: Distribution of phase durations. X-axis: covariates – popularity percentile (20 values) and 15 content categories. Y-axis: duration in days (log-scaled bins). Color intensity: the fraction of phases having property x and duration y. We can see from (a) that popular videos have long and sustained ($>$ 100 days) increasing phases, and from (b) that unpopular videos have longer decreasing phases ($>$ 300 days). In (c), entertainment-related videos are more likely to have long increasing phases. In (d), while *news* videos have by far the most amount of decreasing phases over a year (also see Figure 5.7), long decreasing phases exist across all categories.

5.6 Phase lengths

Figure 5.8 examines the distribution of phase durations, broken down into increasing and decreasing phases, with popularity and category as co-variates. In Figure 5.8(a), we can see that popular videos tend to have longer increasing phases, while the increasing phases for videos in the least popular bins tend to be short. In Figure 5.8(b), while there is a fair amount of long ($\geq$ 160 days) decreasing phases across the entire popularity scale, the least popular videos are still the most likely to have a long and dominant decreasing phase, which is consistent with Figure 5.7(left). In (c) and (d), on the other hand, we can see that the probability of having longer phases of either type spread over different categories. *Music* is slightly more likely to have longer increasing phases than other categories, while *News* is more likely to have a decreasing phase lasting more than 320 days, consistent with Figure 5.7(right).

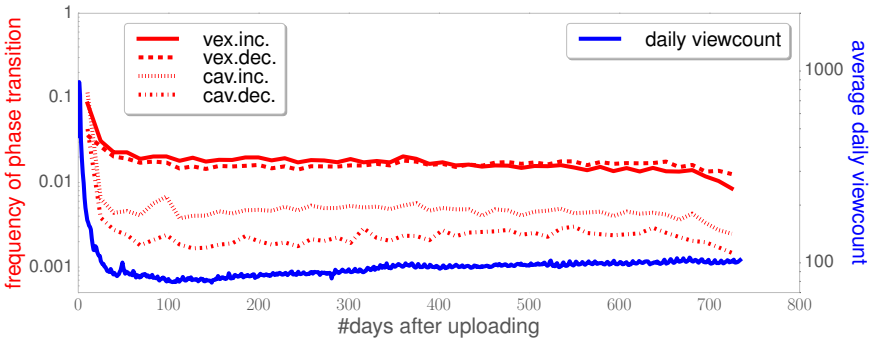


Figure 5.9: Red: The probability, in a 15-day interval, of a video entering a new phase broken down by phase type. Blue: Average daily viewcount for all videos.

5.7 Phase over time

Since new phases tend to be triggered by external events, one may ask whether older videos attract less attention – in other words, are they *forgotten*? Surprisingly, the data says no. Figure 5.9 plots, by age of a video, the likelihood it will enter a new phase (calculated as the probability it will change phase over a 15-day period). The red curves show the four different phase types, and, for comparison, the blue curve shows the total attention (number of views) received by all videos as they age. We can see that after an initial period of about 90 days where there is a high probability of receiving new phases and views, $\sim$ 2% of the 172K videos change, in any given 15-day period, to a new convex increasing or decreasing phase. It is notable that (1) this trend holds constant from about 3 months to 2 years into a video's lifecycle and (2) the number of new convex-increasing phases is roughly the same as the number of new convex-decreasing phases, despite the latter being much more popular (see Table 5.3). The same temporal trend holds for concave-increasing or decreasing phases, except with lower incidence.

5.8 How do videos become viral?

Figure 5.10 explores the relationship of the most popular videos and the phases they went through over time. We examine the top 5% (or 8,642) videos at 180, 360,

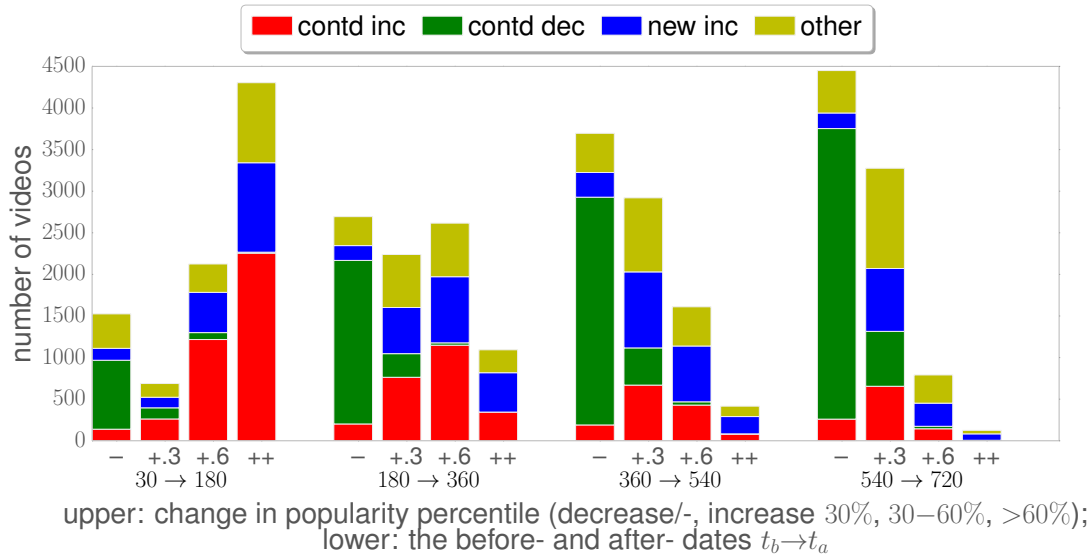


Figure 5.10: Evolution of the most popular videos according to popularity and phase history. See Section 5.8 for explanation and discussions.

540, and 720 days of age (called the "after" dates t_a), and collect statistics about their popularity percentile on a "before" dates t_b about 6 months prior to t_a, and the phases that are present between t_b and t_a. We graph the data according to four types of change in popularity percentile decreased (-), increased by (0-0.3%, 0.3%-0.3%, or >0.6%); and four types of phase history: having (one) continued increasing phase, continued decreasing, with at least one new increasing phase, and other (one or more decreasing phases). We can see that the most popular, "viral" videos are highly volatile, with more than half *jumped* more than 60% in percentile to join this group between 30 and 180 days. Furthermore, new and increasing *phases* plays an important role in the videos whose rank increase significantly. Among the 5,948 videos with improved popularity percentile between 180 and 360 days, for example, only 5% (312 videos) is in a continued decreasing phase, the majority either had a new phase (75%), or are in in a continued increasing phase (20%). In other words, the most popular videos tend to have (new and) increasing phases.

5.9 Phase transitions

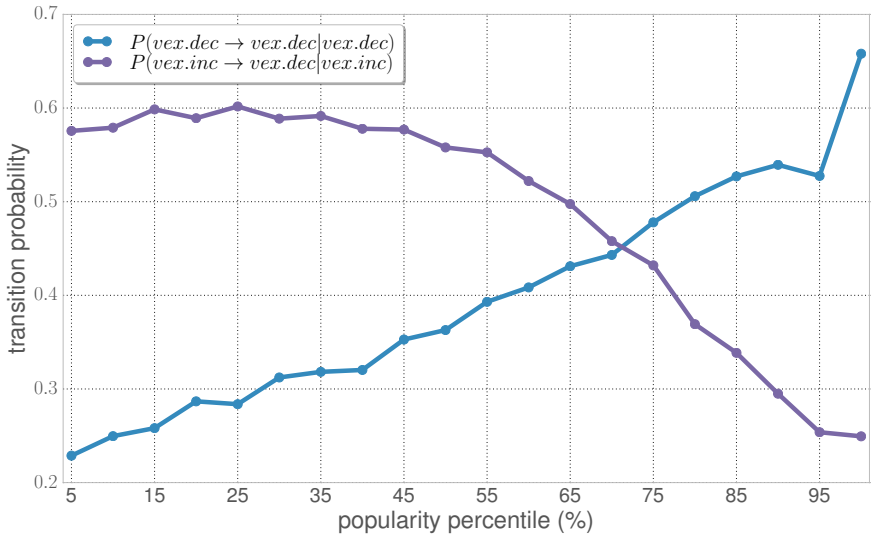


Figure 5.11: Comparison of two kinds of phase transitions for videos of different popularity.

We observed that the statistics of phase transitions are also clearly different for different types of videos. As can be seen in Figure 5.11, for unpopular videos, a convex decreasing phase is more likely to be followed by another convex decreasing phase. For popular videos, a convex increasing phase is more likely to transit into a convex decreasing phase (not to mention that popular videos actually have relatively less convex decreasing phases). To explain these clear trends in viewing patterns, further research is called for.

5.10 Summary

This chapter has presented descriptive statistics of video phases according to popularity and content category. We directly relate phases to popularity, content types, and the evolution of popularity over time. On a dataset containing the 2-year history of over 172,000 YouTube videos, we saw that phases are directly related to content type and popularity change, e.g., nearly 3/4 of the top 5% popular videos have 3 or more phases, whereas only 1/5 of the least popular 5% have that many; More

than 60% of News videos are dominated by one long power-law decay, whereas for Music videos, this number is only 20%; And 75% of videos that made a significant jump to become the most popular videos had been in increasing phases. In general, this chapter has shown that multi-phase representation has the potential to become a tool for understanding the dynamics of other online media, such as hashtags and online memes, and we believe it is a promising avenue to further uncover the laws governing online collective behavior. In the next chapter, phase information will be used as the basis for clustering viewcounts and building predictive models.

Chapter 6

Phase-aware Viewcount Prediction and Clustering

In this chapter, I present two new applications of the viewcount phases represented in Chapters 4 and 5. One of the applications is a new algorithm for predicting future video popularity; the other is a new analytical tool called phase-sketch clustering.

6.1 Phase-aware viewcount prediction

6.1.1 Introduction

Predicting the popularity of an online item is important when dealing with content recommendation, avenue estimation, and the design and evaluation of systems. In the case of YouTube, video viewcount prediction can greatly help smart advertisement deployment, video pre-cache and the like. From the aspect of social science research, prediction is also a good way to investigate the factors which affect viewcount dynamics. In this section, we explore how the segmentation algorithm described in Chapter 4 can be used to improve prediction and we propose a phase-aware method.

6.1.2 Problem formulation

Denoting the daily viewcount of a video v after uploading as $\mathbf{x}[1:T]$ where T is the number of days considered, we want to use the viewcount of $\mathbf{x}[1:t_p]$ to predict the total viewcount in the next Δt days, i.e., $\chi = \sum_{t=t_p+1}^{t_p+\Delta t}(x[t])$ (t_p is called the pivot

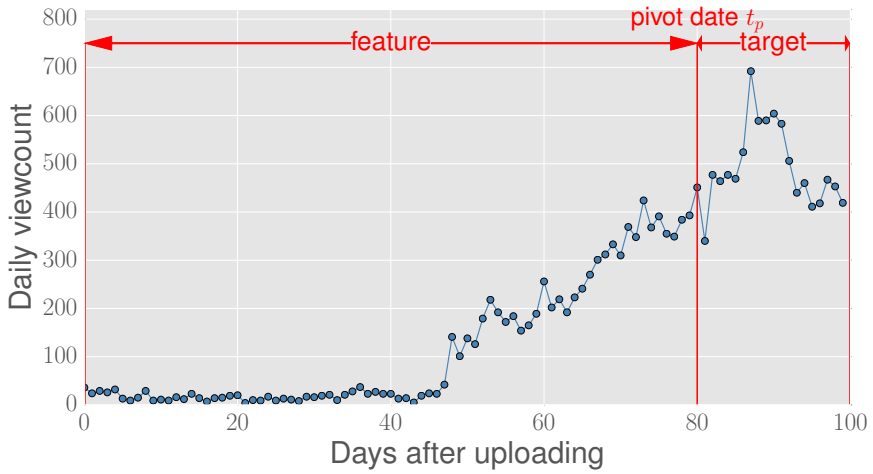


Figure 6.1: Terminology for the viewcount prediction problem.

date, see Figure 6.1). Since the viewcount of videos varies from 10^1 to 10^8, to prevent the popular videos' fitting error from dominating the total error, we first normalize the daily viewcount $\mathbf{x}$ by its maximum value, i.e., let $x_{max} = max\{\mathbf{x}_{1:t_p}\}$, denote $\hat{x} = x/x_{max}$, and define $\hat{\chi} = \chi/x_{max}$. Then we use the normalized MSE (mean-square-error)

$$\epsilon = \frac{1}{\Delta t|\mathcal{V}|}\sum_{v\in\mathcal{V}}(\chi^* - \hat{\chi})^2 \tag{6.1}$$

as the prediction evaluation criterion.

6.1.3 Baseline method

We now choose a linear regression predictor. The prediction output is

$$\chi^* = \mathbf{w}^T\mathbf{x}^* + w_0 \tag{6.2}$$

Here $\mathbf{x}^*$ is a feature vector, and $\mathbf{w}$ and w_0 are weights and a bias term learned from training data (with L2 regularization). That is, the loss function on video set $\mathbb{V}$ is,

$$L(v) = \sum_{v\in\mathbb{V}}(\mathbf{w}^T\mathbf{x}_v^* + w_0 - \hat{\chi})^2 + \alpha(|\mathbf{w}|_2^2 + w_0^2) \tag{6.3}$$

The *baseline* algorithm (Pinto et al. [2013]) uses $\mathbf{x}_{1:t_p}$ as feature vector, and learns

Last shape	Performance	#videos
All	0.3157 ± 0.0701	161,013[1]
convex increasing	0.4074 ± 0.0063	19,176
convex decreasing	0.26227 ± 0.1118	108,784
concave increasing	0.4717 ± 0.0108	16,335
concave decreasing	0.5060 ± 0.0149	16,718

Table 6.1: Performance of multi-linear regression on different videos.

#Segment	Performance	#videos
$1 \leq \# \leq 3$	0.0934 ± 0.1271	58,426
$1 \leq \# \leq 5$	0.1648 ± 0.1202	82,332
$3 \leq \# \leq 5$	0.1388 ± 0.0045	43,381
$6 \leq \# \leq \infty$	0.4239 ± 0.0038	78,681

Table 6.2: Performance of multi-linear regression on videos with different #segments.

one set of parameters $\{\mathbf{w}, w_0\}$ for all videos. The hypothesis behind the baseline is that the weighted average dynamics in the past correlates directly with future popularity. But as we have seen from the many examples before (e.g., Figure 4.1), viewcount dynamics are very diverse. Our hypothesis is that phase shapes plus weighted average history dynamics more strongly correlate with future viewcount. We then make use of phases to help account for such wide diversity in popularity.

6.1.4 Phase-aware viewcount prediction

By exploratory analyses, we found that the shape of the last phase and the total #phases in the first t_p days strongly correlated with the performance of the baseline.

From Tables 6.2 and 6.1, we see that normally, when the last phase is convex and decreasing, model (6.3) performs better than if it is concave and increasing. Also, the prediction result is much better when the training data has less #phases. These facts show that phase information strongly correlate with prediction performance. Our phase-aware prediction utilizes phase information to group similar videos together, allowing better predictions to be made (see Figure 6.2). Based on whether #phases is more than 4, we first split the data into 2 sets[2]. Then for each set, we split it

[1]Some videos are removed as outliers, see Section 6.1.6

[2]In experiments, we found the prediction performance is insensitive to this parameter. We chose 4

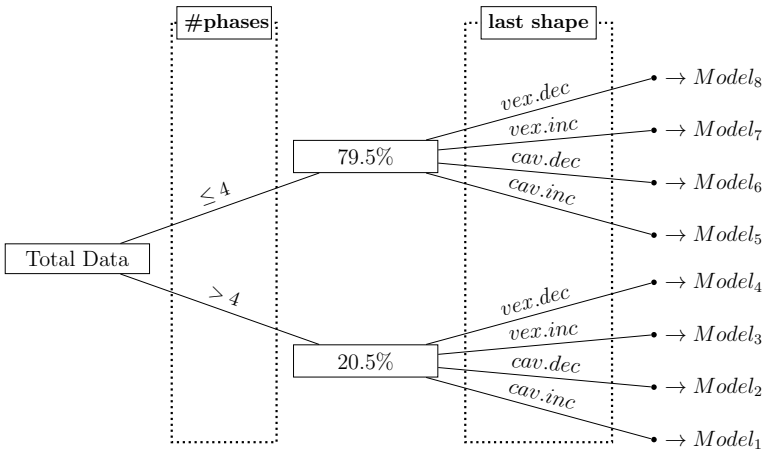


Figure 6.2: Illustration of phase-aware prediction

		#phase ≤ 4 (79.5% videos)			
		vex.inc	**vex.dec**	**cav.inc**	**cav.dec**
15	baseline	0.2450 ± 0.0103	0.0370 ± 0.0038	0.2745 ± 0.0447	0.2402 ± 0.0216
	phase	$\mathbf{0.2232 \pm 0.0093}^{*\dagger}$	$\mathbf{0.0337 \pm 0.0037}^{\dagger}$	0.2614 ± 0.0432	$\mathbf{0.1969 \pm 0.0208}^{*\dagger}$
30	baseline	0.5013 ± 0.0386	0.0852 ± 0.0027	0.5953 ± 0.0562	0.5085 ± 0.0552
	phase	$\mathbf{0.4642 \pm 0.0373}^{*\dagger}$	$\mathbf{0.0771 \pm 0.0011}^{*\dagger}$	0.5734 ± 0.0598	$\mathbf{0.4241 \pm 0.0428}^{*\dagger}$
		#phase > 4 (20.5% videos)			
		vex.inc	**vex.dec**	**cav.inc**	**cav.dec**
15	baseline	0.2555 ± 0.0105	0.1754 ± 0.0075	0.2722 ± 0.0095	0.2676 ± 0.0138
	phase	0.2456 ± 0.0134	0.1745 ± 0.0072	0.2670 ± 0.0090	0.2654 ± 0.0124
30	baseline	0.5146 ± 0.0288	0.3880 ± 0.0095	0.5719 ± 0.0388	0.5633 ± 0.0108
	phase	0.4948 ± 0.0286	0.3865 ± 0.0106	0.5559 ± 0.0321	0.5594 ± 0.0118

Table 6.3: Mean normalized MSE on different video subsets, with $\Delta t = 15, 30$ days, $t_p = 60$ days. * denotes a significant improvement (t-test, $p < 0.05$); † denotes relative error reduction $> 5\%$.

again into 4 sets by their last shape in training data (in total the data is split into 8 subsets). Then we train one predictor on each subset in turn. In this way, our method is adaptive to the phase characteristics of each subset, i.e., we can use larger value of hyper-parameter α on videos with lots of phases (where data is noisy) while using a smaller value for those having few phases (and having a clear trend).

6.1.5 Prediction result

Table 6.3 summarizes prediction performance across all phase-induced subsets. We can see that in all subsets, the phase method reduces the prediction error. The im-

as to make the resulting two datasets not extremely imbalanced but also be able to show the prediction performance of most videos can be significantly improved by my our method.

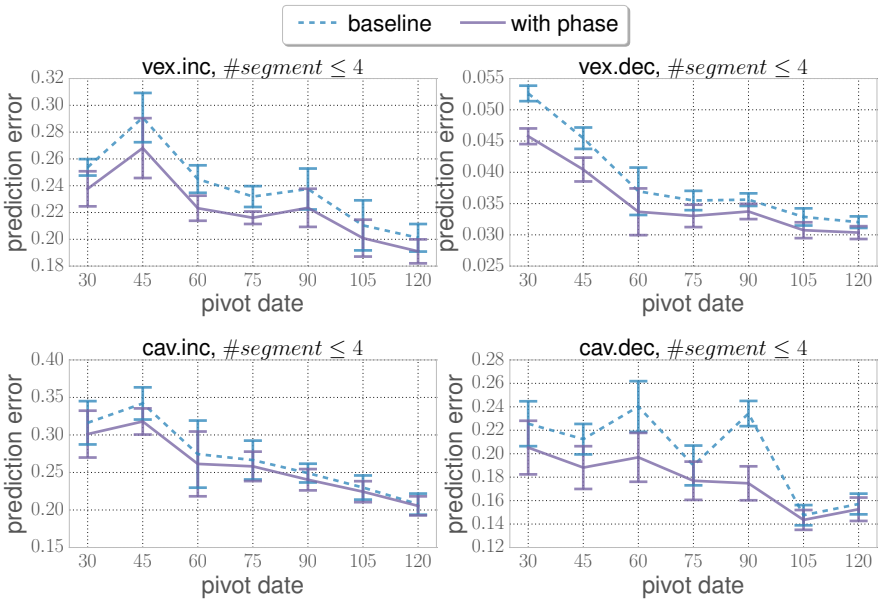


Figure 6.3: Mean normalized MSE for the baseline and phase-aware prediction over different pivot dates (x-axis) for videos with less than or equal 4 phases, broken down by the shape of the last phase of $\mathbf{x}_{1:t_p}$, Δt=15 days.

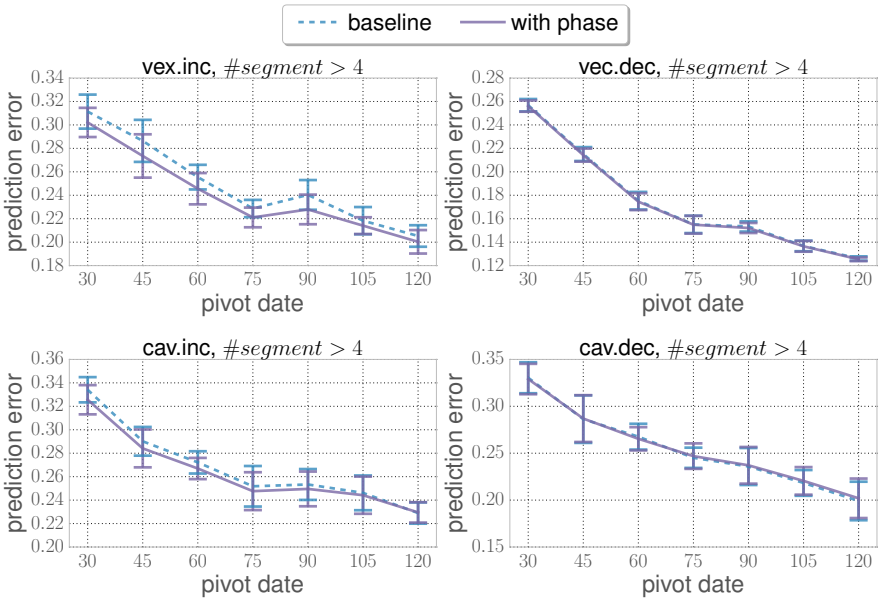


Figure 6.4: Mean normalized MSE for the baseline and phase-aware prediction over different pivot dates (x-axis) for videos with more than 4 phases, broken down by the shape of the last phase of $\mathbf{x}_{1:t_p}$, Δt=15 days. It can be seen the performance improvement (smaller the better) is much smaller than that in Figure 6.3

provement is most significant for videos that end in convex-increasing or concave-decreasing phases – this shows that the additional phase information indeed helps predict future viewcount. Both methods yield higher error when the number of phases is > 4 and the performance improvement is also much smaller (Figure 6.4) – this is the small fraction of videos with highly complex dynamics, indicating that predicting popularity is still a challenging problem. Figure 6.3 shows prediction performance across pivot dates of 30 to 120 days when the number of phases each video has is less than or equal 4 – the *phase* method outperforms baseline significantly in all cases. Figure 6.5 (a)(b)(c) contains representative examples where the phase predictor works much better than the baseline, attributable to the phase change that happened just before the pivot date; (d) contains an example where the baseline works better, in this case because the sharp decline just before the pivot date seems to be noise on a long-term rising trend, rather than a new phase.

6.1.6 Difficult cases

In our analysis, no matter what kinds of model/features we used, there were always some videos with extremely large fitting errors. To explore these cases, we did a 2-fold cross-validation on the dataset and Figure 6.6 shows the distribution of fitting error (in testing data) when the classic multi-linear regression was used.

After examining the videos with largest prediction errors, it was found that, these videos all had at least one "viewcount jump" during the "target time range". To avoid these kinds of videos from dominating the prediction measures, we removed them as outliers by using the following criterion,

$$\frac{\max(x[\tau : t])}{\max([x[1 : \tau]])} < \alpha \tag{6.4}$$

In the evaluation report in Table 6.3, we set $\alpha = 2$ to remove 3.6% of the total data as outliers. These filtered outliers are often caused by strong external intervention and hard to predicted from using only former viewcount trends (actually a change of

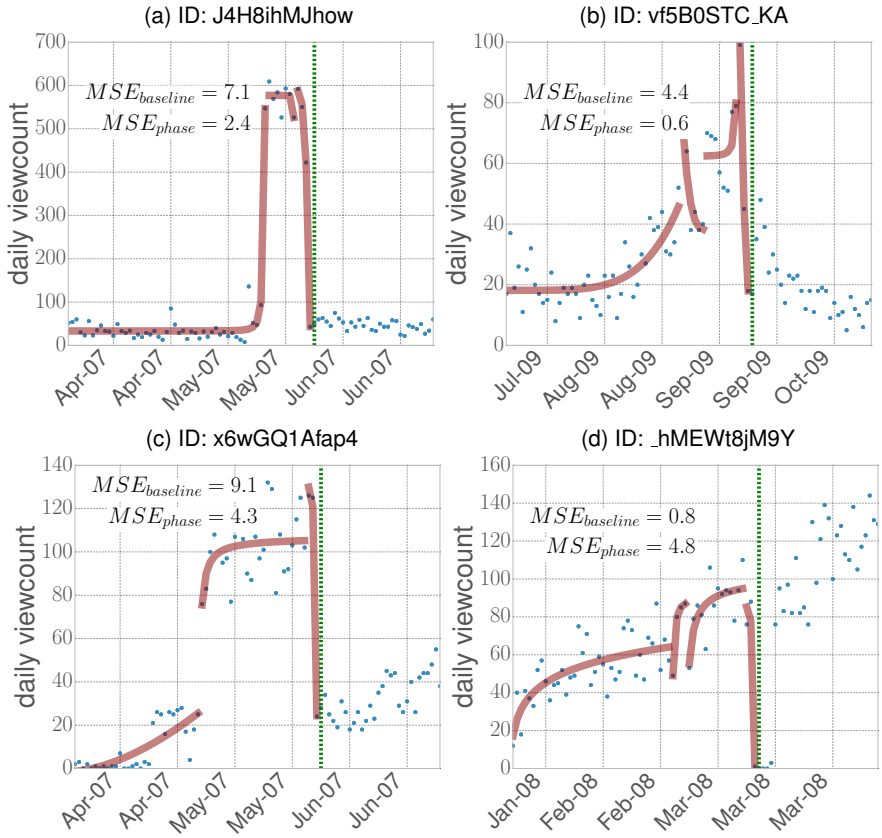


Figure 6.5: (a)(b)(c): Three examples showing that phase-informed prediction performs much better than the baseline; (d): An example where our method performs worse than the baseline ($t_p = 60, \Delta t = 30$). Blue dots: daily viewcounts; Red curves: phase segments detected; Green lines: indicating the pivot dates.

trend). We try to predict such cases with the help of Twitter information in Chapter 7.

6.2 Viewcount clustering based on phases

6.2.1 Introduction

Viewcount clustering is an important way to uncover patterns and gain knowledge from large scale datasets. The main challenge in clustering viewcounts (or more generally, any time series data) is how to define the similarity across different time series (Aggarwal and Reddy [2013]). The difficulty is that the series may be scaled or translated along both temporal and viewcount dimensions. Different definitions of simi-

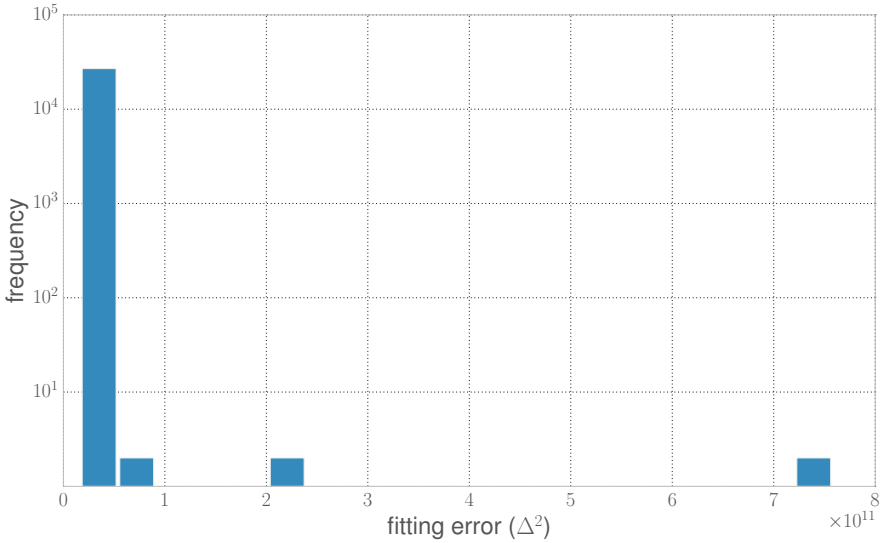


Figure 6.6: Fitting error on total data (pivot date=90 day; prediction date=120 day)

larity between time series can result in very different clusters found by algorithms. It is often hard to determine which clustering method is best since their emphases are often different and the clusters they find are, in certain sense, all meaningful. In this section, we first describe our phase-aware viewcount clustering method and apply it to our dataset. We then review and compare our method with one of the most influential time series clustering method published recently on online content popularity analysis. The result shows that, our phase-aware clustering method is more capable of capturing the prevailing pulse patterns in viewcount data and can find more meaningful clusters. It could become an important addition to the existing time series clustering toolbox.

6.2.2 Phase-sketch viewcount clustering

One application of viewcount phases is to better summarize and visualize large collections of videos. Here we present one such method using novel features constructed from phases, called *phase sketch clustering* (PSC).

For each phase $\rho = \{t^s, t^e\}$ and optimal phase parameters $\theta = a, b, c$, define the

fitted and normalized version of viewcount series as

$$\hat{x}[t] = \frac{a(t - t^s + 1)^b + c}{\max_{t=1:T} \mathbf{x}[t]}, \text{ for } t^s \leq t \leq t^e \tag{6.5}$$

We define a five-dimensional *phase sketch* feature vector over both viewcount magnitude and the timing of the phase.

$$\mathbf{u} = [\ \hat{x}[t^s],\ \hat{x}\lfloor\frac{t^s + t^e}{2}\rfloor,\ \hat{x}[t^e],\ \lambda t^s,\ \lambda t^e\ ] \tag{6.6}$$

In particular, generalized power-law curves are monotonic and are either convex or concave, and can be conveniently *sketched* based on three data points in time: the starting point $\hat{x}[t^s]$, the ending point $\hat{x}[t^s]$ (to determine the increasing or decreasing direction); and the middle point $\hat{x}\lfloor\frac{t^s+t^e}{2}\rfloor$ (being either above or below the average of the starting and end points, thereby, denoting either convexity or concavity) – hence the name, *phase sketch*. The hyper-parameter λ controls the relative importance of magnitude and timing of phases, and is chosen to be 0.2 in this work. Several examples of *phase sketches* are in the left-most column of Figure 6.7.

We consider the set of videos containing the same number of phases $\mathcal{V}_n = \{v \mid n_v = n\}$, and compose a feature vector $\mathbf{u}_v$ for each video v by concatenating features from each of its phases $\rho_{v,i}, \forall i = 1, \ldots, n_v$:

$$\mathbf{u}_v = [\mathbf{u}_{v,1}, \ldots, \mathbf{u}_{v,n}] \tag{6.7}$$

A set of *phase sketch* clusters are then obtained by running the k-means algorithm (Hastie et al. [2009]) over $\mathcal{V}_n$ with feature vectors $\mathbf{u}_v$.

Figure 6.7 contains an example outcome of a PSC, and Figure 6.8 those of k-spectral clustering (KSC) (Yang and Leskovec [2011]), on the same subset of videos $\mathcal{V}_3$ that contain 3 phases. The total number of videos in this set is 33,703, and we extracted 5 clusters with both algorithms. The first column for both PSC and KSC shows the cluster centroids, drawn as *phase sketches* and time-series, respectively.

Note that since the PSC centroids come from k-means, they can be drawn as valid sketches, but themselves need not be either convex/concave, or temporally continuous. The rest of the four columns contain the viewcount traces for four videos which are closest to their respective centroids. We can see that PSC clusters, based on an explicit phase representation, capture the bursting and timing of the viewcount behavior, whereas KSC clusters, based on a scale- and shift-invariant Euclidean distance function, only capture the long-term smooth trends. In particular, we notice that PSC clusters 2 and 3 capture the revival in popularity of videos around 200 and 500 days, respectively. Cluster 4 captures early popularity surge within the first 200 days and a decrease in attention. Cluster 5 contains cases of a minor popularity surge followed by a major one much later on in the lifecycle.

In KSC results, clusters 1 and 2 capture a long-term rising trend, cluster 3 describes a slow temporal decay, and clusters 4 and 5 contain relaxation processes over time. Note that the short-term dynamics of the videos (e.g. cluster 3, example 2; or cluster 4, example 3) are not represented in the cluster. Note that KSC works on the T-dimensional representation of the time series, whereas PSC represents the series in $5n$ dimensions, which is usually much smaller than T. In general, the burst-like temporal PSC clusters contrast with the smooth trends of the KSC clusters, and we envision that different cluster schemes can be used together.

Figure 6.9 presents a summary of PSC clusters over different popularity bins (middle row) and content categories (bottom row). We measure the log-odds-ratio (LOR) of a cluster c with respect to property z (e.g., content type) as follows:

$$LOR(c,z) = \log_{10} \frac{\#(c,z)/\#c}{\#z/|\mathcal{V}|}$$

Here $\#(c,z)/\#c$ is the fraction of videos with property z in cluster c, and $\#z/|\mathcal{V}|$ is the fraction of videos with property z across all clusters. A positive value of $LOR(c,z)$ means (a random video in) cluster c is more likely to have property z than not, and a negative value means property z is less likely to be present in cluster c than not.

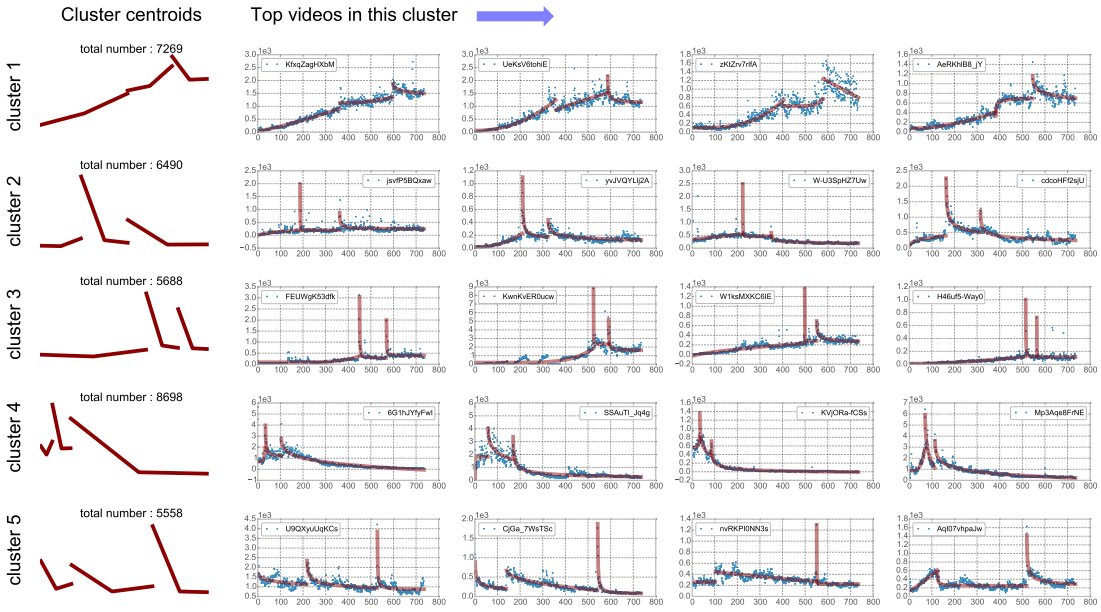


Figure 6.7: Example of *phase sketch* clusters from 33,703 videos having 3 phases. The left-most column contains the clustering centroids plotted as a phase sketch according to features in Eq (6.6). The remaining four columns are viewcounts traces (in blue) closest to the respective centroids, with overlaid phase curves (in red). x-axis: t, days since upload; y-axis: viewcount volume. Best viewed in color.

We can see from Figure 6.9 that cluster 1 is more likely to contain popular videos, clusters 2 and 3 contain videos of medium popularity, while clusters 4 and 5 are more likely to contain the least popular videos. Across the common video categories, we see the following trends: cluster 1 with a persistent rising dynamic is more likely to contain *music* videos; cluster 2 with a late take-off but multiple bursts is more likely to contain *technology* discussions; and cluster 4 with small bursts but slow decay is more likely to contain *game* videos.

Note that PSC assumes a fixed number of phases per video. While results in Figures 6.7–6.9 are for 3-phase videos, this restriction of phase numbers can be lifted by developing extensions with standard time-series techniques such as dynamic time-warp (Berndt and Clifford [1994]). The number of clusters is set to 5 in this example, however standard model selection techniques can be used to automatically determine the number of clusters.

In addition to being correlated with popularity or categories as Figure 6.9 shows,

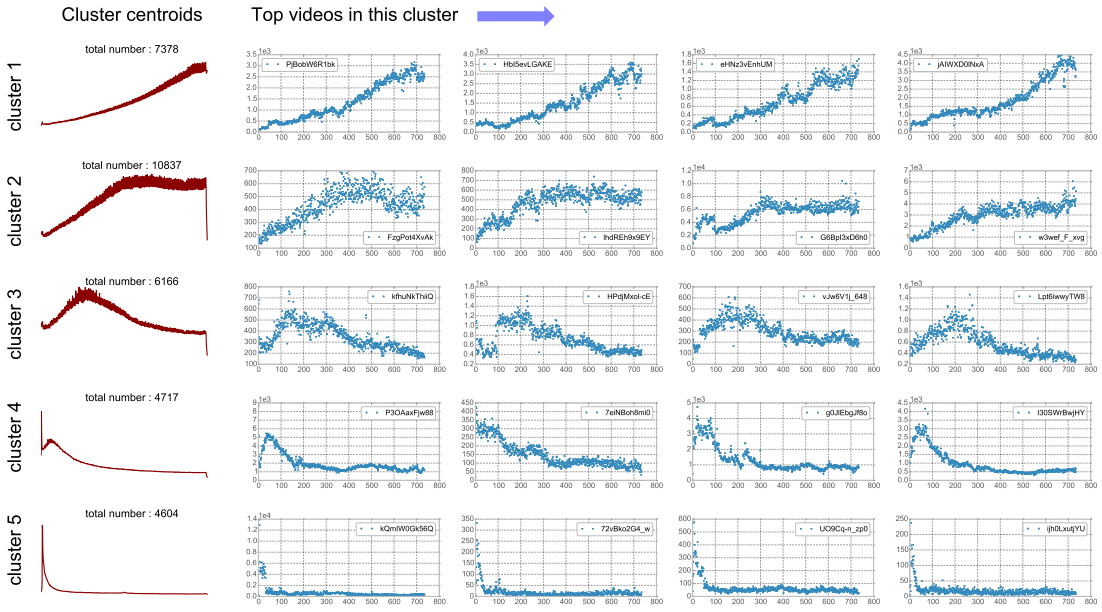


Figure 6.8: Results of KSC clustering on the same 33,703 videos as in Figure 6.7. The examples are nearest to corresponding centroids by the shift and scale invariant distance function (Yang and Leskovec [2011]). The left-most column contains the clustering centroids. The remaining four columns are viewcount traces closest to the respective centroids. x-axis: t, days since upload; y-axis: viewcount volume. Comparing Figure 6.7 and 6.8, PSC captures the volume and timing of the popularity bursts, while KSC tends to capture smooth trends.

we envision that these phase-sketch clusters can be used in a wider range of applications to reveal richer lifecycles of videos rather than a single dominant increasing or decreasing dynamic.

6.3 Summary

This chapter has proposed a phase-aware viewcount prediction method which significantly improves the performance over baseline. A phase-aware viewcount clustering method has also been proposed, a method which is more capable of capturing a common "pulse-pattern" seen in a YouTube video viewcount dataset. These two techniques demonstrate the great practical meanings of viewcount phases.

Nevertheless, as seen in Section 6.1.6, there are some videos whose viewcounts are very hard to predict. Delving into them, it has been discovered that they mostly

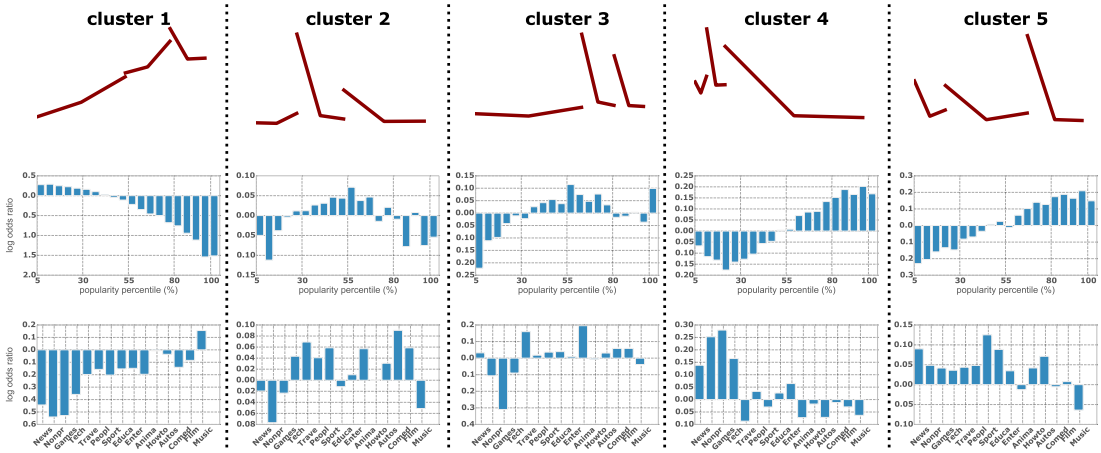


Figure 6.9: First row: *phase sketches* of 5 clusters. Second row: log-odds-ratio of videos with different popularity percentile in each cluster. Third row: log-odds-ration of videos of different category in each cluster. We can see that although they were obtained from popularity traces alone, the PSC clusters are highly informative of the popularity percentile and type of videos.

contain viewcount "sudden jumps" within the range being predicted. Also, to predict future viewcount, no matter we use phases or not, there must exist a non-trivial number of days of history viewcount (i.e., the pivot date must be sufficiently larger), which makes predicting the early popularity of a video hardly possible. In the next chapter, we will use external (Twitter) information to try to deal with these two problems.

Chapter 7

Twitter Driven Viewcount of YouTube Videos

In the Section 6.1, we saw there were videos whose future popularity is very hard to predict. The lifecycles of these videos often include sudden jumps in viewcounts, which are usually caused by external interventions. In this section, we try to utilize information from Twitter, one of the largest online social networks, to handle such cases. In addition, we also look at the scenario where the video had been recently uploaded and there was not enough history available to predict its future popularity. We show how the concurrent Twitter information can be used to help predict its popularity. And this chapter also contains interesting observations from feature importance analysis and case studies.

7.1 Introduction

This chapter focuses on studying the interaction between two of the largest social information networks – Twitter the microblog service, and YouTube the online video platform – with the goal of predicting future video viewcounts on YouTube (Figure 7.1). Understanding and predicting popularity on social media has been a very active area of new research. Social information networks such as YouTube and Twitter contain rich information about content and user profiles, as well as user actions and interactions, available in large quantities and evolving at a rapid pace. These data sources presents exciting opportunities and new challenges to understand the

underlying mechanisms behind message diffusion and user actions, as well as to build systems that can predict user actions individually and in aggregate.

Figure 7.1: Problem overview: using user activities on Twitter to predict video popularity on YouTube.

Video viewcounts on YouTube is a reliable metric for aggregate popularity, making it a good target for prediction. Recent studies have established correlations between viewcount history and number of views in the immediate future (Pinto et al. [2013]; Szabo and Huberman [2010]), but in these cases the effect of external networks are disregarded, and there are two important scenarios where predictions from viewcount history will fail. The first is a sudden viewcount change (called JUMP), as shown in Figure 7.2(Upper). Approaches that rely on viewcounts (such as Pinto et al. [2013]; Szabo and Huberman [2010]) tend to fit smooth trends well, but cannot predict jumps since jumps are often caused by external events such as referrals from an external site. The effect of such referrals is visible in the top blue trace of Figure 7.2(Upper) – there are more than 250 tweets about this video right around the time of the viewcount jump. The second scenario is predicting the popularity of newly uploaded videos (called EARLY) where have no available viewcount history. This challenging task can also be tackled by using other information from external networks. An example is shown in Figure 7.2(lower), the volume of tweets concerning this video started on August 21st, 2009, and a significant viewcount increase started a few days after. JUMP identifies most popular videos that became popular later after upload, and EARLY identifies the videos that became very popular

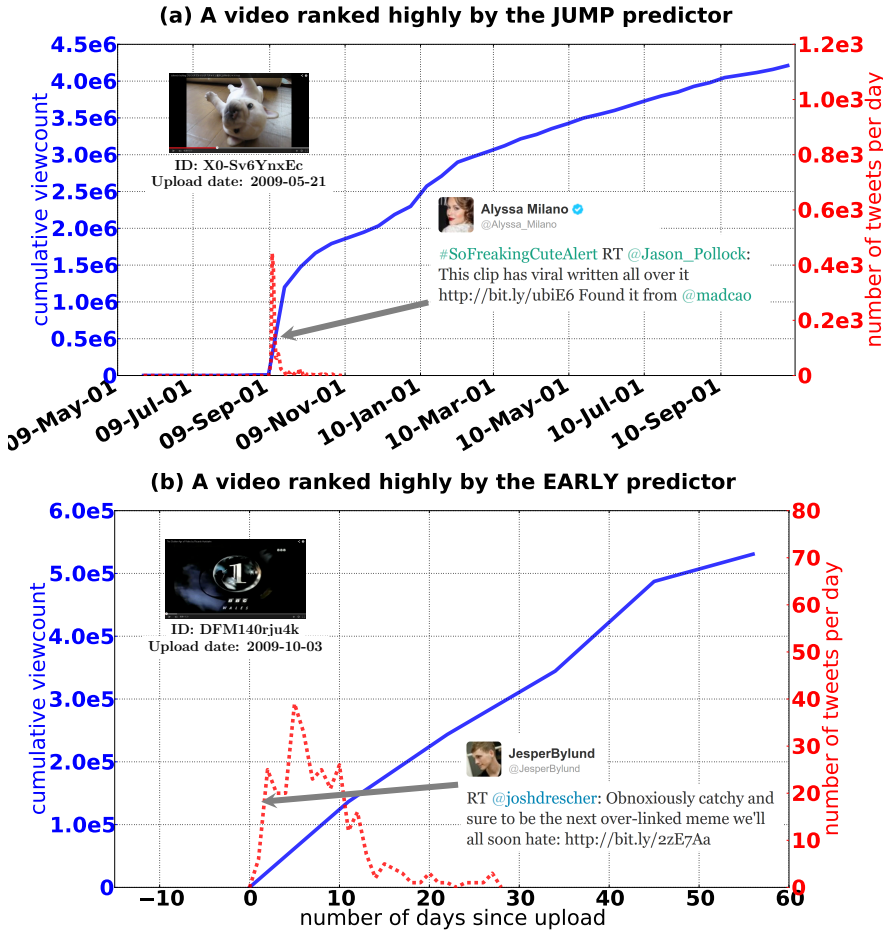


Figure 7.2: Examples of top predictions for JUMP and EARLY. (a) A video having less than 9000 views in its first 3 months, and then gaining 1.2 million views within 15 days (date format of x-axis: yy-mmm-dd). The insert shows a tweet linking to this video by celebrity user Alyssa Milano. (b) A video with a few dozen Twitter mentions and nearly 2×10^5 views in its first 15 days. Note that the video popularity continues to rise even after the tweet volume has tapered off, illustrating the prediction power of early tweets.

in the early stage – both are important scenarios for a range of applications including content recommendation and advertising. Note that they both cannot be handled by existing approaches as they are essentially outliers or suffer from lack of information. Moreover, most studies on content popularity have focused on information within a single network (Borghol et al. [2012]; Hoang and Lim [2012]; Yang et al. [2012]; Bakshy et al. [2011]), while interpretation of these two important scenarios benefits from examining effects across multiple networks.

There are three key contributions in this work.

- We demonstrate that sudden viewcount changes and the popularity of new videos on YouTube can be predicted. We define two predictive tasks JUMP and EARLY, by partitioning the relative and absolute viewcount in to a few distinct classes. We use recent viewcount history (when available) along with four different types of features from Twitter – information about tweets, about the Twitter "following" graph, and Twitter user interaction behaviors (both active and passive).

- We demonstrate that Twitter features have a measurable, and significant impact on predicting Youtube viewcount. We report prediction performance using a logistic regression classifier on a Youtube videos tweeted in a 3-month period August-October 2009. The prediction performance of JUMP is 0.10 better in accuracy than random, and EARLY is 0.25 better than random.

- Twitter user network and activity features are generally more predictive than tweets that mention the video. That is, all Twitter features out-perform the viewcount feature for JUMP, and combining all features leads to significant improvements in prediction performance.

7.2 Related work

This work is related to three active research areas on social media and online social networks. The first area is descriptive and predictive analysis of YouTube content and its popularity. Cha et al. [2007], Chatzopoulou et al. [2010] and Cheng et al. [2008] performed large-scale measurement and descriptive analysis on YouTube datasets and revealed basic statistics of YouTube video views. Crane and Sornette [2008] analyzed the dynamics of viewcount accumulation and proposed different mechanisms by which a video spreads. Brodersen et al. [2012] found a strong geographic effect on the popularity of YouTube videos. Szabo and Huberman [2010] found a strong

correlation between a video's historic and future viewcounts and used a linear model for prediction. Pinto et al. [2013] improved this model by using detailed viewcount traces as features. Borghol et al. [2012] used near-duplicate videos as a distinctive marker to isolate content-agnostic factors that affect a video's popularity. These research effort have provided valuable insight, yet they only consider factors within YouTube.

The second research area is in measuring Twitter users' influence and analyzing tweet diffusion. Two concurrent measurement studies (Cha et al. [2010]; Kwak et al. [2010]) each analysed a very large sample of the Twitter archive, and both found the number of followers reveals little about retweeting influence. Hoang and Lim [2012] proposed methods to model the diffusion of viral topics on Twitter by considering the mutual dependency of users and items. Yang et al. [2012] performed a comprehensive study to predict hashtag adoption on Twitter, based on the topical and community roles that hashtags play. Bakshy et al. [2011] tracked diffusion of URLs on Twitter and found that the largest cascades of events tend to be initiated by users with a large number of followers, but the most cost-effective diffusion strategy is expected to be one where a campaign targets individuals who exert only average or even less-than-average influence. Work in this area gave us good starting points for analyzing tweets and Twitter user behavior. We note that these studies of influence are focused on events within Twitter, and not on the external effects that Twitter may have on the outside world, such as the sale of a product or the viewcount of a video.

The third area is on the interactions between different online social networks. Bhagat et al. [2007] did one of the first descriptive studies between blog, the web, and instant message networks. Cha et al. [2012] performed a measurement study on the spread of media contents (including YouTube videos) through blogs and found different categories of videos have different propagation patterns. Myers et al. [2012] proposed an information diffusion model of social networks considering both internal and external effects and showed that 29% of Twitter information diffusion can be attributed to external events. Wang et al. [2012a,b] analysed the propagation of

online videos on microblogs in China and used location predictions to improve video caching. To the best of my knowledge, no correlation have been established between Twitter activities and sudden changes in YouTube video popularity. We intend to taking the first steps towards filling this gap.

7.3 Processing dataset

We used three data sources: the YouTube video history providing total viewcounts over time, a subset of tweets over a 3-month period, and the Twitter user graph from the same period.

We obtained **YouTube viewcount history** from a video's webpage, when it is made available by the video owner. This history contains the number of views a video has received since its upload, in 100 evenly spaced time intervals, with daily viewcount obtained by temporal interpolation[1]. We used a collection of 467-million tweets of from 2009 (Yang and Leskovec [2011]); this sample is estimated to contain about 20-30% of all posts published on Twitter during the 6-month period, and was authored by about 20 million users. Each tweet is represented by three fields: *author*, e.g., http://twitter.com/annieng; *timestamp*, e.g., 2009-06-07 02:07:42; and *tweet content*, e.g. "in LA now".

We used a snapshot of the **Twitter user graph** from 2009 (Kwak et al. [2010]), with each user as a node, and each following relationship as a directed edge from a user to one of his/her followers.

We extracted URLs from all tweets and resolved shortened URLs, retaining references to YouTube videos. We found that 1,624,274 Twitter users tweeted YouTube video links at least once and that there were 2,350,881 unique videos, of which 1,549,532 (65.9%) are still online (as of Oct 2013). Within this subset, 1,067,895 (68.9%), had their viewcount history publicly available. We call the subset of tweets containing videos with available history *video tweets*.

[1]The temporal granularity of viewcount history is about 12 days (with 1,100+ days between August 1, 2009 and data collection in Oct 2012) and varies depending on the video upload date.

We matched the tweets and the Twitter user graph dataset in order to extract the user graph information of the observed video tweets. About 80% of the users could be identified by matching the username directly, although for 20% of the users we did not find a match. A tweet was dropped from the collection if its author could not be identified.

We processed tweets to extract tags and user interactions. We relied on text processing for this since our historical tweets collection does not contain the full Twitter API feed (where many interactions are already encoded). We extract **hashtags** and **mentions** by finding words prefixed with # and @ symbols. We also extracted non-broadcasting tweets (**nbcTweet**) – when a tweet starts by mentioning a user, it was treated as a targeted interaction between the author and the user being mentioned, so that the followers of the author will not see this tweet in their timeline. We extracted variations of retweets (**RT**). Symbols for retweeting have evolved since the early days of Twitter (Kooti et al. [2012]), and there are still a diverse set of symbols in use in the 2009 data. We extracted 10 major variants, i.e., *RT, R/T, via, HT, H/T, OH, retweet and ret*, plus two variants of the "recycle" symbol (♲).

7.4 Methodology overview

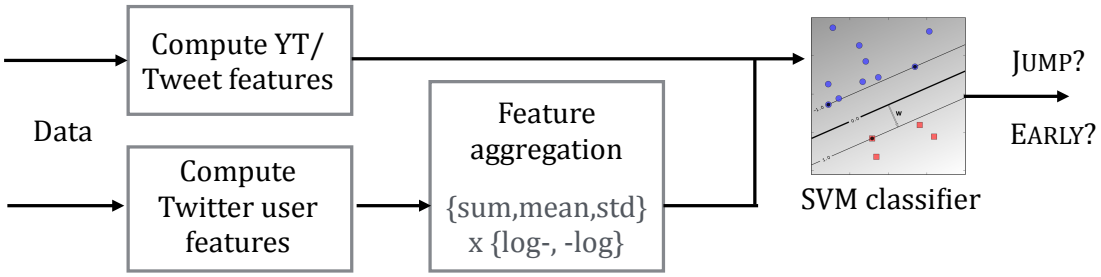


Figure 7.3: Overview of our method for predicting viewcounts using Twitter information.

The overview of our methodology is shown in Figure 7.3. After collecting the data, we computed the features from both YouTube and Twitter datasets. Since there

are usually various numbers of Twitter users who have tweeted about the same video, we use 6 statistics to summarize each Twitter user features. Then we used all these features, plus support-vector-machine classifiers, to successfully predict viewcounts in two tasks, namely JUMP and EARLY , noting that such predictions are usually very hard to make without external information.

In the next section, I will describe the features in details. Then I will formulate the two predictions tasks. Finally, I will discuss the results of the experiments, which will include evaluation of predictions, analysis of the importance of each feature and case study.

7.5 Features from YouTube and Twitter

We begin describing the features by defining units of data over time. We take a sliding time window of length τ as a unit for feature extraction and viewcount prediction. We define time index $t \in \{0, 1, \ldots, T\}$ with increments of τ. With slight overload of notation, index 0 may represent the real-world interval $[0, \tau)$, or time point $t = 0$, the interpretation should be clear from the context. For a YouTube video v, denote its total viewcount (i.e. number of views received since upload) on time t as $c_v(t)$, and the viewcount increment between t and $t + 1$ as $\Delta c_v(t)$. We use $U_v(t)$ to denote the set of Twitter users who tweeted video v in time interval t. In this work, the prediction targets are videos tweeted between August and November 2009 with $T = 93$ *days*, and $\tau = 15$ *days* due to viewcount data granularity. Tweets published before August are used to compute features.

We extract one set of YouTube features and four sets of Twitter features for prediction. There are two general types of aggregated Twitter features we investigate: those directly involving TWEETS on the video and those involving the users who have tweeted on a video. Among the latter type, we further make distinctions among ACTIVE, PASSIVE, and social GRAPH features of those users.

7.5.1 YouTube features

YT-views is the number of views a video v receives in two time intervals before time t on which we are making a prediction, i.e. $[\Delta c_v(t-2), \Delta c_v(t-1)]$. Historical viewcount is shown to highly correlate with future viewcounts (Szabo and Huberman [2010]), and using more than one historical interval is shown to further improve prediction (Pinto et al. [2013]). We chose two intervals via cross-validation. This feature is comparable to those used in prior work (Szabo and Huberman [2010]; Pinto et al. [2013]), and used as the baseline for predicting JUMP .

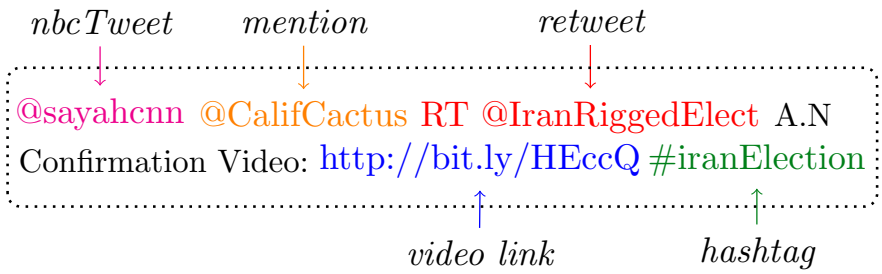


Figure 7.4: Illustration of the special fields of a tweet.

7.5.2 Tweet features

TWEET includes five counting metrics that describe the properties of video tweets about video v in interval $[t_0, t]$ as follows (t_0 is the first date of which we have tweet data, i.e., June 1st 2009; And t is the date on which prediction is made). $T.tweet(v,t)$ is the number of video tweets; $T.hashtag(v,t)$ is the number of times a hashtag is used; $T.mention(v,t)$ and $T.nbcTweet(v,t)$ are the numbers of broadcasting and non-broadcasting mentions, respectively; and $T.RT(v,t)$ is the number of retweets for each of the 10 variants. Intuitively, videos are likely to obtain more views when they are tweeted or are part of twitter interactions (via hashtags, mentions, or retweets).

7.5.3 Twitter user features

We have grouped the Twitter user features into three main categories, namely Graph Active and Passive, which are, explained in the following.

7.5.3.1 Graph

In this section, we describe Twitter user features related to their following graph. We consider the Twitter user network as a directed graph in which each node represents one unique Twitter user. So there is one edge from user u_i to u_j if and only if u_j follows u_i. By this convention, the directions of edges conform to those of the information flow (u_i's tweets are received by u_j). We will use centrality scores from the Twitter user network as features to describe the influence of Twitter users. First, let us briefly review the classic graph centrality scores and then describe our graph features (denoted as Graph) in details.

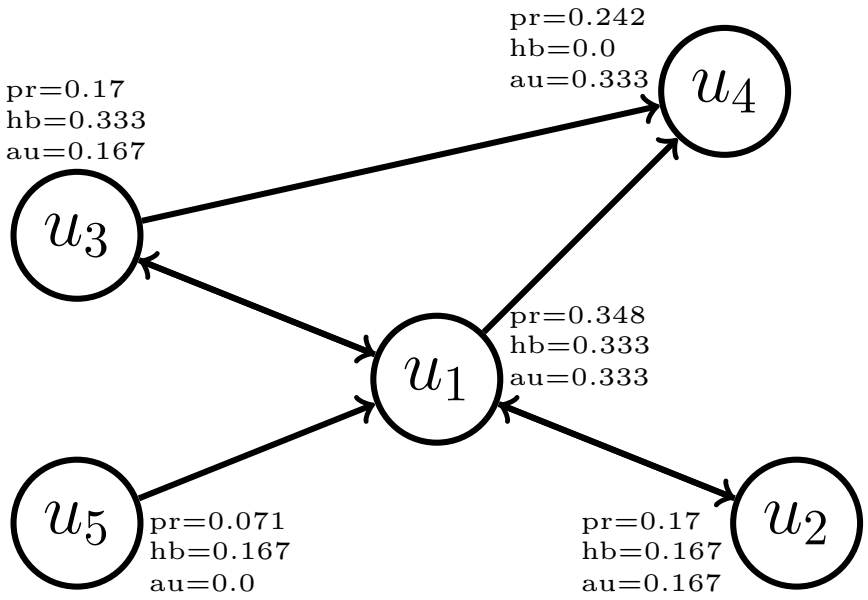


Figure 7.5: An example graph and their nodes' pagerank (pr), hub (hb) and authority (au) scores.

Graph centrality scores

Outdegree The *outdegree* of a node n_i is the number of edges having it as its head. It represents how many followers a user has on Twitter and measures how many users will receive his/her tweet. This is the most basic measure of a user's influence.

Page-rank One of the main flaws of *outdegree* in measuring users' influence in a network is that, it treats all the followers equally. For example, assume that u_i and u_j both have 10 followers. But the followers of u_i are all celebrities with millions of followers whereas the followers of u_j are all grassroots people with very few followers. The importance of user u_i and u_j in terms of information propagation in the network are apparently different, but it is not reflected by their outdegrees, which are both 10.

Page-rank improves this in a way that nodes with higher page-rank scores contribute more to the page-rank scores of the nodes they follow. Denoting the adjacent matrix of a graph as $\mathbf{A}$ and $\mathbf{D} = \mathbf{diag}\{o_1, o_2, ..., o_n\}$ where o_i is the outdegree of node i ($i = 1, ..., n$), the page-rank score $\mathbf{x} = (x_1, ...x_n)^T$ of all the nodes should satisfy the eqaution (Newman [2010]),

$$\mathbf{x} = \alpha \mathbf{A}\mathbf{D}^{-1}\mathbf{x} + \mathbf{fi} \tag{7.1}$$

where α is called the damping parameter and is usually chosen as 0.85 (Page et al. [1999]). The solution of Equation 7.1 is

$$\mathbf{x} = \mathbf{D}(\mathbf{D} - \alpha\mathbf{A})^{-1}\mathbf{fi} \tag{7.2}$$

So, if we choose $\mathbf{fi}$[2] as an identical vector, it will not affect the final score ranks. Page rank scores of large social network can be solved through iterative algorithms

[2]In the random walk explanation of page rank scores, **fi** is the probability of a surfer on the graph making a *teleport* operation (Manning et al. [2008b]), and α is the probability the surfer will move his/her position from the current node by following edges (rather than staying still or making a teleport operation).

(Berkhin [2005]). Here we use the SNAP package[3] to compute it.

Hub-Authority score Hub-Authority score was first proposed by Kleinberg [1999]. The hypothesis is that a node's importance may also depend on the nodes that direct to it. The (HITS) algorithm then assigns two scores, namely *hub* and *authority* scores to each node. High authority nodes are those containing useful information and nodes with high hub scores means they can tell us where the best authorities are to be found (Newman [2010]). In the case of Twitter, authority scores are the measurement of value of a user, whereas hub scores measure his/her taste. Denoting hub and authority scores by $\mathbf{x}$ and $\mathbf{y}$ respectively, their relationship is as follows,

$$\mathbf{x} = \alpha \mathbf{A}\mathbf{y}, \ \mathbf{y} = \beta \mathbf{A}^T \mathbf{x} \tag{7.3}$$

where α and β are positive constants and $\mathbf{A}$ is the adjacent matrix. Then one can easily see that $\mathbf{x}$ and $\mathbf{y}$ are actually the eigenvectors of $\mathbf{A}\mathbf{A}^\mathbf{T}$ and $\mathbf{A}^\mathbf{T}\mathbf{A}$ of the same eigenvalue $(\alpha\beta)^{-1}$. According to Newman [2010], hub and authority scores are not as widely used as page-rank but similar algorithms are used by *Teoma* and *Ask.com*.

Other centrality scores *Closeness* and *betweenness* are also common centrality scores used to measure nodes' importance. But since they are not computable for very large networks, we will not use them as features.

Twitter user graph features Let us now explain the concrete Twitter user network feature set, namely GRAPH. It consists of three features computed from the Twitter user graph. As written above, for a Twitter user u, $G.outdegree(u)$ is the number of followers he/she has. $G.pagerank(u)$ contains the pagerank score of a user, a robust measure of a user's influence in adopting hashtag (Yang et al. [2012]). $G.hubauthority(u)$ contains a pair of hub and authority scores (Kleinberg [1999]). A Twitter user has a high hub score if her followees have high authority scores; she

[3] https://github.com/snap-stanford/snap

has a high authority scores if her followers have high hub scores.

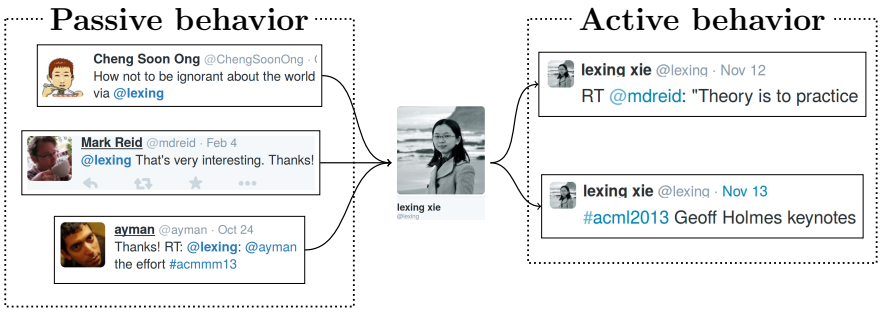


Figure 7.6: Illustration of a user's *active* and *passive* behavious (@lexing as example).

7.5.3.2 Active behavior

Active Behavior consists of five types of behavior features for Twitter user u up to time $t-1$. They are denoted as $A.tweet(u)$, $A.hashtag(u)$, $A.mention(u)$, $A.nbcTweet(u)$ and $A.RT(u)$ to respectively capture the a user's tweet volume, use of hashtags, sending of broadcasting and non-broadcasting mentions and retweet behavior. Moreover, each feature type is represented with a number of metric variants. $A.tweet(u)$ includes four variants: the total and per-day average of all and unique tweets. Each interaction feature (hashtag, nbcTweet, mention, RT) includes four variants: the total number of interactions, its average per day; the number of unique user-to-user interactions, and its average per day.

7.5.3.3 Passive Behavior

Passive Behavior consists of three behavior features that Twitter user u receives from other users up to time $t-1$. Denoted as $P.nbcTweet(u)$, $P.mention(u)$ and $P.RT(u)$, they represent interactions where user u is mentioned in broadcasting tweets, non-broadcasting tweets, and retweeted, respectively. Each of these features of Twitter user interactions has the same metric variants as those for Active features.

Both active and passive features have been recognized as capturing user influence within Twitter (Yang et al. [2012]); here we use them to infer YouTube popularity.

Table 7.1: YouTube and Twitter feature summary (Sec 7.5)

Feature group	Feature name	# of dimensions
YT-Views	viewcount	2
Tweet	T.tweet	1
	T.hashtag	1
	T.mention	1
	T.nbcTweet	1
	T.RT	10
Graph	G.outdegree	6
	G.pagerank	6
	G.hubauthority	12
Active	A.tweet	24
	A.hashtag	24
	A.mention	24
	A.nbcTweet	24
	A.RT	240
Passive	P.mention	24
	P.nbcTweet	24
	P.RT	240

Note that features from tweets are computed from dataset inception, i.e. 2009-05-31 (Yang and Leskovec [2011]). Furthermore, we aggregate the Graph, Active and Passive features from the set of users tweeting about the same video into six summary statistics. These statistics incorporate three kinds of aggregation (sum, mean and standard deviation (std)); over two scaling variants (log-aggregate or aggregate-log). The method to compute them can be found in Table 7.2. This accounts for the variable number of users tweeting each video, and is able to be generalized across users. An overview of all features is in Table 7.1. Note that the feature dimensionality includes summary statistics for all user features and all metric variants, e.g., A.RT (and P.RT) has 10 RT literals $\times$ 4 metric variants $\times$ 6 summary statistics, totaling 240 dimensions.

7.6 Two prediction tasks

In this section, I describe our two video viewcount prediction tasks namely Jump and Early.

Jump captures cases when a video gains a large number of views in a relatively

Table 7.2: Six summary statistics for user features. u: a Twitter user; U: a set of Twitter users; $f(u)$: a user feature.

Name	Description
sum-log	$\sum_{u \in U} \log(f(u)+1)$
log-sum	$\log(\sum_{u \in U}(f(u))+1)$
mean-log	$\frac{1}{\lvert U \rvert}(\sum_{u \in U} \log(f(u)+1))$
log-mean	$\log((\frac{1}{\lvert U \rvert}(\sum_{u \in U} f(u))+1))$
std-log	$std(\{\log(f(u)+1)\}_{u \in U})$
log-std	$\log(std(\{f(u)\}_{u \in U})+1)$

short period of time. For video v, denote the total viewcount gained between time 0 and T as $\Delta c_v(0,T)$; we compute the *normalized* gain during interval t as $r_v(t) = \Delta c_v(t)/\Delta c_v(0,T)$. For a video v, a *jump* is deemed to have occurred in time t if $\Delta c_v(t)$ has more than 50 views; $\Delta c_v(t-1)$ is not more than $\Delta c_v(0,T)/T$, the average gain over interval $[0,T)$; and $r_v(t) \geq \alpha$ with predefined threshold α. Defining jumps using such normalized increments allows us to compare videos that undergo popularity changes at very different levels, e.g., from hundreds to millions of views.

Early captures cases when a video receives a significant number of views just after being uploaded. We take the most popular videos as prediction targets, i.e., those having the most viewcount in their first $\hat{\tau}$ days, denoted as $\Delta c_v(0,\hat{\tau}) > \beta$, with a pre-defined threshold β. Prior approaches that rely on historical viewcount (Pinto et al. [2013]; Szabo and Huberman [2010]) cannot be used to analyze such a phenomenon.

Binary classifiers are trained with linear support vector machines for each task. We use $\alpha = 0.5$, and $\beta = 10^4$, high thresholds that yield popular videos which are likely to be mentioned in tweets. The empirical YouTube viewcount distributions are long-tailed, and do not show a clear separation around these (or any other) values. To this end, thresholds separating the top few percent of videos are equally valid conceptually. Figures 7.2 (a) and (b) contain examples of videos in the respective Jump and Early classes, which are ranked highly by our algorithm.

7.7 Experiments

7.7.1 Prediction result

We evaluate JUMP and EARLY prediction with the following settings. For JUMP, each time interval t with at least one tweet about video v becomes an instance, and there are 6,156 positive JUMP instances with a random guess prior of 1.2%. The five feature groups are specified in Section 7.5, and ALL is a result of concatenating features from all available groups. For EARLY, each video (in its first $\hat{\tau}$ days) becomes an instance. Results are reported on 29,998 videos, out of which about 5.3%, or 1,591 are positive examples. We report average precision (AP) (Manning et al. [2008a]) and Precision@100, with the average and the 95% confidence interval over 5-fold cross-validation with stratified sampling (preserving the random guess probability).

Table 7.3 summarizes the performance of different features for JUMP . We can see that among the four types of Twitter features, each improves upon results using viewcount history only. The best predictor doubles the AP and nearly quadruples the Precision@100 vs. viewcounts, and with Precision@100 at 0.46, almost half of the top-ranked videos actually contain a JUMP . In addition, differentiating users and taking into account user history (with GRAPH, ACTIVE and PASSIVE) make the predictor perform significantly better than only using viewcounts or tweet properties as features.

Table 7.4 summarizes the prediction performance of EARLY , with the same feature groups as JUMP except that YT-VIEWS is unavailable for newly uploaded videos. The prediction is done for the first 15 days, and then $\hat{\tau} = 30, 60$ and 90 days. Longer term predictions are done with ACTIVE features, because (1) it is the best-performing feature group – only 0.05 away from ALL in prec@100; (2) these features only need user history for the video tweets, and do not need Twitter GRAPH or PASSIVE interactions, which are expensive to obtain. It is encouraging to see that the top 5% most popular videos can be predicted with an AP of more than 0.40, and there are 70+ correct entries in the top 100. Moreover, this accuracy is maintained from 15 to 90

Table 7.3: Performance for JUMP prediction. See Sec 4.

Features	Avg Prec	Prec@100
Random	0.012	0.012
YT-VIEWS	0.056 ± 0.006	0.125 ± 0.028
YT-VIEWS+TWEET	0.058 ± 0.002	0.204 ± 0.041
YT-VIEWS+GRAPH	0.097 ± 0.007	0.406 ± 0.023
YT-VIEWS+ACTIVE	0.105 ± 0.003	0.432 ± 0.057
YT-VIEWS+PASSIVE	0.104 ± 0.005	0.444 ± 0.044
ALL	0.113 ± 0.008	0.460 ± 0.053

Table 7.4: Performance for EARLY prediction. See Sec 4.

$\hat{\tau}$	Feature	Avg Prec	Prec@100
all	Random	0.053	0.053
15-d	TWEET	0.248 ± 0.142	0.450 ± 0.229
15-d	GRAPH	0.382 ± 0.030	0.646 ± 0.044
15-d	ACTIVE	0.441 ± 0.027	0.702 ± 0.058
15-d	PASSIVE	0.375 ± 0.055	0.656 ± 0.088
15-d	ALL	0.463 ± 0.029	0.750 ± 0.045
30-d	ACTIVE	0.421 ± 0.023	0.686 ± 0.060
60-d	ACTIVE	0.435 ± 0.024	0.722 ± 0.018
90-d	ACTIVE	0.424 ± 0.026	0.720 ± 0.043

days.

7.7.2 Feature importance analysis

We perform an analysis of the informativeness of individual feature dimensions described in Sec 7.5. We compute the mutual information between the target class $Y \in 0, 1$ and each feature X, as

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} . \tag{7.4}$$

The larger the mutual information, the more informative a feature is towards predicting the target. Fig 7.7 contains box plots of such mutual information on user features (GRAPH, ACTIVE and PASSIVE) grouped by the three feature aggregation methods: std, sum, and mean. We can see that the majority of most informative features (e.g. top 1/6 above the median for std) are std-features, with sum features moderately informative and mean features the least informative. A high standard deviation for a

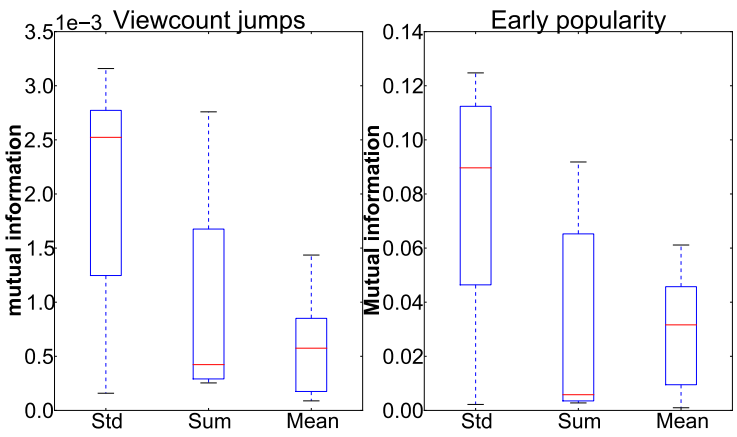


Figure 7.7: Box plots of mutual information grouped by feature aggregates. The most informative features are generated by std aggregation for both JUMP and EARLY .

feature implies that there is broad interest across a spectrum of users with a range of activity and interaction levels. This concurs with a recent observation on hyperlinks in Twitter (Bakshy et al. [2011]) – that having a diverse set of users (std) mentioning an item is helpful for improving its popularity.

7.7.3 Case study

Figure 7.2 gives examples of the videos with high rank score in our EARLY and JUMP predictors. We can see tweeting behaviors not only strongly correlate with viewcount increase but also have predictive power.

An interesting aspect of the video (b) in Figure 7.2 is that its author actively promoted his/her video through Twitter after its uploading (see Figure 7.8). This is a typical example of successful internet promotion.

7.8 Summary

In this chapter, we demonstrated that user and content information from Twitter can be effectively used to predict content popularity on YouTube, as shown by two challenging tasks – predicting viewcount JUMP and EARLY popularity – both with

Figure 7.8: The author of video DFM140rju4k recommended his video to five celebrities on Twitter after upload. The video received 2×10^5 views in the first two weeks.

significant and quantifiable performance gains. These results are encouraging in that they show viewcounts are predictable from one external source alone, without taking into account the influence from many other social and traditional media sources (Reddit, Tumblr, Pinterest, Facebook,...). Furthermore, the results show the predictive power of different features and aggregation methods and reveal that having a diverse range of users and associated tweeting activities is more informative than the total or average volume of activity of these users and also more informative than other features – including those based on social network derived measures of influence. This work raises many interesting avenues for future work, such as leveraging diffusion patterns on Twitter to further improve popularity prediction and quantifying the roles of influencers vs. grassroots users.

Chapter 8

Conclusion and Future Work

In this chapter, I first summarize the main contributions of this thesis from three perspectives, namely as a measurement study, as a means for predicting viewcounts and as an advance in viewcount clustering. Then I discuss the possible future research directions and extensions of this thesis.

8.1 Measurement study

This thesis contains a large measurement study on YouTube video lifecycles based on a high-quality dataset. The video IDs from a large Twitter repository were extracted by filtering the URLs from tweets. These videos received more than a minimal amount of attention, assuming people who tweeted the video probably watched it. Then by using powerful data crawlers developed for this thesis, the full daily viewcount history of videos while they were still online and publicly available were collected. Investigating these data, we observed the followings: Viewcounts are distributed exponentially over their relative popularity rankings. Popularity percentiles are a good way of representing a video's popularity. Some old videos like "Music" and "Comedy" are much more likely to be discussed by Twitter users than old "News" and "Games" videos; Most videos obtain their views right after upload; The variance of viewcount series is not homogeneous over time; The viewcount data of some videos (e.g., related to weather) clearly have weekly or yearly periodicity. There are strong correlations between increases in a video's viewcount and the concurrent tweets about it. In all, these results expand on previous measurement studies on

YouTube video popularity.

More importantly, based on previous research, we have proposed a new way of representing popularity phases in order to handle the temporal complexity of viewcount dynamics. We have also developed an efficient phase detection algorithm which simultaneously computes phase parameters and boundaries. It can also automatically determine the number of phases each video has gone through. We applied this algorithm to a large daily viewcount history dataset. By examining the relationship between phases and their two main co-variates (video popularity and user-assigned category), a number of novel observations have emerged, such as that popular videos are more likely to have more phases. Thus, in our dataset, nearly 3/4 of the videos in top 5% of popularity have 3 or more phases, whereas only 1/5 of the least popular videos (bottom 5%) do. Videos of some categories, e.g., "News", are more likely to have a long power-law decreasing phase. For example, more than 60% news videos are dominated by one long power-law decay, whereas only 20% of music videos do. We also observed phase profiles which can not be explained by previous models of collective attentions. Using phase properties, we observed, for the first time in detail, how viral videos became viral. In general, our research clearly demonstrates that multi-phase representation has the potential to be a useful tool for analyzing the way in which the popularity of online media evolves.

8.2 Viewcount prediction

In this thesis, we have proposed two new viewcount prediction methods, namely phase-aware viewcount prediction and predicting YouTube video viewcounts with Twitter feeds. The first method was based on the observation, repeated seen in many experimental analysis: the performance of baseline method in predicting viewcounts correlated strongly with a video's phase properties. Then I proposed a new phase-aware viewcount prediction algorithm, one which groups videos by the number of phases they have and the shape of their last phases. Subsequently, a baseline model

is trained for each of the groups. Experiments show that such a method significantly improves the performance of the baseline method over all subgroups.

In viewcount prediction experiments, it was found that some videos were very hard to predict. These videos usually contained sudden jumps in viewcount which were often caused by external events like news or a celebrity' recommendation on Twitter or Facebook. In such cases, we tried to utilize information from Twitter to improve prediction. Another difficult scenario is where a video has been newly uploaded and there is not enough viewcount history available to make a prediction. We designed an method which used 5 types of tweet features and 11 types of Twitter user features to deal with the situation. Using these features with support-vector-machine classifiers, we have successfully predicted viewcounts in both of these cases. Our results are encouraging in that they demonstrate that viewcount sudden jumps and a video's early popularity can be predicted using only a single external source. Furthermore, by comparing the predictive power of different features, we discovered that the best predictions come from looking at Twitter users' active behavior features. This may have practical significance in that such Twitter features are easier to compute than other features like a Twitter user's pagerank score. Our results also show that having the associated Twitter activity of a diverse range of users – both grassroot members and celebrities – is more predictive than using the total/average volume of activity of these users. Overall, this research can claim to be not only a pioneering study of how external information can be used to predict YouTube video popularity, but also a landmark analysis of how Twitter has a major effect on the popularity of things – notably videos – which become the subject of tweets.

8.3 Viewcount clustering

Beside being important features in predicting viewcounts, viewcount phases are also effective tools to summarize and visualize large collection of videos. This was the inspiration for proposing a new viewcount clustering method called phase sketch

clustering. In this method, each viewcount series is represented by the parameters and boundaries its phases. Experiments have shown that the clusters found by phase sketch clustering correlate with a video's popularity and category. In comparison with previous method, we found it is much better in capturing the prevalent pulse-pattern in viewcount datasets and reveals that videos have much richer lifecycles rather than just predominant increasing/decreasing trends. In addition, by turning the hyper-parameter, users can change the relative importance of phase shape similarity and phase position similarity. Lastly, this method also provides a neat sketch to represent each cluster. In summary, we are strongly convinced that phase sketch clustering can be an important tool for analyzing the popularity dynamics of online media and will find a place in a wide range of applications.

8.4 Future work

Concluding this thesis, there are 5 aspects that deserve further study,

1. **Investigate and compare the popularity phase profiles of more kinds of YouTube video datasets**. Sampling bias is always a challenging problem for empirical research. YouTube is so huge and diverse that it is difficult to fully analyze. An outstanding question in understanding people's collective behavior is: do people behave differently depending on the type of online contents? In our research, due to limited time and resources, we only considered videos discussed by Twitter users (which we thought were the most meaningful set) and then evaluated the viewcount dynamics in terms of video popularity and video category. However, previous research has also considered, for example, deleted videos, top videos in YouTube webpage and random sampled video IDs. Are there significant differences among those different types of videos? And do videos related to different hashtags/events/topics have different phase properties? Analyses like these might help us understand collective attention in more dimensions.

2. **Investigate the popularity phases of other types of online media.** Beside video viewcount, online items like hashtags, short textual phrases in tweets, popularity of blog posts, and so on, can all be seen as aggregate measures of collective attention. So what are the popularity phase characteristics of these data? Is there any difference among them and what factors cause the difference? These analyses will give us a broader view of the characteristics of people's collective attention to online items.

3. **Explore the evolution of online users' behavior by comparing the popularity dynamics of old and current videos**. The videos in our dataset were mostly around 5 years old. Over these years, the patterns of link spam and the like have changed significantly, which may affect the conclusions in Chapter 7. Also many apps now auto-tweet content for users, which increases tweets and decreases the effort involved per user, which again may change how one approaches Chapter 7 or at least introduce the need for additional features and/or tweet filtering. Lastly, one of the most prominent changes from these 5 years may be that more and more Internet traffic now comes from mobile devices rather than desktops. So is there any difference between the viewcount dynamics of newly uploaded videos and the "old" videos? Such research could reveal the evolution of online user behavior of not only YouTube but also Twitter.

4. **Build new mathematical models to explain the diverse viewcount dynamics**. In this thesis, it has been shown that viewcount dynamics differ depending on the type of videos. A natural follow-up question is: what kind of factors/mechanism causes such differences? To answer this question, more empirical research is needed, but we also need more mathematical models to explain the data patterns we have found. For example, our research reveals some novel observations that can not be explained by existing models, e.g. 1) Multiple peaks. 2) A non-trivial number of concave shapes, particularly, in Music or popular videos. These intriguing anomalies call for further study.

5. **Seasonality**. In our research, we found the daily viewcount of many videos had strong weekly or seasonal periodicity. But what kind of videos are most likely to have such periodicity? Of these, how many can be better predicted by making use of periodicity? Besides temperature and day of the week, What other factors or events can also cause viewing periodicity? Answering questions like these definitely deserves further research.

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 7th ACM International Conference on Web Search and Data Mining*, WSDM '14 (New York, New York, USA, 2014), 593–602. ACM, New York, NY, USA. doi:10.1145/2556195.2566588. http://doi.acm.org/10.1145/2556195.2566588. (cited on pages 13 and 70)

Aggarwal, C. C. and Reddy, C. K., 2013. *Data clustering: algorithms and applications*. CRC Press. (cited on page 89)

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 Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13 (Rome, Italy, 2013), 607–616. ACM, New York, NY, USA. doi:10.1145/2433396.2433473. http://doi.acm.org/10.1145/2433396.2433473. (cited on pages 23 and 70)

Alexa.com, 2015. Most popular websites. http://www.alexa.com/topsites. (cited on page 3)

Allocca, K., 2011. Why videos go viral. http://www.ted.com/talks/kevin_allocca_why_videos_go_viral. (cited on pages 4 and 42)

Asur, S. and Huberman, B. A., 2010. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, 492–499. IEEE Computer Society, Washington, DC, USA. doi:10.1109/WI-IAT.2010.63. http://dx.doi.org/10.1109/WI-IAT.2010.63. (cited on page 1)

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J., 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74. ACM. (cited on pages 48, 99, 101, and 114)

Bass, F. M., 1969. A new product growth for model consumer durables. *Management Science*, 15, 5 (1969), 215–227. doi:10.1287/mnsc.15.5.215. http://dx.doi.org/10.1287/mnsc.15.5.215. (cited on page 21)

Bellman, R., 1961. On the approximation of curves by line segments using dynamic programming. *Commun. ACM*, 4, 6 (Jun. 1961), 284–. doi:10.1145/366573.366611. http://doi.acm.org/10.1145/366573.366611. (cited on pages 15 and 51)

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J.; and Ross, K., 2009. Video interactions in online video social networks. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5, 4 (Nov. 2009), 30:1–30:25. doi:10.1145/1596990.1596994. http://doi.acm.org/10.1145/1596990.1596994. (cited on page 14)

Berkhin, P., 2005. A survey on pagerank computing. *Internet Mathematics*, 2 (2005), 73–120. (cited on page 108)

Berndt, D. J. and Clifford, J., 1994. Using dynamic time warping to find patterns in time series. In *KDD workshop*, vol. 10, 359–370. Seattle, WA. (cited on page 93)

Bhagat, S.; Rozenbaum, I.; Cormode, G.; Muthukrishnan, S.; and Xue, H., 2007. No blog is an island analyzing connections across information networks. In *In Int. Conf. on Weblogs and Social*. (cited on page 101)

Borghol, Y.; Ardon, S.; Carlsson, N.; Eager, D.; and Mahanti, A., 2012. The untold story of the clones: content-agnostic factors that impact youtube video popularity. KDD '12. (cited on pages 23, 47, 50, 70, 99, and 101)

Borghol, Y.; Mitra, S.; Ardon, S.; Carlsson, N.; Eager, D.; and Mahanti, A.,

2011. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*, 68, 11 (2011), 1037–1055. (cited on page 11)

Brodersen, A.; Scellato, S.; and Wattenhofer, M., 2012. Youtube around the world: Geographic popularity of videos. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 (Lyon, France, 2012), 241–250. ACM, New York, NY, USA. doi:10.1145/2187836.2187870. http://doi.acm.org/10.1145/2187836.2187870. (cited on pages 12 and 100)

Broxton, T.; Interian, Y.; Vaver, J.; and Wattenhofer, M., 2013. Catching a viral video. *J. Intell. Inf. Syst.*, 40, 2 (Apr. 2013), 241–259. doi:10.1007/s10844-011-0191-2. http://dx.doi.org/10.1007/s10844-011-0191-2. (cited on page 12)

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P., 2010. Measuring user influence in twitter: The million follower fallacy. In *ICWSM '10: Proceedings of international AAAI Conference on Weblogs and Social*. (cited on page 101)

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 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07 (San Diego, California, USA, 2007), 1–14. ACM, New York, NY, USA. doi:10.1145/1298306.1298309. http://doi.acm.org/10.1145/1298306.1298309. (cited on pages 6, 9, 10, 29, 32, 50, and 100)

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S., 2009. Analyzing the video popularity characteristics of large-scale user generated content systems. *Networking, IEEE/ACM Transactions on*, 17, 5 (Oct 2009), 1357–1370. doi:10.1109/TNET.2008.2011358. (cited on page 1)

Cha, M.; PÃl'rez, J.; and Haddadi, H., 2012. The spread of media content through blogs. *Social Network Analysis and Mining*, 2, 3 (2012), 249–264. doi:10.1007/s13278-011-0040-x. http://dx.doi.org/10.1007/s13278-011-0040-x. (cited on page 101)

Chatzopoulou, G.; Sheng, C.; and Faloutsos, M., 2010. A first step towards understanding popularity in youtube. In *INFOCOM Workshops*, 1–6. (cited on pages 10, 14, 34, 50, and 100)

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J., 2014. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, 925–936. International World Wide Web Conferences Steering Committee. (cited on pages 24, 48, and 51)

Cheng, X.; Dale, C.; and Liu, J., 2008. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, 229–238. doi:10.1109/IWQOS.2008.32. (cited on pages 3, 10, 47, 50, and 100)

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*, 105, 41 (2008), 15649–15653. doi:10.1073/pnas.0803685105. http://www.pnas.org/content/105/41/15649.abstract. (cited on pages 6, 20, 21, 47, 48, 50, 52, 53, and 100)

Crane, R.; Sornette, D.; et al., 2008. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *AAAI Spring Symposium: Social Information Processing*, 18–20. (cited on pages 4, 21, 34, 50, and 70)

Das, G.; ip Lin, K.; Mannila, H.; Renganathan, G.; and Smyth, P., 1998. Rule discovery from time series. 16–22. AAAI Press. (cited on page 18)

Esling, P. and Agon, C., 2012. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45, 1 (2012), 12. (cited on pages 19 and 51)

Figueiredo, F., 2013. On the prediction of popularity of trends and hits for user generated videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13 (Rome, Italy, 2013), 741–746. ACM, New York, NY,

USA. doi:10.1145/2433396.2433489. http://doi.acm.org/10.1145/2433396.2433489. (cited on pages 3 and 22)

Figueiredo, F.; Almeida, J. M.; Gonçalves, M. A.; and Benevenuto, F., 2014. On the dynamics of social media popularity: A youtube case study. *ACM Trans. Internet Technol.*, 14, 4 (Dec. 2014), 24:1–24:23. doi:10.1145/2665065. http://doi.acm.org/10.1145/2665065. (cited on page 6)

Figueiredo, F.; Benevenuto, F.; and Almeida, J. M., 2011. The tube over time: Characterizing popularity growth of youtube videos. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11 (Hong Kong, China, 2011), 745–754. ACM, New York, NY, USA. doi:10.1145/1935826.1935925. http://doi.acm.org/10.1145/1935826.1935925. (cited on pages 3, 6, 10, 29, 50, and 70)

Fu, T.-c., 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24, 1 (2011), 164–181. (cited on pages 15, 16, 19, and 51)

Gill, P.; Arlitt, M.; Li, Z.; and Mahanti, A., 2007. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 15–28. ACM. (cited on pages 10, 32, and 50)

Golub, G. and Pereyra, V., 2003. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems*, 19, 2 (2003), R1. (cited on pages 54 and 55)

Gummadi, K. P.; Dunn, R. J.; Saroiu, S.; Gribble, S. D.; Levy, H. M.; and Zahorjan, J., 2003. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, SOSP '03 (Bolton Landing, NY, USA, 2003), 314–329. ACM, New York, NY, USA. doi:10.1145/945445.945475. http://doi.acm.org/10.1145/945445.945475. (cited on pages 1 and 9)

Guo, L.; Tan, E.; Chen, S.; Xiao, Z.; and Zhang, X., 2008. The stretched exponential distribution of internet media access patterns. In *Proceedings of the Twenty-seventh ACM Symposium on Principles of Distributed Computing*, PODC '08 (Toronto, Canada, 2008), 283–294. ACM, New York, NY, USA. doi:10.1145/1400751.1400789. http://doi.acm.org/10.1145/1400751.1400789. (cited on page 32)

Hastie, T.; Tibshirani, R.; and Friedman, J., 2009. *The elements of statistical learning*, vol. 2. Springer. (cited on page 91)

Himberg, J.; Korpiaho, K.; Mannila, H.; Tikanmaki, J.; and Toivonen, H., 2001. Time series segmentation for context recognition in mobile devices. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 203–210. doi:10.1109/ICDM.2001.989520. (cited on page 18)

Hoang, T.-A. and Lim, E.-P., 2012. Virality and susceptibility in information diffusions. In *International AAAI Conference on Weblogs and Social Media*. (cited on pages 99 and 101)

Huberman, B., 2013. Social computing and the attention economy. *Journal of Statistical Physics*, 151, 1-2 (2013), 329–339. doi:10.1007/s10955-012-0596-5. http://dx.doi.org/10.1007/s10955-012-0596-5. (cited on page 2)

Keogh, E., 1997. Fast similarity search in the presence of longitudinal scaling in time series databases. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, 578–584. doi:10.1109/TAI.1997.632306. (cited on pages 16 and 51)

Keogh, E.; Chu, S.; Hart, D.; and Pazzani, M., 2001. An online algorithm for segmenting time series. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 289–296. doi:10.1109/ICDM.2001.989531. (cited on pages 16, 17, and 19)

Keogh, E. and Kasetty, S., 2003. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Min. Knowl. Discov.*, 7, 4

(Oct. 2003), 349–371. doi:10.1023/A:1024988512476. http://dx.doi.org/10.1023/A:1024988512476. (cited on page 19)

Keogh, E. J.; Chu, S.; Hart, D.; and Pazzani, M., 2004. Segmenting time series: A survey and novel approach. In *Data Mining In Time Series Databases* (Eds. M. Last; A. Kandel; and H. Bunke), vol. 57 of *Series in Machine Perception and Artificial Intelligence*, chap. 1, 1–22. World Scientific Publishing Company. ISBN 978-981-238-290-0. (cited on pages 16, 17, 51, 60, and 62)

Keogh, E. J. and Pazzani, M. J., 1998. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. (cited on page 18)

Kleinberg, J. M., 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46, 5 (1999), 604–632. (cited on page 108)

Kooti, F.; Yang, H.; Cha, M.; Gummadi, P. K.; and Mason, W. A., 2012. The emergence of conventions in online social networks. In *ICWSM*. (cited on page 103)

Kwak, H.; Lee, C.; Park, H.; and Moon, S., 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10 (Raleigh, North Carolina, USA, 2010), 591–600. ACM, New York, NY, USA. doi:10.1145/1772690.1772751. http://doi.acm.org/10.1145/1772690.1772751. (cited on pages 32, 101, and 102)

Li, H.; Ma, X.; Wang, F.; Liu, J.; and Xu, K., 2013. On popularity prediction of videos shared in online social networks. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13 (San Francisco, California, USA, 2013), 169–178. ACM, New York, NY, USA. doi:10.1145/2505515.2505523. http://doi.acm.org/10.1145/2505515.2505523. (cited on page 24)

Lin, J.; Keogh, E.; Lonardi, S.; and Chiu, B., 2003. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th*

ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD '03 (San Diego, California, 2003), 2–11. ACM, New York, NY, USA. doi:10.1145/882082.882086. http://doi.acm.org/10.1145/882082.882086. (cited on pages 15 and 19)

Manning, C. D.; Raghavan, P.; and Schütze, H., 2008a. chap. Evaluation in information retrieval. Cambridge University Press. (cited on page 112)

Manning, C. D.; Raghavan, P.; and Schütze, H., 2008b. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. ISBN 0521865719, 9780521865715. (cited on page 107)

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 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12 (Beijing, China, 2012), 6–14. ACM, New York, NY, USA. doi:10.1145/2339530.2339537. http://doi.acm.org/10.1145/2339530.2339537. (cited on page 21)

Mitra, S.; Agrawal, M.; Yadav, A.; Carlsson, N.; Eager, D.; and Mahanti, A., 2011. Characterizing web-based video sharing workloads. *ACM Transactions on the Web (TWEB)*, 5, 2 (2011), 8. (cited on page 13)

Myers, S. A.; Zhu, C.; and Leskovec, J., 2012. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12 (Beijing, China, 2012), 33–41. ACM, New York, NY, USA. doi:10.1145/2339530.2339540. http://doi.acm.org/10.1145/2339530.2339540. (cited on page 101)

Newman, M., 2010. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA. ISBN 0199206651, 9780199206650. (cited on pages 107 and 108)

Newman, M. E. J. and Park, J., 2003. Why social networks are different from other

types of networks. *Phys. Rev. E*, 68 (Sep 2003), 036122. doi:10.1103/PhysRevE.68.036122. http://link.aps.org/doi/10.1103/PhysRevE.68.036122. (cited on page 32)

Page, L.; Brin, S.; Motwani, R.; and Winograd, T., 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/. Previous number = SIDL-WP-1999-0120. (cited on page 107)

Pavlidis, T., 1973. Waveform segmentation through functional approximation. *Computers, IEEE Transactions on*, C-22, 7 (July 1973), 689–697. doi:10.1109/TC.1973.5009136. (cited on pages 16 and 51)

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 Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13 (Rome, Italy, 2013), 365–374. ACM, New York, NY, USA. doi:10.1145/2433396.2433443. http://doi.acm.org/10.1145/2433396.2433443. (cited on pages 3, 23, 29, 37, 47, 51, 70, 84, 98, 101, 105, and 111)

Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 2 (1989), 257–286. (cited on pages 49 and 56)

Romero, D. M.; Meeder, B.; and Kleinberg, J., 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, 695–704. ACM. (cited on page 48)

Sornette, A. and Sornette, D., 1999. Renormalization of earthquake aftershocks. *Geophysical Research Letters*, 26, 13 (1999), 1981–1984. doi:10.1029/1999GL900394. http://dx.doi.org/10.1029/1999GL900394. (cited on pages 20 and 52)

Sornette, D. and Helmstetter, A., 2003. Endogenous versus exogenous shocks in systems with memory. *Physica A: Statistical Mechanics and its Applications*, 318,

3-4 (2003), 577 – 591. doi:http://dx.doi.org/10.1016/S0378-4371(02)01371-7. http://www.sciencedirect.com/science/article/pii/S0378437102013717. (cited on pages 6, 20, 22, and 52)

Stone, H., 1961. Approximation of curves by line segments. *Mathematics of Computation*, (1961), 40–47. (cited on pages 15 and 51)

Szabo, G. and Huberman, B. A., 2010. Predicting the popularity of online content. *Commun. ACM*, 53, 8 (Aug. 2010), 80–88. doi:10.1145/1787234.1787254. http://doi.acm.org/10.1145/1787234.1787254. (cited on pages 23, 37, 47, 51, 70, 98, 100, 105, and 111)

Terzi, E. and Tsaparas, P., 2006. Efficient algorithms for sequence segmentation. In *SDM*, 316–327. SIAM. (cited on page 17)

Wang, Z.; Sun, L.; Chen, X.; Zhu, W.; Liu, J.; Chen, M.; and Yang, S., 2012a. Propagation-based social-aware replication for social video contents. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12 (Nara, Japan, 2012), 29–38. ACM, New York, NY, USA. doi:10.1145/2393347.2393359. http://doi.acm.org/10.1145/2393347.2393359. (cited on page 101)

Wang, Z.; Sun, L.; Wu, C.; and Yang, S., 2012b. Guiding internet-scale video service deployment using microblog-based prediction. In *INFOCOM, 2012 Proceedings IEEE*, 2901–2905. doi:10.1109/INFCOM.2012.6195726. (cited on pages 1 and 101)

Wattenhofer, M.; Wattenhofer, R.; and Zhu, Z., 2012. The youtube social network. In *ICWSM 2012*. (cited on pages 3 and 11)

Wu, F. and Huberman, B. A., 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104, 45 (2007), 17599–17601. doi:10.1073/pnas.0704916104. http://www.pnas.org/content/104/45/17599.abstract. (cited on page 34)

Wu, F. and Huberman, B. A., 2008. Popularity, novelty and attention. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, EC '08 (Chicago, Il, USA, 2008), 240–245. ACM, New York, NY, USA. doi:10.1145/1386790.1386828. http://doi.acm.org/10.1145/1386790.1386828. (cited on page 34)

Xie, L.; Natsev, A.; Kender, J. R.; Hill, M.; and Smith, J. R., 2011. Visual memes in social media: Tracking real-world news in youtube videos. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11 (Scottsdale, Arizona, USA, 2011), 53–62. ACM, New York, NY, USA. doi:10.1145/2072298.2072307. http://doi.acm.org/10.1145/2072298.2072307. (cited on page 29)

Yang, J. and Leskovec, J., 2011. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11 (Hong Kong, China, 2011), 177–186. ACM, New York, NY, USA. doi:10.1145/1935826.1935863. http://doi.acm.org/10.1145/1935826.1935863. (cited on pages xxiv, 6, 21, 22, 28, 69, 91, 94, 102, and 110)

Yang, J.; McAuley, J.; Leskovec, J.; LePendu, P.; and Shah, N., 2014. Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 (Seoul, Korea, 2014), 783–794. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland. doi:10.1145/2566486.2568044. http://dx.doi.org/10.1145/2566486.2568044. (cited on page 50)

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q., 2012. We know what @you #tag: does the dual role affect hashtag adoption? WWW '12 (Lyon, France, 2012), 261–270. (cited on pages 24, 99, 101, 108, and 109)

YouTube.com, 2015. Statistics of youtube. https://www.youtube.com/yt/press/statistics.html. (cited on pages 2 and 3)

Yu, H.; Xie, L.; and Sanner, S., 2014. Twitter-driven youtube views: Beyond individual influencers. In *Proceedings of the ACM International Conference on Multime-*

dia, MM '14 (Orlando, Florida, USA, 2014), 869–872. ACM, New York, NY, USA. doi:10.1145/2647868.2655037. http://doi.acm.org/10.1145/2647868.2655037. (cited on page 4)

Zhu, C.; Byrd, R. H.; Lu, P.; and Nocedal, J., 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23, 4 (Dec. 1997), 550–560. doi:10.1145/279232.279236. http://doi.acm.org/10.1145/279232.279236. (cited on page 55)

Zink, M.; Suh, K.; Gu, Y.; and Kurose, J., 2008. Watch global, cache local: Youtube network traffic at a campus network: measurements and implications. In *Electronic Imaging 2008*, 681805–681805. International Society for Optics and Photonics. (cited on page 12)