# Characterizing and Modeling Popularity of User-generated Videos

Youmna Badr Borghol

A Dissertation
Presented to the Faculty of Engineering at
the University of New South Wales
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
Advisers:
Dr. Sebastien Ardon
Dr. Anirban Mahanti

June 2012

# Abstract

User-generated video content sites such as Youtube have become extremely popular. Understanding video online popularity is of great value for network and service providers, marketing industries, entertainment businesses and content creators. In this thesis, we develop novel frameworks to evaluate the impacts of different content-agnostic factors on Youtube videos popularity and use the resulting insights to study and model their popularity growth patterns.

The first significant subject of our thesis is developing and applying a novel methodology that is able to accurately assess, both qualitatively and quantitatively, the impacts of various content-agnostic factors on video popularity. When controlling for video content, we observe a strong linear "rich-get-richer" behavior, with the total number of previous views as the most important factor except for very young videos. We analyze a number of phenomena that may contribute to rich-get-richer, including the first-mover advantage, and search bias towards popular videos. Our findings also confirm that inaccurate conclusions can be reached when not controlling for video content.

The second central topic of our research is performing a characterization and modeling of the videos popularity dynamics using only the total view count for analysis. We develop a framework for studying the popularity dynamics of user-generated videos, present a characterization of the popularity dynamics, and propose a model that captures the key properties of these dynamics. Using a dataset that tracks the views to a sample of recently-uploaded Youtube videos over the first eight months of their lifetime, we study the popularity dynamics. We find that the relative popularities of the videos within our dataset are highly non-stationary, owing primarily to large differences in the required time since upload until peak popularity is finally achieved, and secondly to popularity oscillation. We propose a model that can accurately capture the popularity dynamics of collections of recently-uploaded videos

as they age. Another important aspect of our research is illustrating the biases that may be introduced in the analysis for some choices of the sampling technique used for collecting data.

# Acknowledgements

I would like to thank the many people that have helped me on the path towards this dissertation.

First and foremost, I would like to thank my advisers Sebastien Ardon and Anirban Mahanti. Without their guidance and encouragement, none of the work presented in this thesis would have been possible. They both have made my PhD experience enlightening and enjoyable at the same time. Sebastien, I greatly appreciate your support from the very beginning, throughout and up until the completion of this thesis. I also want to thank you for the time and freedom that you gave me so that I can find my own path in the field. Your constant reassurance, enthusiasm, knowledge and go-for-it attitude have been immensely inspiring qualities. Anirban, thank you for always being ready to offer your guidance when needed and thank you for making me push myself harder. I greatly appreciate your invaluable advice and criticism as well as your sense of humor. I am also grateful for the many hours you patiently put into reading my work.

Secondly, I wish to express my warm and sincere thanks to Niklas Carlsson. He has been an amazing co-author and collaborator in these last four years. This work would not have been possible without his invaluable feedback and support throughout the process. I would especially like to thank Niklas, for offering me the internship opportunity at Linköping University in Sweden. During this time, a major part of this thesis was completed. Derek Eager also deserves many thanks. His expertise and attention to detail has helped improve the quality of this work.

I would also like to thank my colleagues and friends at NICTA for their friendship over the years. Special acknowledgment goes out to Feisilia Tan and Rema Zogabe, whose company and support made my research experience much more enjoyable.

This research was kindly supported by a PhD scholarship from National ICT Australia (NICTA). I would like to thank NICTA for their amazing support and

most importantly their financial support which has allowed me to complete my thesis and participate in numerous national and international conferences.

I acknowledge and appreciate the love and support of my family, and I return it to them tenfolds. My dear parents, Badr and Salwa Borghol, you are the main reason for me being where I am today. Without your continued support I could never have accomplished so much. Thank you for helping me become the independent and confident person that I have grown into. I love you! Furthermore, I could not have done this without the constant love of my amazing brother and sister. Mohamad, I'm lucky to have a brother like you by my side. Ghina, you mean the world to me. I love you both, and I am so proud you.

Finally, I would like to thank my wonderful and loving husband who offered his endless support on an everyday basis and who was by my side through all the ups and downs of this journey. Jamil Helweh, thanks for always believing in me and thanks for being my rock. I am so lucky to have you.

Dedicated to Mom, Dad, Mohamad, Ghina and Jamil. I love you.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

> 'look at 2006 through a different lens and you'll see another story, one that isn't about conflict or great men. It's a story about community and collaboration on a scale never seen before. It's about the cosmic compendium of knowledge Wikipedia and the million-channel people's network YouTube and the online metropolis MySpace. It's about the many wresting power from the few and helping one another for nothing and how this will not only change the world, but also change the way the world changes.'
>
> You  Yes, You  Are TIME's Person of the Year. By Lev Grossman
> Time Magazine, 25 December 2006

When Time magazine selected 'You' as person of the year in 2006, where 'You' represented the Internet users, it obviously recognized the significance and impact of user-generated content (UGC). UGC is any online content produced and published by the users themselves. Different types of UGC, including images, videos, articles, audio files, and reviews, are generated and published continuously on a myriad of platforms such as social networking sites, blogs, podcasts, wikis and social media sites.

The emergence of UGC was the outcome of a new phase in the life of the Web. Until the early 2000's (a period referred to as Web or Web 1.0), a relatively small number of organizations created and published content and most end-users were restricted to passive viewing of the available content. Since then, the Web experienced a tremendous transition in how it was used by businesses, software developers, and individuals. It evolved from a static Web into a more dynamic, interactive, collaborative, and social Web.

The term 'Web 2.0' was coined by Tim O'Reilly to describe this second generation of the World Wide Web. Web 2.0 is user-centric; it's focused on empowering the user to take part in producing and shaping the content. If Web 1.0 was a monologue to users, Web 2.0 is a dialogue with users. This architecture of participation has led to the development and evolution of social media, 'a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content' [40].

The phenomenal success of online media sharing and social networking services has resulted in massive volumes of user-generated content being created and has spawned new content consumption approaches. One important and influential form of UGC is user-generated video. Popular video-sharing services such as YouTube host a large and rapidly growing catalogue of user-generated videos, and serve large numbers of video streams each day. Any video uploaded on YouTube has an enormous potential for reaching audiences and thus could have widespread impacts on the technical, social, and political sphere. The extent and implications of the evolution of user generated content gave rise to a great research interest in this area and raised many fundamental research questions and problems. In this thesis, we focus on one such problem: understanding the online popularity of user-generated video content.

In this chapter, we begin by providing background information about user-generated content in general, and video content in particular. We then discuss the

research questions we raise, the problems we are trying to solve, the motivation behind this research, and the contributions of this work. Section 1.1 examines the drivers contributing to the emergence of UGC. Sections 1.2 and 1.3 discuss on a broad scale the motivation behind our interest in the area of user-generated video content. We first examine the potential structural impacts of UGC on the social, political sphere and economics, respectively. Next, we discuss the scale of video content generated by users on social media platforms such as YouTube. Section 1.4 highlights the motivation of our work. Section 1.5 presents the main objectives in this thesis and Section 1.6 describes our key contributions. We conclude with an outline of the remainder of the thesis.

## 1.1 Drivers of user-generated content

Tim O′reilly defined Web 2.0 as 'a set of economic, social, and technology trends that collectively form the basis for the next generation of the Internet – a more mature, distinctive medium characterized by user participation, openness, and network effects'. The appearance of a series of technologies and services was a key factor contributing to the emergence of UGC. User-generated content has been rapidly growing with the the global penetration and development of broadband and software technologies and the dramatical decrease in production costs. High speed connections translated into higher Internet usage and spawned new user content production. Users became able to upload larger media files with less time and were no longer limited to low quality graphics and texts. The increasing popularity of wireless broadband technologies was another factor impacting the growth of UGC as it allowed users to be connected from literally everywhere. The rise of accessible and advanced software tools facilitated the process of content creation and publishing without the need of prior professional knowledge.

In addition to the technological factors, several social factors influenced UGC adoption: the desire of people to express themselves and their thoughts, their need to connect, and their aspiration to collaborate and develop communities. Given the chance to connect, onlines users accommodated to the new Web transformations and changed their consumption habits; they did not only connect, they engaged.

UGC adoption was also influenced by economic drivers. Diverse commercial players have acknowledged the new revenue opportunities resulting from UGC; monetization of UGC has began and continues to rapidly grow [36]. Fearing the loss of profits in traditional media, and exploiting the opportunity to leverage the 'long tail' [6] were the primary motivations for these investments.

## 1.2 Implications of user-generated content

The rise of UGC offered new opportunities to individuals to participate in powerful ways. User participation and collaboration on a massive scale has been changing the world. UGC has revolutionized our lives today and resulted in radical social, political, and economic transformations.

### 1.2.1 Social Impacts

UGC has shaped a new structure of communication and caused a shift from a passive culture of content consumption to a participatory one. The increased level of communication and interactivity has positively influenced the quality of life of individuals. When offered the opportunity to produce an online identity, users have expressed themselves, shared their experiences, connected with individuals around the globe, and found support through online communities. In addition, users have experienced more active social relationships built around exchange, without space or time boundaries. UGC has made it easier to meet new people, discover others with

similar interests, and sustain friendships. UGC has also influenced the personal and societal attitudes, values and behaviors.

UGC has democratised the media production environment by diffusing the power to generate cultural goods such as music. Today, almost anyone can unleash their creativity and show their skills and talents. Any Internet user can become a comedian, a music star, or a filmmaker.

UGC has also encouraged a culture of sharing. Sharing knowledge and skills has positively impacted the educational development worldwide. One of the greatest examples is the birth of Wikipedia. This huge online user-generated and user-edited encyclopedia is a 'cosmic compendium of knowledge' [32] built as a result of collective intelligence.

The collaborative site has sustained an enormous growth rate; it has currently 3,865,591 articles, 16,213,885 users, and the total number of user edits exceeds 500 millions [2]. An additional implication of the culture of sharing is the appearance of online communities used to gain knowledge, exchange information, and discuss ideas. 'My language exchange' is an interesting example of an educational online community where individuals practice and learn foreign languages from each other. The community has more than 1 million members from 133 countries, practicing 115 languages [1]. Online health communities are another example; here users find health care information and gain clinical knowledge from other patients and medical professionals. In fact, in 2011, a survey found that 80% of Internet users search for health information online [37].

A key feature of user-generated content is the potential to rapidly reach a global audience. This feature has been empowering users to accomplish any goal through building public will, raising public awareness and mobilizing people to a cause. In January 2010, news about the tragic earthquake of Haiti spread rapidly worldwide via countless videos on Youtube, microblogs on Twitter, images and status updates

on Facebook, etc. Online activity was at the heart of spreading awareness and information and at the center of a mass mobilisation that raised millions of dollars [25].

### 1.2.2 Political Impacts

UGC has positively impacted transparency in politics, which led to reduced corruption and increased deliberative democratic process. Launched in 2007, OpenCongress.org. is an example of a popular website for government transparency in the federal U.S. Congress.

The era of social media has increased user involvement in the political process. Users utilize the UGC platforms to initiate public debates and exchange opinions, analysis, and political commentary. The powerful user engagement can influence the political selection process. Obama′s use of social media is considered to have contributed to his presidential win. Some even stated that his win was driven more by social media than by his Harvard degree and his political ideology [13]. 'Were it not for the Internet, Barack Obama would not be president' said Arianna Huffington, editor-in-chief of The Huffington Post, while speaking on a panel at the Web 2.0 Summit in San Francisco [52]. Through the approximately 2,000 YouTube videos, Obama spoke to the citizens about his views and plans to raise the USA.

As user-generated services gained popularity, audience engagement in news production increased dramatically and led to rise of citizen journalism. The power of established media institutions, editors, and publishers shifted to the online user. Today every citizen is a reporter. Today it's the people who are in control. Users participation and active role in delivering the news has changed and democratized online professional journalism. Adopting and recognizing the strong user engagement in news creation became a key to the success and continuity of the traditional mass media. The Telegraph and many other newspapers provide their readers with services to blog, upload photos and exchange opinions on forums. The CNN's iReport website

and the Guardian's CommentIsFree separate the user-generated content by providing different platforms for the readers to blog and discuss the issues of the day.

The developed publishing technology available to users through social media platforms gave citizens the power to impact and shape political events. Most recently, user-generated content has played a critical role in the pro-democracy upsurges of protests and demonstrations currently occurring across North Africa and the Middle East, commonly referred to as the 'Arab Spring' uprising [50]. From Tunisia and Egypt to Libya, Bahrain, Syria, and Yemen, the social media and UGC have been the power tools in the propagation of democratic revolution. The liberation technology was used by protesters to speak up, virally spread information, coordinate actions, and produce intense public activism. In Tunisia, after Mohammed Bouazizi set himself ablaze, it was videos about the abusive state in the country, uploaded and watched on YouTube and Facebook, that inspired and mobilized people across the country to stand up for democracy and freedom [50]. In Egypt, the massive-scale revolts that caused President Hosni Mubarak's authoritarian regime to fall were sparked by a Facebook page created by Google executive Wael Ghonim. The page was in memory of a man arrested and beaten to death by the country's police in 2010. The extent of public support for the 2011 Egyptian revolution grew tremendously through the social media and shaped the future of the country. In an interview with CBS's 60 Minutes Ghonim said: 'Because the whole thing before the revolution was the most critical thing. Without Facebook, without Twitter, without Google, without You Tube, this would have never happened.' The government of many authoritarian regimes (e.g Iran, Tunisia and Egypt), made huge efforts to suffocate social media by banning Twitter, Facebook, and video sites such as YouTube and DailyMotion because they recognized the political impact of the combination of social networks and user generated content. This combination has been the trigger and catalyzer of many revolutions and has changed forever the political future of many countries. In

contrary to Gladwell's assertion in late 2010 that 'the revolution will not be tweeted', the revolution has been tweeted.

### 1.2.3 Economic Impacts

The culture of participation has direct economical impact on various industries; it is a disruptive force that created new economic opportunities and introduced new challenges. UGC can impact the profits of entertainment business, marketing and advertising industries. In fact, the survival of many commercial products, media, recreation and technologies rely on a significant user base. The power of UGC to affect purchase decisions forced different industry sectors and firms to transform the way they deal with their customers and do business. The consumer became a co-creator with an increased business value and increased power. And as a consequence, the world of web marketing and advertising has been radically altered. The new concept involves the users and communities to create the campaign or brand identity. One of the massively successful Internet campaign is the Australian Government's 'Best job in the world' campaign to promote tourism in Queensland. Applicants were requested to post videos on a special page on Youtube to explain why they were the right candidate for the job. The person chosen for the position was required to post video blogs every week to promoting the Great Barrier Reef. It even scooped the two top awards at the Cannes Lions International Advertising Festival. The campaign produced more than $70 million worth of worldwide publicity.

Another example demonstrating the effect of user-generated content on economy is the significance of reviews on the consumers buying behavior [72, 31]. Reviews are becoming one of the key factors for success on travel websites such as Trip Advisor [47].

## 1.3 User-generated video content: scope and scale

User-generated video is one dominant form of user-generated content. The arrival of YouTube, the world's largest video-sharing website, in February 2005, marked the beginning of the online video era. In 2006, after YouTube has been acquired by Google Inc. for $1.65 billion, the online video market began to explode. The number of viewers and the volume of video content consumed have increased dramatically. By mid-2006, YouTube had approximately 65,000 video uploads and 100 million video requests per day [66]. The bandwidth consumed by YouTube in 2007 is evaluated to be approximately equal to the entire Internet in 2000 [62]. Since 2006, the online video industry has experienced a 706% impressive growth [28], to become nowadays a tremendous global market. YouTube today is the most powerful online video provider worldwide, with more than 90 billion videos viewed in October 2011 and a market share of approximately 45% [26]. In January 2012, YouTube reported that users around the world upload an average of one hour of video to the site every second, and that the site has surpassed four billion global views a day (more than half the world's population) [14]. More recently, the most popular video-sharing site with revenues of approximately $10 billion a quarter [65], invested more than $100 million into original content channels, from partners including the Wall Street Journal, the online magazine Slate and Madonna [64].

Today, over half of all Internet traffic is web video. Based on the current trends, Cisco predicted in November 2011 that online video will soon account for more than 90% of all Internet traffic [63]. The prediction was later confirmed by YouTube [55], the God Father of the online video revolution.

So far, we have identified various reasons why UGC is an area of significant research interest. We have discussed the substantial implications of user-generated content in general and demonstrated the enormous scale of user-generated video con-

tent in particular. Next, we discuss on a narrower scale the motivation behind this work.

## 1.4 Motivation

With the tremendous and rapidly growing number of online user-generated videos, there is interest in understanding the popularity characteristics of user-generated video content and understanding the characteristics and processes governing their popularity dynamics. There is interest as well in understanding what factors cause some videos to become more popular than others.

Such insights are valuable for myriad reasons. First, content popularity may have significant impact on system design issues. New and efficient content distribution approaches can be developed using workload models developed from characterization of user-generated content usage [18, 7]. Understanding the popularity characteristics of user-generated video content can be helpful in identifying potential bottlenecks in discovering content [18]. Second, understanding online popularity is crucial to detect the onset of a video becoming popular. Online popularity impacts the profits of entertainment business, marketing and advertising industries. In fact, the survival of many commercial products, media, recreation and technologies survives on grounding popular content. Thus, such insights can be useful for the design of marketing and advertising campaigns [40, 45]. Third, from a theoretical point of view, the massive amount of data available from these online services provides an unprecedented opportunity to understand the underlying social behavior and collective human dynamics governing content creation and consumption processes [27, 59].

Vast amounts of new video, audio, image, and text content are created each year. What determines which items become popular and which do not? Several factors can impact the popularity of user-generated videos. Although the content of the

item (is it interesting, is it topical, is it high-quality, and so on) plays an important role, it has been widely recognized that other "content-agnostic" factors can also have a substantial impact on popularity. For videos shared through a site such as YouTube for example, content-agnostic factors that may impact a video's current viewing rate include the video uploader's social network size, the video's current view count, the video's title, the keywords associated with the video, the time that has elapsed since the video was uploaded (the video "age"), and the service provider's search and featuring algorithms that also influence a video's popularity. Such factors can directly impact the choices of potential viewers, as well as indirectly impact these choices through their influence on the service provider's search and featuring algorithms. Since both content-related and content-agnostic factors impact video popularity, understanding how content-agnostic factors influence popularity, and separating the influence of content and non-content related factors has been challenging. For instance, videos uploaded by users with large social networks may tend to become more popular because they generally upload more interesting content, not because social network size has any direct impact on popularity. Prior studies have used datasets consisting of videos with widely-varying contents, and thus are unable to rigorously distinguish the impacts of content-agnostic factors on popularity, from the impacts arising from differing contents.

In the first part of this thesis, we provide a qualitative and quantitative answer to this fundamental question [15]. We develop and apply a methodology that is able to accurately assess the impact various content-agnostic factors have on popularity. Our methodology is based on studying the dynamics of identical copies of a video content; we refer to such videos as *clones*. This approach allows us to control the bias introduced when studying videos that do not have the same content. When controlling for video content, one of our key findings was observing a strong linear "rich-get-richer" behavior, with the total number of previous views as the most important

factor, except for very young videos. When looking across different contents, we demonstrate that the rich-get-richer behavior gets weaker and becomes inaccurate, and thus rich-get-richer type of models do not accurately capture content popularity.

In the second part of this thesis we perform a characterization and modeling of the videos popularity dynamics using only the total view count for analysis [16]. We study and model how viewing rates of user-generated videos change over time, which we refer to as *popularity dynamics* or *popularity evolution*. Studying the popularity dynamics, however, is challenging because of the extremely large and rapidly growing number of videos available from such services. We make several contributions that address this challenge. Given the huge volume of content available from popular services, and given that sampling may yield datasets biased towards content with elevated short-term and/or long-term popularity, we first tackle the issue of sampling techniques. We then examine the popularity dynamics and churn of user-generated videos using seemingly unbiased datasets collected from YouTube[1], and we show that current popularity is not a reliable predictor of future popularity. Based on this key finding, we propose a characterization of popularity evolution and a new model that can capture the popularity dynamics of a collection of videos.

## 1.5 Objectives

Fundamentally, we address in this thesis two aspects of content popularity. The first aspect deals with understanding the popularity characteristics and the factors impacting users' behavior in the choice of specific online content. Modeling the popularity of content in video-sharing services has been challenging due to the complex interactions among content quality, content highlighting and discovery chosen by the social media platform, and social influence among peers. While these factors make it hard to forecast popularity a priori, most previous works on popularity prediction have mainly

[1] http://www.youtube.com

proposed rich-get-richer types of models, where future success is based on early measurements of popularity [30, 73, 49, 19]. In the first part of this thesis, we aim to take the first step in isolating the influence of the impossible to measure factor, content quality. We attempt to understand how content-agnostic factors influence popularity. Our high level goals in the first part include the following:

- Developing a methodology that allows us to providing a wide ranging analysis that covers nearly all endogenous and exogenous content-agnostic factors impacting videos' popularity on YouTube.
- Investigating to what extent, the rich-get-richer model [51], explains the popularity evolution of user-generated video content.

The second part of this thesis deals with the second aspect of popularity; specifically, the long-term temporal dynamics of user-generated videos popularity. We use our results and insights to suggest a model that captures the temporal dynamics of views to videos. We also aim at addressing the issue of sampling biases introduced by the commonly utilized sampling techniques. While this issue could have direct impact on the reported results, no prior work has approached it. Our high level goals in the second part of this work include the following:

- Developing a framework that allows us to provide an in-depth analysis and characterization of the temporal aspects of content popularity.
- Developing a model that captures the key properties of the observed popularity dynamics.
- Developing methods that allow us to collect a dataset seemingly unbiased towards popular content.

## 1.6 Contributions

Our main contributions in this thesis are centered upon understanding the different factors that impact the evolution of popularity over time and the long-term temporal dynamics of user-generated videos popularity.

### 1.6.1 Factors impacting videos' popularity on YouTube

- **We develop a data collection methodology that is able to accurately assess the impacts various content-agnostic factors have on video popularity.**

  Our methodology is based on studying popularity differences among videos that have essentially the same content; i.e., can be considered as "clones" of each other. Popularity differences among clones can only be due to content-agnostic factors. Through manual exploration, search, and viewing of YouTube videos, 48 sets of videos were identified, with each set containing between 17 and 94 videos that are sufficiently similar in content that they can be considered clones. We used the YouTube developer's API and HTML scraping to extract video information and statistics for each of these 1,761 videos. This data was collected twice, with a week separation between the two collections, so as to capture both the "current" popularity of each video (as measured by the new views acquired over that week) as well as lifetime statistics.

- **We apply a multivariate linear regression and other statistical methods on our dataset to systematically determine the content-agnostic factors that most influence a video's current popularity.**

  In particular, by analyzing a large number of explicit measurable factors that are provided through the YouTube API, we find that the most significant content-agnostic factors are the total number of previous views and the video age. We

also show that determining the relative importance of these factors without controlling for video content (i.e., ignoring clone set memberships) would result in inaccurate results; in particular, the relative importance of factors such as video age and the number of followers of the uploader would be significantly overestimated.

- **We find that when controlling for video content, "rich-get-richer" preferential selection based on the total number of previous views appears to provide a good model of video popularity evolution.**

  Specifically, using regression analysis and statistical hypothesis testing we show that current video popularity, among videos of similar "generation" (age within a multi-year window), follows a scale-free rich-get-richer model with power-law exponent of approximately one. We also show that carrying out this analysis without controlling for video content would result in erroneously concluding that preferential selection is significantly weaker, not scale-free, with a power-law exponent smaller than one. We investigate a number of possible contributors to the observed rich-get-richer behavior, including the "first-mover" advantage and search bias towards popular videos.

- **We demonstrate that the total number of previous views becomes less significant for very young (newly-uploaded) videos that have not yet accumulated many views.**

  For such videos, we show that other factors such as uploader characteristics and the number of keywords become much more significant. Their significance is substantially underestimated, however, when not controlling for video content.

### 1.6.2 Temporal dynamics of user-generated videos popularity

- **We examine the popularity dynamics and churn, for our sample of recently-uploaded videos, over the first eight months of their lifetime.**

  An important observation is that the relative popularities of the videos are highly non-stationary. One cause of the observed non-stationarity is the presence of large differences in when videos peak in popularity. While a majority of the videos peak in popularity, as measured by weekly viewing rate, within the first six weeks of their lifetime, many others do not peak until much later. Another cause of non-stationarity is the presence of oscillations in video popularity.

- **We propose a three-phase characterization of popularity evolution for our sample of recently-uploaded videos.**

  This characterization is motivated by the observed non-stationarity in the relative popularity of these videos and the differences in how long it takes video popularity to peak. For each week, the videos are partitioned into three disjoint sets, based on whether they are *before*, *at*, or *after* their observed popularity peak. Grouping the videos in this manner, we identify several interesting properties of video popularity evolution. First, we find that within each set of videos, the distribution of the number of weekly views is heavy tailed, where the tails may be approximated by a lognormal distribution. Second, we find that these distributions are approximately week-invariant. Third, as a specific consequence of the second property, we find that the viewing rate at peak popularity is approximately independent of how long it takes videos to attain their peak popularity.

- **We develop a model, based on our three-phase characterization, that can capture the popularity evolution of newly-uploaded videos.**

  In particular, using only a small number of distributions based on the three-phase characterization, our model is able to generate synthetic datasets in which key characteristics and consequences of the video popularity dynamics match those observed in the empirical data, including the distribution of the weekly viewing rate for videos at a particular age, the distribution of total accumulated views to videos at a particular age, and measures of churn in the relative popularity of videos. The model is developed in two stages. We first present a basic model in which popularity churn results *only* from the movement of videos (at varying times) between their before-peak, at-peak, and after-peak phases. We find that this model successfully captures the first-order dynamics of popularity evolution and yields results matching most of the characteristics of the empirical data. So as to better capture churn characteristics, in particular hot set evolution, we present an extended model that adds a tunable degree of additional popularity variation by shuffling the popularities of the videos within each phase. Our model can be considered a first step towards a synthetic workload generator for user-generated video services.

### 1.6.3 Sampling biases

- **Our last contribution concerns biases that may be introduced in the analysis of user-generated video popularity owing to use of sampling techniques.**

  Sampling is necessary as popular services host millions of videos with restrictions on the rate at which data may be fetched from the service. Furthermore, sampling is not straightforward because services often restrict how videos may be discovered from these services. From YouTube, for example, videos may

be sampled from various "most-popular" lists (such as most viewed today, this week, this month, or all time most popular), the "recently-uploaded" list, or by searching using keywords. Evidently, sampling from any of the most-popular lists provides a set of videos that are biased towards popular content. In this work, we used the YouTube developer's API to collect two datasets, one based on sampling from the *recently-uploaded* videos, and another based on *keyword searches*. We tracked the views to these videos over an eight month period. Perhaps not surprisingly, we find that sampling based on keyword searches yields a dataset biased towards more popular content. Fortunately, however, our results suggest that the YouTube API call that returns details on recently-uploaded videos gives an unbiased sample of such videos.

## 1.7 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 presents related work within the context of the contributions of our work. Chapter 3 first describes our novel clone-based methodology and analysis framework. Second, it presents an analysis for the relative impacts of the measured content-agnostic factors on current video popularity and shows the importance of controlling for video content in this analysis. This Chapter also studies the applicability of rich-get-richer preferential selection models, and examines contributors to rich-get-richer behavior. Finally, Chapter 3 analyzes the content-agnostic factors impacting the popularity of newly-uploaded videos. Chapter 4 studies the temporal evolution of videos popularity with emphasis on understanding data collection strategies impact the findings. The Chapter begins by describing our data collection methodologies, the basic characteristics of our datasets and our measurement and analysis framework. Afterwards, some initial analyses concerning possible biases in the datasets owing to use of sampling techniques is

introduced. Chapter 4 presents our three-phase characterization of popularity evolution and provides the underpinnings for the model proposed in this work. The chapter ends by presenting the basic model, its extension, its validation, and insights drawn from the model. Chapter 5 concludes the thesis and gives directions for future work.

# Chapter 2

# Literature Review

The great success of online media sharing and social networking services has resulted in massive volumes of user-generated content being produced and spawned new research interests and activities related to the generation and evolution of the content consumption patterns. One area that has attracted interest from the scientific community lately has been the analysis of content popularity in social media platforms [60, 8, 70, 27, 29, 17, 18, 59, 67, 56]. In particular, the success of video-sharing services has led to a surge in research on various facets of YouTube and other similar services. There has been considerable prior work concerning measurements, analyses of these measurements, and/or models, for various user-generated video properties including popularity. Studies have examined the characteristics of user-generated video files [30, 21, 18, 49, 73], use of social networking features in video-sharing services [35, 49], the structure of YouTube's "friend" network [48], the use of the "video response" feature of YouTube [11], the different aspects of user-generated video metrics [19, 49, 21], the popularity characteristics of user-generated videos [18, 73, 30, 49, 17, 29], and also models for user-generated video popularity prediction [43, 60].

In this thesis, we are also interested in the analysis of user-generated video content popularity, with primary focus on the analysis of factors impacting users' behavior in social media applications, the characteristics of online content popularity, and the the temporal patterns of content consumption by users. We also focus on the issue of dataset biases introduced by the sampling techniques used for the selection of a suitable sample for study and analysis.

Many prior works have been centered upon characterizing online popularity and understanding the factors impacting users' behavior in the choice of specific online resources, which is of central importance in anticipating online content popularity [19, 20, 30, 71, 49, 73, 17, 29]. A large number of studies addressed the endogenous (internal to the online content concerned) factors influencing popularity on YouTube. Overall, these works focused mostly on one popularity metric: view count. Cha et. al [19] studied the correlation between the user participation and the view count of YouTube videos, but the user participation was represented by only the number of ratings. Chatzopoulou et al. [20] took the first step in studying popularity metrics other than view count, but still limited to about four other variables only, the number of comments, favourites, ratings, and the average rating. A few studies analyzed the exogenous (external to the online content concerned) factors influencing popularity. For instance, Figueiredo et al. [29] and Zhou et al. [71] both studied the significance and effect of video referrers, through which videos are being discovered, on the popularity evolution of Youtube videos. They reported that YouTube's internal search and recommendation engines are the central sources of views for videos [71, 29].

Most characterization studies have reported that the total view count of an online content is the key factor affecting its future popularity [30, 73, 49, 19]. It has also been established that the total view count distribution of user-generated content is heavy-tailed [30, 73, 49, 18], and subsequently, cumulative advantage and rich-get-richer types of models, where the probability that a content experiences an increase

in popularity is directly proportional to its current popularity, has been suggested to model content popularity [19, 60]. All the aforementioned studies, however, have disregarded the influence of the content quality, simply because it is hard to measure in a convincing manner. Our work is complementary as we are also interested in understanding the characteristics of user-generated videos' popularity and we address how different factors impact the users' behavior in the choice of specific content. We propose a new clone-based methodology that allows us to isolate the influence of the video quality. Our unique dataset offers the opportunity to study the significance and impact of content-agnostic factors while controlling for differences in video content. To the best of our knowledge, no prior work has separated out the impacts of content-related and content-agnostic factors on popularity. Using our novel dataset, we provide a wide ranging analysis that covers nearly all endogenous and exogenous metrics that could affect videos' popularity on YouTube. Our investigation of the content-agnostic elements validates the intuition that view count is the central factor. Our systematic study supports the hypothesis that popularity is governed by a strong rich-get-richer behavior, however, only when controlling for video content. When looking across different contents, we demonstrate that the rich-get-richer behavior gets weaker and becomes inaccurate, and thus rich-get-richer type of models do not accurately capture content popularity.

The temporal aspects of user-generated content popularity has been another major research interest in the last few years. There has been a parallel increase in interest in characterizing and modeling how UGC evolves over time. Works addressing the popularity dynamics have frequently relied on crawling the social media networks over and capturing a set of snapshots. Some studies have used only few snapshots and provided a short-term analysis of how popularity changes over time, while other works dealt with the longer-term evolution. Cha et al. [19] has performed a large-scale analysis addressing the user-generated video content; they analyzed the consumption

pattern and popularity distribution of videos on YouTube and Daum. Papadopoulos et al. [53] provided insights into content popularity evolution while considering the social aspects represented by the process of content rating by the user community. Many recent works have focused on predicting the future popularity of a content based on early popularity measurements, motivated again by rich-get-richer types of models. For example, using early measurements of user comments in online debates, different studies have proposed models to predict the popularity of online articles [61, 41], discussion threads on Slashdot [39], news on the social news portal Digg [44, 38], online discussion forums [43], and more. Huberman et al. proposed a model to predict the number of views to a Youtube video using its early view count [60]. Our work is complementary as we provide, in the second part of this thesis, a systematic study of the long term popularity evolution of videos. We show that although there is a strong correlation between early and future popularity measurements, individual video popularity is highly unstable and unpredictable. This observation in the long term popularity analysis, along with the observations in the first part of our work, motivate the need for a new model that captures the evolution of content popularity in time. We propose a model for how the popularity statistics of a collection of recently-uploaded videos evolve over time, instead of considering popularity evolution for individual videos.

An overview of previous research efforts addressing user-generated video content popularity are summarized in Sections 2.1, 2.2, 2.3, and 2.4. In this chapter, we restrict our attention mostly to related prior work on characterizating the popularity of user-generated videos, analyzing the impacting factors, and modeling the phenomenon of popularity evolution in video-sharing platforms.

## 2.1 Characterization of the UGC popularity distribution

There has been many research studies concerned with the analysis and characterization of user-generated video popularity. Previous studies on characterization has commonly used network traffic traces from a network gateway [30, 73] or meta-data sampled from video sharing services [18, 49, 29, 17, 4].

Both Gill et al. [30] and Zink et al. [73] analyzed YouTube video requests from a campus network and observed that the video requests follow a Zipf-like distribution and substantial network bandwidth savings would be feasible if large proxy caches were used. Zink et al. [73] performed a trace-driven study of caching policies confirming the benefits of local edge network caching.

Halvey et al. [35] investigated the use of social networking features on YouTube to understand community behavior. Their results showed that many users do not form social networks in the online media sharing platforms such as YouTube. Instead, social connections are formed by only a few users who frequently utilize the social interaction facilities available within the site. Halvey et al. did not observe a Zipf-like behavior in the popularity distribution of Youtube; however, they reported that the number of favourite videos and the number of uploads are best fit by a Zipf distribution.

Abhari et al. [4] provided a workload characterization study of YouTube video characteristics and user behavior using meta-data collected through crawling YouTube for a five months period. The authors then developed a synthetic workload generator for YouTube to study different proxy cache approaches for popular YouTube files. Their results suggested again that reduced network traffic and increased scalability of YouTube could be achieved if proxy caching of YouTube popular videos is used.

Plissonneau at al. [54] characterized the effect of YouTube traffic on an ADSL platform of an ISP in France. The authors investigated the users' behavior on YouTube and found that video viewing abortions happen mainly due to the lack of interest for the content, and not because of low network throughputs.

Several prior works have been centered upon studying the properties and structure of YouTube social networks. Mislove et al. [48] provided a comprehensive study of the network of YouTube friends, and compared this network of users to the network of Web pages and to two other online social networks. Biel [12] presented a large-scale static analysis of the network of YouTube subscriptions, which represents how users with mutual interests are connected to each other. The author also described the small-world, power-law, and reciprocity characteristics of the network of YouTube subscriptions.

Haddad et al. [33] presented a survey of many classes of caching policies on YouTube while investigation and discussing how to improve the scalability and performance on YouTube. The authors conclude that access time and boot delay in watching videos could be reduced by using local caching combined with prefetching techniques.

The aforementioned characterization studies revealed interesting properties regarding the popularity of user-generated videos and the traffic demand on media-sharing platforms. They confirmed the popularity skewness among online content and the heavy tailed popularity distributions, and they addressed YouTube performance issues such as scalability and large bandwidth demand. However, they do not address the temporal aspects of online videos popularity.

## 2.2 Temporal aspects of UGC popularity

Many recent studies have addressed the temporal aspects of the popularity dynamics, and were based on meta-data sampled from video sharing services. Few studies have analyzed different aspects of user-generated video metrics such as total views, total ratings, total comments, and uploader social network size (e.g., see [19, 49, 21]). Cha et al. [19] presented one of the first large-scale experimental work that studied the process of popularity evolution of user-generated videos. The authors used meta-data collected by crawling videos from Daum,the most popular UGC service in Korea, and YouTube. analyzed the consumption pattern and popularity distribution of uploaded videos. They found that the total views since upload for videos of various ages is best fit by a power law distribution with exponential cutoff. The exponential decay observed in the tail of the total views distribution was explained by information filtering through recommendation and search engines where only a small number of popular items is returned, and by a 'limited fetch' model where some users do not request the same content many times. Cha et al. [19] demonstrated a strong correlation between the view count and the user participation, but they limited the user participation to one popularity metric, the number of ratings. They reported the absence of any correlation between the video view count and its upload time, which is confirmed in our results. However, their analysis is limited to pairwise correlation to evaluate the impact of the video age on its popularity.

Mitra et al. [49] compared four popular video sharing workloads other than YouTube and established the presence of "invariants" among their characteristics, such as heavy-tailed total view count distributions and positive correlation between total views and total ratings to a video. In addition, Mitra et al. distinguish between lifetime and short-term popularity measures, and evaluate their respective degrees of relevance for cache management. They found total views popularity to be ineffective for making caching decisions due to variations in the viewing rates of videos.

Chatzopoulou et al. [20] investigated the popularity evolution of a collection of videos (instead of individual videos) by grouping videos into age bins. Starting from the standard feeds, the authors recursively collected the related videos of every video in the list. They took the first step in studying popularity metrics other than view count, but still limited to about four other variables only, the number of comments, favourites, ratings, and the average rating. They showed that all metrics, except the average rating, are highly correlated with the video view count and that the correlation increases with video popularity. They used linear regression to develop a model that evaluates the video view count as a function of the number of favorites and ratings. The model could be used to detect artificial boosting of video popularity. Chatzopoulou at al. report the small world network properties for the related video graph by focusing on a subgraph of only popular nodes. The authors took the first step in studying popularity metrics other than view count, but still limited to about four other variables only, the number of comments, favourites, ratings, and the average rating. They showed that all metrics, except the average rating, are highly correlated with the video view count and that the correlation increases with video popularity. Our work again is complementary as we study all the possible video variables provided by YouTube to determine which factors influence shaping the success trajectory of a content.

Cheng et al. [23] analyzed Youtube from an internal and external perspectives. Using meta-data crawled in a 1.5 year span, the authors provided an in depth study of the characteristics of YouTube. The authors presented an active life span model to study popularity trends and predict its future growth. In addition, the impacts of the external links of YouTube were investigated in [23], using information from approximately 1 million videos' external link. The authors reported that videos benefit from external links more in their early lifetime, an observation that we confirm in our analysis. Cheng et al. [21, 22, 23] also studied the relationship of related videos on

YouTube and showed that the related video graph has small-world characteristics and has a large clustering coefficient. They suggest subsequently the possibility of using new caching techniques and peer-to-peer content distribution systems to deliver the videos.

More recently, Zhou et al. [71] studied the influence of some of YouTube's referrers on views to videos: the featured, search and related videos referrers. They observed that Youtube search and the related videos recommendation system are the most important source of traffic to videos. The analysis was based on two datasets, one consisting of video requests at a university network gateway, and the other consisting of an initial set of featured videos and their related videos, collected recursively for three levels. They also observed a strong correlation between the views to a video and the number of views of its related videos, and suggest to consider it as a factor in video popularity prediction.

In recent work, Figueiredo et al. [29] studied popularity dynamics using three different YouTube datasets, specifically, a sample of most popular videos, a sample of deleted videos, and a sample of videos obtained via keyword searches. Similar to our work, the authors utilized a recently available feature of the YouTube API ('Insight Data') that provides high-level and limited statistics of how clicks to a particular video grows over time, along with some information on sources of these clicks. Figueiredo et al. show that the popularity characteristics of the three datasets are different, and that among different types of referrers, search and internal mechanisms are the most important sources of traffic to Youtube videos.

Papadopoulos et al. [53] presented an analysis framework to investigate the empirical properties of a Social Bookmarking System. Besides verifying the heavy-tailed behavior of popularity and studying the dynamics of content popularity in social media, the authors quantified the influence of the social factor on it. The proposed framework was validated using data from Digg, a social news website. The effect of

social influence on the evolution of popularity has been investigated in an interesting experiment by researchers at Columbia University [59]. The authors created an artificial music market in a form of website, where users can listen to 48 unknown songs from unknown bands, download them for free, and rate them. In the experiment, 14341 participants were randomly allocated to either a 'social influence' group, where they can see how many times the song has been previously downloaded, or an 'independent' group, where they chose which songs to listen to based only on the names of the bands and their songs. The results showed that social influence was the key factor impacting the popularity of content instead of its quality. The authors argue that popularity is unpredictable.

Unlike these prior works, our novel dataset of YouTube video clones allows us to study the significance and impact of content-agnostic factors while controlling for differences in video content. In addition, our analysis is more comprehensive, as we study all the possible internal and external video metrics provided by YouTube to determine which factors influence shaping the success trajectory of a content.

Cha et al. [19] first noted the presence of YouTube video clones which they referred to as "aliases". They observe that aliases tends to "dilute" popularity, as the views for the same content are spread out over several videos. The authors definition of a clone is not necessarily an identical copy of a content as it includes videos having non-overlapping parts up to one min. The authors did not use aliases to study how different factors influence content popularity, as we have done in our work.

## 2.3 Modeling the popularity evolution of user-generated videos

Understanding which newly-uploaded content will become popular has been a major research problem that attracted much attention. Many recent studies on popularity dynamics have focused on classification, modeling and prediction of online popularity.

### 2.3.1 Classification models

There has been interest in clustering and classifying user-generated videos based on the diversity in changes to the viewing patterns [17, 27]. For instance, Crane and Sornette [27] developed a method for classifying collective dynamics of a social system based on whether the factor responsible for a viewing activity was caused by internal or external influences. The authors implicitly define internal and external events as video referrers; they consider external links, embeds, and featured links as exogenous factors and search links, and internal links as endogenous factors. Crane and Sornette applied their model to a time-series of views to 5 million YouTube videos, and based on the diversity in video popularity growth patterns, they labeled videos as *viral*, *quality*, and *junk*. Figueiredo et al. [29] characterized the videos belonging to the different datasets they collected into the classes defined by Crane and Sornette [27]. Based on the videos' popularity evolution patterns, they classify the most popular video lists into the *quality* category, whereas the list of deleted videos, and the list of videos obtained via keyword searches into the *viral* category.

Popularity classes based on the heterogeneity in popularity growth behavior were also proposed in [17], and explained by the level of socialness of a video. Video socialness is defined by assigning the video referrers as social (external links and embeds) or non-social (search and internal mechanisms). Even though different explanations

were used for categorization in [27] and [17], similar growth patterns were found across classes.

Yang et al. [68] presented a time-series analysis associated with online content. The authors showed that content popularity can be classified in six different groups with distinct temporal shapes of attention. They presented a time-series clustering algorithm that finds clusters using their proposed similarity metric and suggested a simple prediction model that forecasts the shape of attention of online content. Yang et al. validated their approach using data from blogging (a set of 170 million blog posts and news media articles), and micro-blogging (a set of 580 million Tweets).

In a more recent work, Asur et al. [8] brought attention to the importance of content in variations among popularity. The authors proposed a model for forecasting a range of popularity of news items on Twitter, prior to their release. Instead of using early measurements as predictors of online popularity, they used the following properties obtained from the content of news articles: the source, the category, and the subjectivity in the language of the article. Asur et al. found that the source of the article is one of the key predictors of it's popularity. Their model achieve 84% accuracy, using classifiers. A similar study on the variations of the spread of content was performed by Romero et al. [57], where only the categories of Twitter hashtags is used for analysis.

### 2.3.2 Prediction models

Many studies addressed the issue of predicting the future popularity of user-generated content using early measurements of popularity. Early measurements are defined by early view counts on video sharing services such as YouTube, early votes on Digg, early number of comments on sites and forums, early number of likes on facebook, etc.

Avramova et al. [9] empirically studied and modeled the change in popularity of online videos. The popularity evolution of online videos is modeled using a closed-form expression. Depending on the form-determining parameter, the cumulative popularity can be represented either as a power-law distribution or an exponential decay distribution.

Lerman and Hogg [44] presented a stochastic model of user behavior to predict the future success of a newly uploaded user-generated content. The simple model is based on extrapolating from the early measurements of user reactions to a new content. The authors validated their approach through a dataset collected from Digg.

Perhaps most closely related to our work are the models by Szabo and Huberman [60] and Raktiewicz et al. [56]. Szabo and Huberman [60] presented a model for predicting the total future view counts gathered by a video based on the total view count received at the time of prediction. They tracked approximately 7,000 recently-uploaded YouTube videos for a period of one month and observed a strong linear correlation between the logarithmically transformed total views early in the lifetime and later in the lifetime of the video. The authors modeled future total view counts as the sum of the current total view count, plus a linear term, that defines the relationships between the logarithmically transformed total view counts accumulated over times, and a noise term, that captures the randomness in the data. The error margin produced by their approach is relatively large because the percentage error is based on the total views since upload. Szabo and Huberman also studied the popularity evolution on Digg. Their results showed that the user-generated contents on Digg have a faster decay in popularity when compared to video contents in YouTube.

Raktiewicz et al. [56] provided an analysis of the temporal aspects of online content popularity in two large-scale systems, the Wikipedia and the Chilean Web space. The authors suggested a model that combines the classical "rich-get-richer" model [10] with random popularity shifts, with the goal of capturing the influence of exogenous

events on content popularity. The model was validated using click-through data for Wikipedia and the Chilean Web.

Our work is complimentary, as after empirically demonstrating that individual video popularity is highly unstable and unpredictable, we proposed a model for how the popularity statistics of a collection of recently-uploaded videos evolve over time, instead of considering popularity evolution for individual videos.

## 2.4 Sampling biases

Social media platforms offer the opportunity to access tremendous amounts of human social dynamics over time. for Recent works on user-generated video content is using the most popular video-sharing site, YouTube, as the key source of large-scale data. A majority of the studies has typically relied either on network traffic traces from a network gateway [30, 73] or meta-data collected via crawling from video sharing services [19, 21, 34, 35, 49, 60, 43, 27, 29, 17]. Cheng et al. [21] indicated that the usual crawling techniques results with datasets biased towards popular content. However, none of the prior works, except [60], considered the issue of sampling biases. When using a biased, non-random sample of videos to study the popularity phenomenon, erroneous results may be attributed to the popularity evolution process, instead of the method of sampling.

A few studies introduced even more bias in the datasets by limiting the analysis to videos with a certain popularity threshold [17] or selecting only the first result in every keyword search entry list when using keyword-search techniques [29]. Our work builds upon this body of work [18, 73, 30, 49, 17, 29] by studying how biases occur owing to sampling. We demonstrate that biases are introduced when keyword-search techniques are used for sampling videos and we show that a seemingly unbiased

dataset for the analysis of UGC popularity could be extracted by tracking a collection of newly uploaded videos.

# Chapter 3

# Content-agnostic Factors that Impact YouTube Video Popularity

When analyzing the factors that most impact video popularity, it is important to take into account the content itself. Some content may attract more views because it simply is better content or because there are more people interested in this content. Analyzing large sets of identical videos, which we refer to as 'clone sets', with many videos of the same content, allow us to control the impact of content, provide an unbiased analysis and focus on capturing the content-agnostic factors that allow a video to attract more views. Our novel dataset allows us to answer questions that could not be answered in previous studies. We identified large clone sets, which allow a number of statistical methods to be applied on each clone set individually, as well as more advanced methods that take into account the impact of differences in contents across a larger set of videos.

The remainder of this chapter is organized as follows. Section 3.1 describes our data collection methodology, basic characteristics of our dataset including the variables captured for each video, and our analysis approach. Section 3.2 presents an analysis for the relative impacts of the measured content-agnostic factors on current

video popularity, while Section 3.3 shows the importance of controlling for video content in this analysis. Section 3.4 studies the applicability of rich-get-richer preferential selection models, and examines contributors to rich-get-richer behavior. Section 3.5 analyzes the content-agnostic factors impacting the popularity of newly-uploaded videos. Finally, Section 3.6 concludes the chapter.

## 3.1 Methodology

In this section, we describe our datasets, the data collection methodologies, the analysis approaches and techniques used for the examination of the factors influencing user-generated video popularity. In the following sections, we provide the analysis details and results.

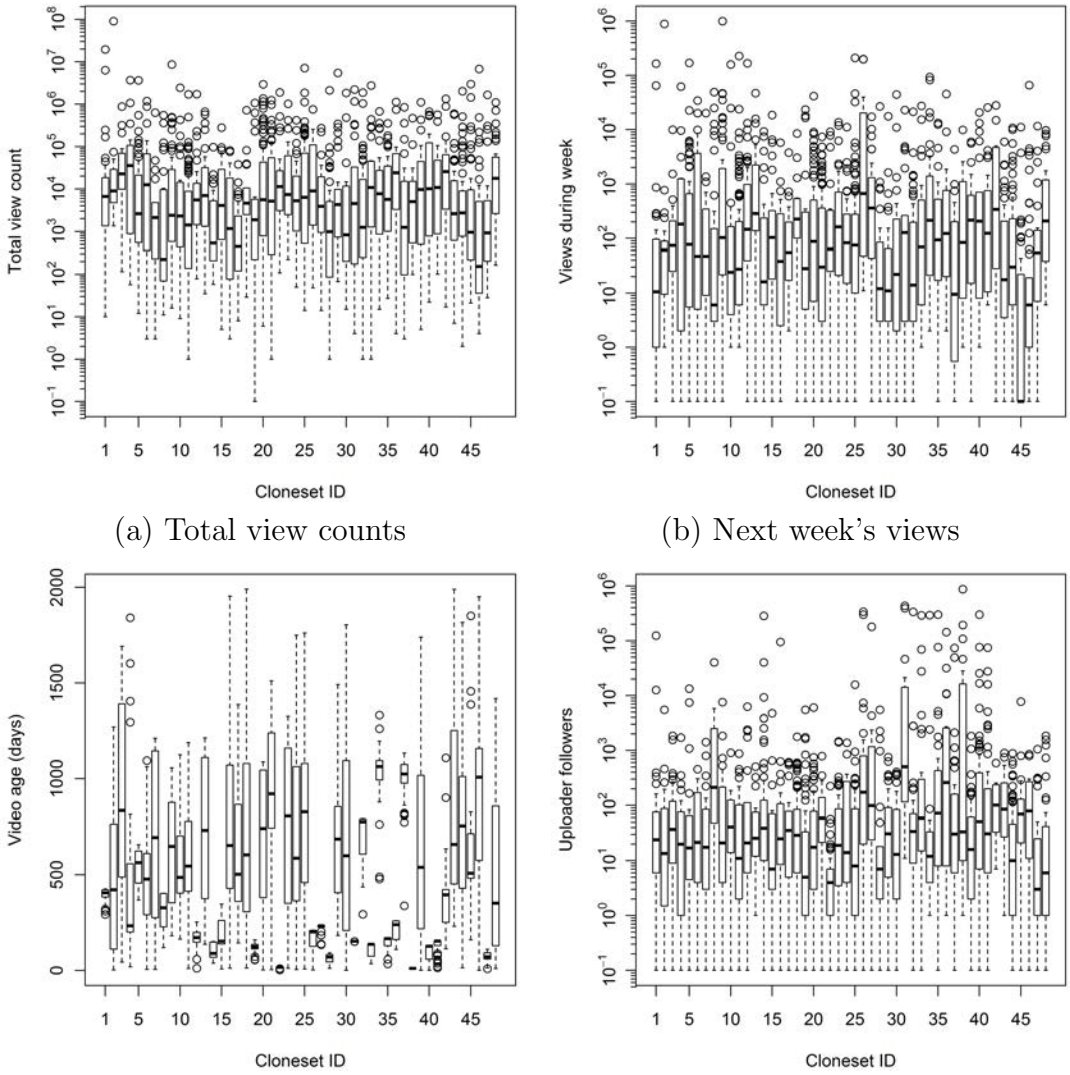
### 3.1.1 Data Collection

To analyse factors influencing video popularity, we start by identifying a large corpus of identical or nearly identical videos on YouTube. By identical we mean the same video content and audio soundtrack. We allow subtitles, variations in encodings (quality), and small variations in video duration. In this paper, we refer to such a set of nearly identical videos as a *clone set* and videos in such a set as *clones*. Through extensive exploration, search, and viewing of YouTube videos, we manually identified 48 clone sets, each of which contain between 17 and 94 clones, with a median size of 29.5. [1] In total we identified 1,761 videos.[2]

We developed a web-based collection system which allows us to easily enter clone video urls in a database. Each video entered is assigned a clone set id and a video id. Once in the database, the system then extracts video and uploader information using

[1] Our dataset is available at http://www.ida.liu.se/ nikca/papers/kdd12.html.

[2] Our initial dataset was somewhat larger, but we removed all videos whose duration deviated more than 15% from the median duration within their clone set.




(a) Total view counts (b) Next week's views

(c) Video age (d) Uploader followers

Figure 3.1: High-level clone set summary.

both the YouTube developer's API [69] and through HTML scraping. The system collects three types of information:

- **Video statistics:** These include statistics such as view count, uploader's followers count, number of comments, "likes" and "favourite" events and average rating. For each clone set, two snapshots were collected, spaced one week apart. For all videos in a clone set, the data collection was done as close together in

time as possible (within minutes). Table 3.1 describes all variables collected, and Figure 3.1 provides an overview of the variations of four example variables.

- **Historical view count:** When available, we extract historical video view counts from the YouTube HTML page. This information is referred to by YouTube as "insight data"; an example is presented in Figure 3.2. We programmatically obtain this historical view count information by intercepting the URL request which the YouTube website uses to plot the graph. This URL contains 100 points with date/view count pairs.

- **Influential events:** The YouTube insight data also contains information on how users discover a video. It reveals the top 10 "most significant" sources of discovery, or where the video was linked from. Common sources of discovery include "discovered through YouTube search" and "embedded on Facebook". We also collect this list of referrers and, for each referrer, the first date of referral and the associated view count.

The dataset used in this work was collected between February 2010 and April 2011.

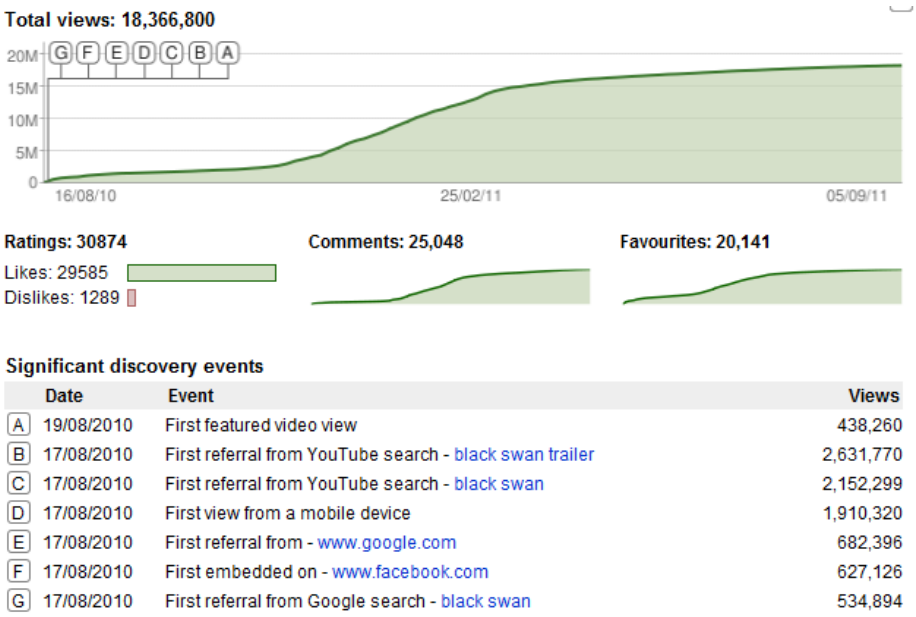



Figure 3.2: Example of YouTube insight data.

| Variable | Description | Type | Scale | Category |
|---|---|---|---|---|
| Clone set ID | Unique clone set identifier | – | – | – |
| Capture time | Time at which this video data was captured | – | – | – |
| Upload time | Time at which the video was first published | – | – | – |
| Update time | Time at which the video was last updated | – | – | – |
| Categories count | Number of categories associated with this video | – | – | – |
| Next week views | Number of views between two weeks | Predicted | log | Video popularity |
| Rating average | Average rating (min and max ratings also measured) | Predictor | linear | Video popularity |
| Total comments | Number of comments | Predictor | log | Video popularity |
| Total dislikes | Number of 'dislike' events | Predictor | log | Video popularity |
| Total favourites | Number of time this video was 'favourited' | Predictor | log | Video popularity |
| Total likes | Number of 'like' events | Predictor | log | Video popularity |
| Total ratings | Number of ratings | Predictor | log | Video popularity |
| Total view count | Number of views | Predictor | log | Video popularity |
| Uploader age | Age of the uploader | Predictor | log | Uploader characteristics |
| Uploader contacts | Number of (YouTube) 'friends' of the uploader | Predictor | log | Uploader popularity |
| Uploader followers | Number of followers for the uploader | Predictor | log | Uploader popularity |
| Uploader video count | Number of videos uploaded by the uploader | Predictor | log | Uploader popularity |
| Uploader view count | Number of time any of the uploader's videos were viewed | Predictor | log | Uploader popularity |
| Video age | Age of the video | Predictor | log | Video characteristics |
| Video keywords | Number of keywords assigned to the video | Predictor | log | Video characteristics |
| Video quality | The best quality (frame size) available for this video (higher is better) | Predictor | linear | Video characteristics |

Table 3.1: Variables collected and analyzed.

As mentioned previously, in this work, two videos are considered identical when they have identical audio and video content. Despite our best effort to identify such identical videos, some error is introduced by this manual data collection. These errors can be classified as objective errors (i.e., when the video or audio content is quite different), or subjective errors (i.e., when the video is close to identical, for example, the video can be pre-fixed by a short clip or logo advertising the uploader's identity or organization). We have taken care to ensure that most objective errors are eliminated. Subjective errors remain in the dataset and are inherent to the data collection method.

In addition, while the YouTube insight data provides valuable information regarding a video's popularity evolution, it also has some limitations. First, this data is not available for all videos, as uploaders can choose to hide it from public view. We could retrieve the insight data for approximately 40% of the videos in our dataset. Second, the historical view count data includes only 100 points, irrespective of the video's age. To extract the view count at a specific point in time, we applied linear interpolation, which introduces an error dependent on video age. Finally, the referrer's data only reveals 10 referrers, with the exact method used by YouTube to select which referrer to include in the list being unknown. This limits the number of views that can be mapped to a specific source, but also leaves some uncertainty in whether there are other more significant sources not accounted for. In our analysis we try to minimize the effect of these limitations.

### 3.1.2 Analysis Approach

In this section, we introduce our analysis approach. Since our dataset contains multiple sets of (near) identical content, we are able to apply a range of techniques, both on individual clone sets and on the overall set of videos across all clone sets. When using the overall set, we can then choose to take the content (clone set id) into consid-

eration or not. This allows us to identify factors impacting video popularity, as well as evaluate the errors of other methods that do not take into account the impacts of differing video contents. Specifically, we focus on the following:

- **Individual clone set statistics**: Calculated for each clone set. We present results for an example clone set and summary statistics across all clone sets.
- **Content-based statistics**: These are calculated across all videos using an extended model that takes into account each video's clone set identity.
- **Aggregate video statistics**: These are calculated across all videos, ignoring clone set identity. These statistics are used for comparison.

Our analysis utilizes statistical techniques such as multivariate linear regression, collinearity analysis and principle component analysis, as well as hypothesis testing when applicable.

Several techniques used in the following assume a linear relationship between variables and normally distributed errors. To validate these assumptions, we first performed a univariate linear regression to examine the relationship between the response variable (weekly view count) and each other variable. Secondly, we examine the residual plots and corresponding tests to check that the conditions for using linear regression are satisfied.

To ensure linearity with regards to the weekly view count, some variables require log transformation. In addition, some other variables clearly are weak predictors, with higher variation in their residuals. To avoid introducing subjective biases, we did not remove such variables. Instead, we allow the analysis to help us identify suitable candidates. This turned out to be important as some variables are weak predictors on their own, but complement other variables well. The resulting variables used in the remainder of this thesis and any transformations used are summarized in Table 3.1.

- **Principal component analysis (PCA)**

  The relationships between variables can often be characterized using PCA [3]. This technique allows us to identify groups of variables (called principal components (PCs)), which explain different parts of the variations in the future popularity. The resulting PCs can be used in following analyses such as regression. A principal component is defined as a linear combination of optimally weighted observed variables (weighted such that the resulting components explain a maximum amount of variance in the data set). To calculate scores on PCs generated using PCA, the following formula is used:

$$PC_1 = \beta_{11}(X_1) + \beta_{12}(X_2) + ...\beta_{1p}(X_p), where:$$

  - $PC_1$= the subject's score on principal component 1 (the first component extracted)
  - $\beta_{1p}$= the regression coefficient (or weight) for observed variable p, as used in creating principal component 1
  - $X_p$= the subject's score on observed variable p

  When a variable is given a great deal of weight in constructing a principal component, we say that the variable loads on that component. The regression weights (loadings) are determined using a type of equation called an eigenequation.

- **Correlation and collinearity analysis**

  Interrelated explanatory variables can have negative effects on regression results. It is therefore important to detect and understand these relationships. For this purpose, we first perform a preliminary analysis to discover any correlations between the predictors themselves, and if there are groups of variables that

provide redundant information and/or explain the same variation. We leverage a number of different statistical techniques, including relative importance of predictors, pair-wise correlation matrices and auxiliary regression.

We investigate the strength of the linear relationships among the variables using Pearson's correlation. We assess the relative importance of each predictors in linear regression. We compute the univariate coefficient of determination $R_i^2$; i.e., how much the variable explains on its own. We then use the "LMG" metric [46] to decompose $R^2$ into contributions that sum to the total $R^2$. When using the LMG method, the $R^2$ contribution is averaged over orderings among regressors.

In addition, we use collinearity analysis techniques to check if there are linear relationships among the set of explanatory variables. To find out which predictor $X_i$ is a linear combination of other predictors, we run auxiliary regressions. We determine the coefficient of determination $R_i^2$ of how well the remaining explanatory variables $X_{j\neq i}$ explains $X_i$.

- **Multi-linear Regression with variable selection**

  When many explanatory variables captures the same effect, it is desirable to reduce the number of explanatory variables. Using multi-linear regression with variable selection techniques, we identify a subset of the variables that captures the majority of the variations and eliminate variables that does not provide much information regarding future popularity.

  In this section, we describe how we use multi-linear regression to determine which factors most influence video popularity. For this purpose, we define the response variable as the weekly view count (difference in view count between our two data collections), and the measured factors (also called predictors) as

all the other variables. The use of linear regression is motivated by the observed linear relationships between the measured predictors and the response variable.

We perform three types of multi-linear analysis. We first use the standard multi-linear regression model

$$Y_i = \beta_0 + \sum_{p=1}^{P} X_{i,p}\beta_p + \epsilon_i,$$

where the response variable $Y_i$ is modeled as a linear function of the independent variables ($X_{i,p}$), and the method of least squares is used to estimate the coefficients $\beta_p$ for the $P$ predictors. *Individual clone set statistics* are obtained by applying the above model on each clone set independently; this allows us to determine which factors are the best predictors for each clone set. We then apply this model on all videos together, regardless of the clone set identity, to obtain *aggregate clone set statistics*. This allows us to evaluate the error when not using our content-aware approach as discussed below.

In order to obtain *content-based statistics*, we design an extended model that incorporates a categorical variable for the clone set identity. This model is useful in understanding the influence of individual clone sets on the regression, and whether or not the classification makes a difference. Assuming that we have $K$ clone sets (or categories), we introduce $K-1$ additional category variables, each capturing the relative difference against a reference clone set. The extended multi-linear model is then given as:

$$Y_i = \beta_0 + \sum_{p=1}^{P} X_{i,p}\beta_p + \sum_{k=2}^{K} Z_{i,k}\gamma_k + \epsilon_i,$$

where $K$ is the number of clone sets; $P$ is the number of predictors; and $Z_{i,k}$ is the category regressor, encoded as $Z_{i,k} = 1$ if clone $i$ is from clone set $k$, and as

0 otherwise. Note that $\gamma_k$ can be interpreted as the relative distance between the regression lines of clone sets 1 and $k$, or in other words, a measure of their relative popularity.

- **Hypothesis testing**

  Throughout the paper we apply standard hypothesis testing techniques to assess the significance of our observations. The problem of statistical hypothesis testing may be stated simply as follows: Is a given observation or finding compatible with some stated hypothesis or not? In the language of statistics, the stated hypothesis is known as the null hypothesis and is denoted by the symbol H0. The null hypothesis is usually tested against an alternative hypothesis denoted by H1. When we reject the null hypothesis, we say that our finding is statistically significant. On the other hand, when we do not reject the null hypothesis, we say that our finding is not statistically significant. Many of these tests are made possible by the large number of large clone set that we identified, and would not be possible with previously reported datasets.

## 3.2 Factors and Their Importance in Prediction

In this section, we present our factor strength and importance analysis and results. We first describe some preliminary analysis which qualitatively explains the variation in weekly view count from the variables, then present our regression model, variable selection, and reduced models.

### 3.2.1 Preliminary Analysis

Before looking at which factors best capture the future popularity, we perform a preliminary analysis to discover any correlations between the factors themselves, and if there are groups of variables that provide redundant information and/or explain the

same variation. First, we we perform a thorough correlation and collinearity analysis: We investigate the strength of the linear relationships among the variables using Pearson's correlation, we assess the relative importance of each predictor in linear regression, and we apply collinearity analysis techniques to check if there are linear relationships among the set of explanatory variables. Second, we apply Principal Component Analysis (PCA) on each of the individual clone sets to take a closer look at the interdependence between the predictor variables.

### Correlation and collinearity analysis

To examine the strength of the linear relationships among the variables we use Pearson's correlation. Figure 3.3 shows the correlation matrix plot for an example clone set. The variables in the matrix plot are ordered based on their correlation with the response variable, and each entry shows the pairwise correlations between the corresponding two variables. The correlation's magnitude is represented by the ellipse symbol, and its sign is represented by colors (and slope), with red (with slope to the left) used for negative values and blue (with slope to the right) for positive values. We note that many of the variables have high pairwise correlation and very similar clustering of the pairwise correlation for most clone sets. In particular, two sets can be identified: (i) the set of variables related to the past video popularity (i.e., the total view count, favourite count, comment count, ratings count, likes and dislikes), and (ii) the set of variables related to the uploader characteristics (e.g., the number of uploader followers, contacts, videos, and views).

Relative importance is the evaluation of an individual predictor's contribution to a multiple regression model. To assess the relative importance of each predictor in linear regression we first compute the univariate coefficient of determination $R_i^2$; i.e., how much the variable explains on its own. We then use the "LMG" metric [46] to decompose $R^2$ into contributions that sum to the total $R^2$. When using the LMG

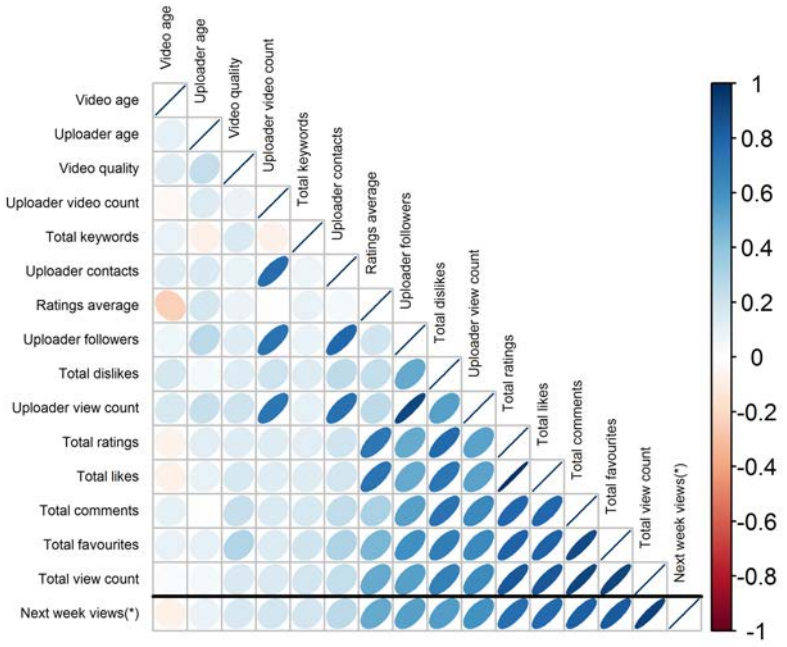
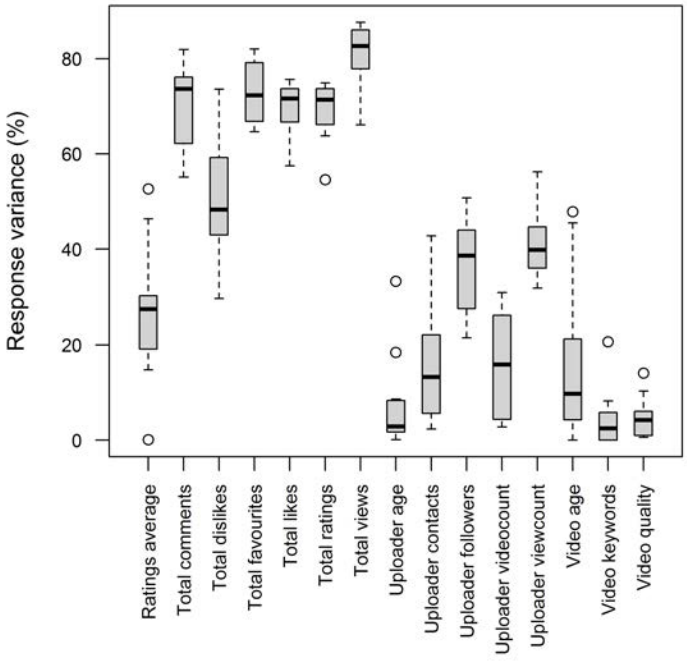


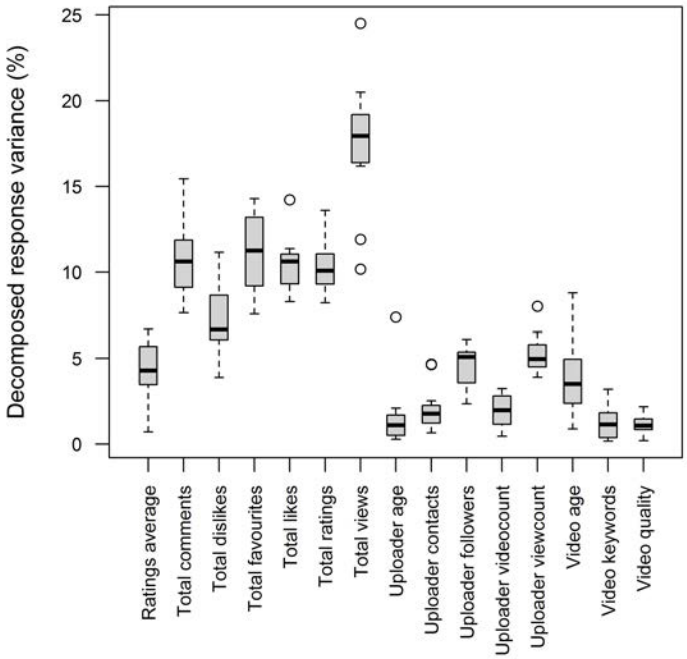

Figure 3.3: Correlation matrix for clone set 41.

method, the $R^2$ contribution is averaged over orderings among regressors. Figure 3.4 shows the boxplots of the relative importance of regressors, using both methods, across the aggregate of all clone sets. One can see that the total view count is the strongest predictor, explaining more than 80% of the response variance on its own.

Collinearity analysis techniques can also be used to check if there are linear relationships among the set of explanatory variables. Understanding and detecting collinearity is important, as interrelated variables could have negative effects on the regression results. To find out which predictor $X_i$ is a linear combination of other predictors, we run auxiliary regressions to determine the coefficient of determination $R_i^2$ of how well the remaining explanatory variables $X_{j\neq i}$ explain $X_i$. The auxiliary regressions on all individual clone sets and on all videos aggregated across clone sets show that the $R_i^2$ values of the following regressor factors exceed the overall $R^2$ value of the model including all the explanatory variables: the total view count, the number of times a video was favourited, the number of comments, the number of ratings, the




(a) The univariate coefficient of determination $R_i^2$




(b) The $R_i^2$ contributions that sum to the total $R^2$.

Figure 3.4: Relative importance of predictors.

number of times the video was "liked" or "disliked", as well as the total number of views to all videos uploaded by the uploader, and the count of the uploader's fol-

lowers. We note that these factors fall into the two previously identified groups of correlated variables. Overall, these results provide evidence of a serious collinearity and its sources.

### Principal Component Analysis

The first step when performing PCA is *extracting all the original components* and their regression weights (loadings) through the eigendecomposition for the correlation (or covariance) matrix of observable variables. The PCs are determined such that the first component is responsible for the highest fraction of total variance in the data set, and the second component is responsible for the highest fraction of variance that was not explained by the previous component(s). Every additional PC is responsible for increasingly less and less significant amounts of variance. The analysis proceeds till the total variance has been accounted for, but only the first few important components are held for rotation and interpretation. To *find the number PCs to retain*, we use several standard criteria: the eigenvalue-one criterion, the cumulative variance accounted for criterion, as well as the scree test. Our results show that typically two to four components account for an important fraction of the variance in our dataset.

Next, we *interpret the meaning of each retained component* by examining the weight (or factor loading) of the observed variables, which is equivalent, in an orthogonal analysis, to the correlations between the component and the variables. We find the two aforementioned variable groups, (i) the set of variables related to the past video popularity, and (ii) the set of variables related to the uploader characteristics), to correspond to the two primary principal components, particularly for younger clone sets. Figure 3.5 shows the scatterplot of factor loadings for the two first principal components for one such clone set. The variables giving a significant weight (higher than 0.5) in constructing principal components 1 and 2 are marked by

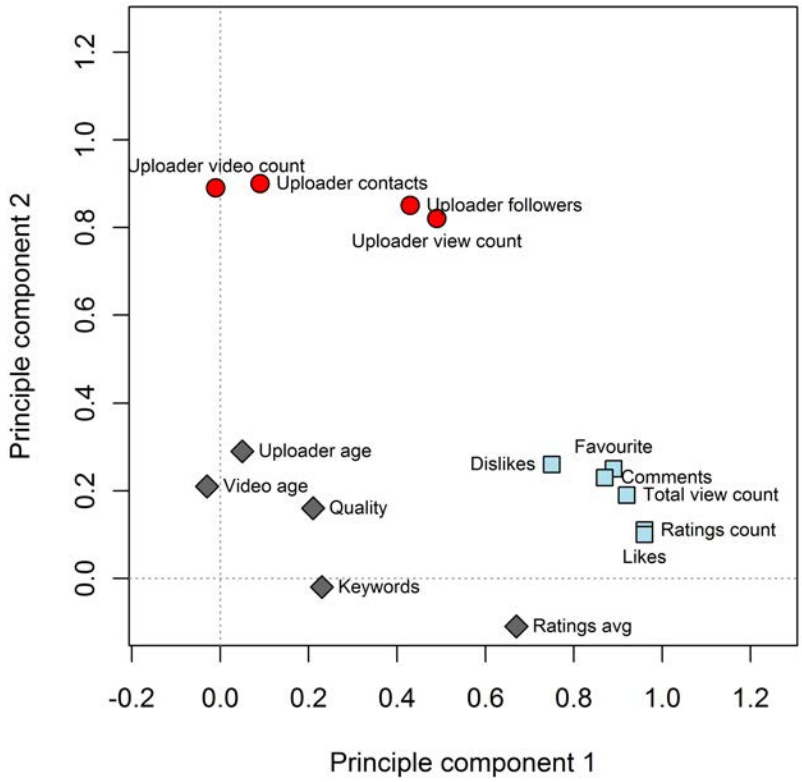



Figure 3.5: Principal components plot for clone set 41.

blue squares and red circles respectively. Referring to Table 3.1, these roughly refer to *video popularity* and *uploader popularity* metrics, respectively. For many other clone sets, particularly clone sets with big variation in video age, other *video characteristics* (such as video age and video quality) forms a third important component.

In summary, the results in this section indicate that there are many explanatory variables in our dataset that are responsible for the same variation in the total view count. In the subsequent sections we proceed to variable selection, to prune unnecessary variables and limit the negative impact of collinearity in the regression.

### 3.2.2 Variable Selection within Clone Sets

To determine which predictor variables have the most impact on the popularity of a clone within a clone set, we applied multivariate regression analysis on individual

clone sets. We use variable selection techniques to identify a good subset of variables that explains most of the variations, and to help eliminate redundant variables.

### Multivariate Regression

This section presents our results for variable selection. Table 3.2 summarizes the results for an example clone set, and presents summary statistics across all clone sets.

We used the standard t-test to assess the significance of individual regression coefficients, and test the hypothesis that the true value of the coefficient is non-zero. A small p-value (0.05 or less) is used to reject the null-hypothesis that the parameter is zero, in favor of the alternative hypothesis which suggests that the parameter in fact is of value to the model. These probabilities are summarized in the p-value column. Significance codes are added to emphasize the importance of predictors, with p-values of (0.001, 0.01, 0.05) represented by (***, **,*) respectively.

We used the F-test to assess the significance of the model as a whole by testing whether the dependent variable (i.e., the weekly view count) has a linear relationship with at least one of the explanatory variables. The large F-value and the corresponding small p-value indicate that there clearly are strong linear relationships, validating our previous observations.

Finally, we calculate the coefficient of determination ($R^2$), which summarizes the portion of the variation in the response variable explained by the fitted model. If a model has perfect predictability, $R^2 = 1$. If a model has no predictive ability, $R^2 = 0$. Note that both multiple $R^2$ and adjusted-$R^2$ suggest a strong predictive relationship between the model and the weekly view count.[3]

The results from the multivariate regression analysis, as shown for an example clone set in Table 3.2, indicate a significant predictive ability with high coefficient of

[3]The adjusted-$R^2$ is an $R^2$-like measure, which penalizes the excess number of regressors that do not add to the explanatory power of the regression. Unlike $R^2$, adjusted-$R^2$ does not increase unless the new variables have additional predictive capability.

| | Clone set 41 | | Median all clone sets |
|---|---|---|---|
| Predictor | $\beta_i$ ($\sigma_i$) | p-value | p-value |
| (Intercept) | -4.467 (3.098) | 0.157 | 0.336 |
| Ratings average | 0.119 (0.230) | 0.607 | 0.449 |
| Total comments | -0.063 (0.218) | 0.775 | 0.603 |
| Total dislikes | -0.024 (0.242) | 0.920 | 0.493 |
| Total favourites | -0.100 (0.211) | 0.638 | 0.416 |
| Total likes | 0.501 (0.497) | 0.320 | 0.533 |
| Total ratings | -0.730 (0.572) | 0.210 | 0.543 |
| Total views | 1.456 (0.236) | 0.000 *** | 0.003 ** |
| Uploader age | 0.333 (0.596) | 0.580 | 0.485 |
| Uploader contacts | 0.050 (0.119) | 0.677 | 0.500 |
| Uploader followers | 0.232 (0.197) | 0.248 | 0.361 |
| Uploader video count | -0.121 (0.277) | 0.663 | 0.522 |
| Uploader view count | -0.137 (0.206) | 0.511 | 0.511 |
| Video age | -0.590 (0.424) | 0.172 | 0.044 * |
| Video keywords | 0.036 (0.242) | 0.884 | 0.552 |
| Video quality | 0.080 (0.183) | 0.666 | 0.498 |

Table 3.2: Summary of multivariate regression results. Clone set 41 have 40 degrees of freedom in comparison to the median clone set which have 13.5 degrees of freedom. Other summary statistics include the residual standard error (1.34 and 1.01), the F-statistic (22.78 and 18.83) using 15 (15) variables, the overall p-value ($5.9 \cdot 10^{-15}$ and $2.1 \cdot 10^{-6}$), the multiple $R^2$ (0.895 and 0.936) and the adjusted-$R^2$ (0.856 and 0.876).

determination values; however, most explanatory variables are statistically insignificant and have high standard errors. In addition, some coefficients have the opposite sign to what would be suggested if using the univariate regression, with one variable at a time. For example, the full model suggests that the coefficients for the number of comments and the number of ratings should be negative. Yet, Figure 3.3 shows that both these variables have strong linear correlation with the predicted variable. These observations are classic symptoms of collinearity. To limit the impact of this collinearity, we next proceed to variable reduction, to eliminate redundant variables and select the best model.

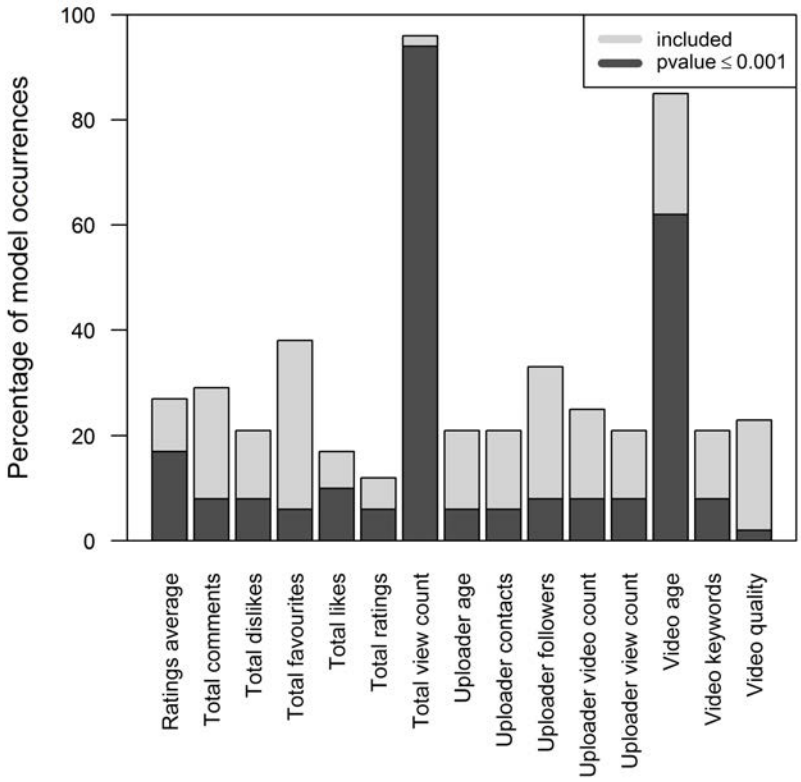



Figure 3.6: Percentage of occurrences in the set of "best models", using the best subset approach with Mallow's $C_p$. Dark color shows fraction of models in which the variable was selected while having a p-value smaller than 0.001 in the final model. In the remaining occurrences the variable was selected, but with a higher p-value.

### Reduced models

The elimination of redundant and unnecessary predictors is crucial to minimize the impact of collinearity and additional noise on the regression models. Variable selection includes a search strategy and a selection criterion or benchmark to compare two models. The two commonly used search strategies are stepwise regression and best subsets regression. Stepwise regression is a greedy step-by-step algorithm that sequentially selects explanatory variables to be included or excluded in the model based on a given criterion. This method proceeds either by forward selection, where one explanatory variable is introduced at a time, or backward selection, where all the explanatory variables are included using multivariate regression and one vari-

able is eliminated at a time. Another approach in stepwise regression is to combine forward and backward selection and test at each step for variables to be added or dropped. The best subsets regression method inspects all the possible models that could be generated from the combination of potential predictors and finds the best model based on a specific criterion. Several criteria could be used with either search method to select the best model. We examined the following: (a) adjusted $R^2$, (b) Akaike information criterion (AIC), and (c) Mallow's Cp criterion. The goal of these criteria is to minimize the residual sum of squares (RSS) while imposing a penalty for adding more regressors. Thus when selecting a model, there is a tradeoff between the model complexity and its goodness of fit.

We proceed to eliminate redundant variables using the best subset search technique and Mallow's $C_p$ as the selection criterion [5]. We have obtained qualitatively similar results using the other commonly used search methods and selection criteria. The results on Figure 3.6 show that the total view count is the most important explanatory variable. It is selected in 92% of the total set of "best models", and is determined to be highly significant. The video age is the second most important predictor, being very significant and having the second most common variable in the models. While the video age did not appear a good predictor on its own, as exemplified by the ordering in Figure 3.3 and low individual $R^2$ values (with a median of 0.081), its frequent inclusion indicates that it accounts for different variations than the total view count. This has also been observed in some of our PCA analyses. It is also interesting to note that other independent variables, such as variables related to the uploader characteristics, did not appear important in the original regression are now significant when selected in the final model, and are often significant for younger clone sets. One factor that seldomly is significant (even when included) is the video quality. In part, this may be a consequence of use of default encodings. However, although our analysis does not find a significant linear relationship, we believe that

quality differences may be important in clone sets with wide variations in video age and associated wide variations in quality. The quality variations in such cases may play a role in making age our second most important predictor.

From Figure 3.6, we can see that the best subset approach with Mallow's $C_p$, on average, reduces the number of variables by about 60%. The multiple $R^2$ values for the chosen models are then only slightly smaller than the original $R^2$ value of the full model.

### 3.2.3 Summary

This section analysed the importance of different variables in explaining video popularity. We first established that several variables are redundant and can be eliminated. We find the most significant variables to be total views and video age. Other factors include the uploader popularity (as measured by the number of its followers). An interesting observation is that the most influential factors also are the only statistics available to the YouTube users, when searching for a video.

In the following sections we take a closer look at the impact of clone identity on the video popularity over time, and how the popularity is influenced by other external factors such as external links/embeds, the video being featured, or more accessible through searches.

## 3.3 Impact of Content Identity

The regression analysis presented in the preceding section was applied on individual clone sets. Using the model extension presented in Section 3.1.2, we perform regression analysis over the entire dataset while taking into account the content identity, and thus by extension study the impact of the video content on popularity dynam-

| | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Total view count ($\beta_i$) | 1.100 | 0.013 | 87.83 | 0.000 *** |
| Video age ($\beta_i$) | -1.008 | 0.039 | -25.80 | 0.000 *** |
| Clone set ($\min_k \gamma_k$) | -0.727 | 0.348 | -2.08 | 0.037 * |
| Clone set ($\max_k \gamma_k$) | 2.802 | 0.345 | 8.08 | 0.000 *** |

Table 3.3: Summary of extended regression analysis using categorical variables for clone set identification. With 95% confidence, the rejection rate of the hypothesis that the category variables ($\gamma_k$) are equal to zero is 94%.

ics. Evaluating the importance of the clone set categorical variable is important as it allows us to separate the impact of content-related and content-agnostic factors.

We perform the content-based regression analysis using the most important explanatory variables identified in Section 3.2.2. We use our default clone set ordering, where sets are numbered from 1 to 48, and choose one clone set as the baseline set.

Summary results are presented in Table 3.3, for the baseline clone set number 1. The coefficients of the category variables ($\gamma_k$) explain by how much the intercept of the selected clone set differs from the intercept of the baseline clone set. The significance of the categorization, i.e. the impact of video content, is then measured by the corresponding p-values. We also report range values $\min_k \gamma_k$ and $\max_k \gamma_k$, across the 47 non-baseline clone sets.

We find that 44 out of 47 category variables have p-value smaller than 0.05. When averaging over all possible baseline clone sets, we found approximately 60% of the category variables to be significant. This illustrates the importance of taking clone identity into consideration.

As a second step to evaluate the importance of video content, we compare the regression analysis results of the content-aware extended model and the regular individual clone set models with the aggregate model which ignores clone identity.

For each model type, we used four different models: three partial models and one full model. The first partial model includes only the view count variable, the second

model includes both the view count and the video age, and the third model further adds the uploader followers.

Table 3.4 shows the coefficient of determination $R^2$ values for each model when running the regression analysis on each clone set individually ("Individual"), across all clones and clone sets as an aggregate ("Aggregate"), and when we take the clone identity into account using category variables ("Content-based"). Comparing the last two rows, we note that the 'Content-based' models consistently explain a larger portion of the variation, as evidenced by higher $R^2$ values. This is another indication that taking into account the clone identity is an important factor in modeling the video popularity.

| | View count (1 variable) | + age (2 var.) | + followers (3 var.) | all (15 var.) |
|---|---|---|---|---|
| Individual (mean) | 0.788 | 0.864 | 0.871 | 0.933 |
| Individual (median) | 0.803 | 0.873 | 0.875 | 0.940 |
| Individual (41) | 0.861 | 0.870 | 0.874 | 0.895 |
| Content-based | 0.792 | 0.850 | 0.852 | 0.855 |
| Aggregate | 0.707 | 0.808 | 0.808 | 0.821 |

Table 3.4: Summary of $R^2$ values for example models.

Table 3.4 also reveals that the view count by itself explains the biggest percentage of the variance, especially when taking into account the clone identity. Adding the video age variable increases the $R^2$ values relatively significantly. Adding the uploader followers variable can result in an occasional incremental increase in the goodness of fit, while the other variables impact is even less important.

However, perhaps more importantly, this table also shows that if one tried to analyze the relative importance of age, followers, etc., without controlling for video content, one would conclude that factors such as age and followers are relatively more important (compared to view count) than they really are. This is illustrated by comparing the difference in values from left to right, for the aggregate and the

content-based models. In one case, $R^2$ is improved by 0.114, and in the other by only 0.063.

The next section will take a closer look at the impact clone identity may have on predictive models, such as the rich-get-richer model.

## 3.4 A Closer Look at Preferential Attachment

Prior works have suggested that video popularity evolves according to rich-gets-richer preferential selection [10] or a variant thereof (e.g., [60, 19]), wherein the current viewing rate of a video is proportional to the total number of views the video has already acquired. In Section 3.4.1, we evaluate whether or not our data is consistent with a rich-get-richer model of popularity evolution. Section 3.4.2 considers a more restricted form of rich-get-richer behavior, the "first mover" advantage. Finally, Section 3.4.3 explores other phenomena that may result in rich-get-richer behavior, including search bias towards popular videos.

### 3.4.1 Models

We consider rich-get-richer models wherein the probability $\Pi(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law

$$\Pi(v_i) \propto v^{\alpha},$$

where $\alpha$ is the power exponent.

Perhaps the most interesting case is when $\alpha = 1$, which was considered by Barabasi and Albert [10]; this corresponds to linear preferential selection, and can be shown to result in a scale-free distribution (in our context, of total view counts). For the more general non-linear case, the underlying distribution depends strongly on the parameter $\alpha$ and the scale-free characteristic no longer holds [42]. The sub-linear case

| | Slope estimate | Confidence intervals | | Hypothesis testing | | |
|---|---|---|---|---|---|---|
| Metric | $\alpha$ ($\sigma$) | 90% | 95% | H0: $\alpha = 1$ | H0: $\alpha \geq 1$ | H0: $\alpha \leq 1$ |
| Individual | 1.027 (0.091) | 0.988-1.065 | 0.981-1.073 | 0.85 | 0.57 | 0.43 |
| Content-based | 1.003 (0.014) | 0.98-1.027 | 0.976-1.031 | 0.81 | 0.59 | 0.40 |
| Aggregate | 0.932 (0.016) | 0.906-0.958 | 0.901-0.963 | **REJECTED** | **REJECTED** | 1.00 |

Table 3.5: Rich-get-richer slope estimates and hypothesis testing.

($\alpha < 1$) results in a stretched exponential distribution. In the super-linear case ($\alpha > 1$), the rich get much richer, and when $\alpha > 2$ a winner-takes-all phenomenon quickly occurs. We note that these cases can be compared against a bank giving differentiated interest rates on their customers savings. The linear case simply corresponds to the case where everybody gets the same interest, but whoever has more money naturally will gain more money than those with less money. Similarly, the sub-linear case corresponds to the case where people with less money will be given a higher interest rate, such as to help them catch up with the richer people, and of course, as suggested by the winner-takes-all phenomenon, for example, the super-linear case have opposite effect.

Basic rich-get-richer models as described above consider only the number of accumulated views as a determinant of the rate of acquiring additional views. With YouTube videos, however, user interest in particular subjects changes over time, causing a deviation from rich-get-richer behavior when one considers a collection of such videos with differing contents. An important question is whether a rich-get-richer model is applicable when one removes the impact of changing user interests, as we are able to do with our clone-based methodology.

To answer this question, we first identify within each clone set videos of similar "generation" (age within a multi-year window). We restrict attention to videos of similar generation to avoid our analysis being impacted by wide variations in video quality (or other generation-related effects). Specifically, for each clone set, we first find the video clone with the highest current popularity (i.e., the video that acquired the most additional views during our one week measurement window). We then

consider only the videos in the clone set that were uploaded within two years of the upload time of this video.

We now take a closer look at the impact differences in video identity can have on the rich-get-richer phenomena. We examine how the rate at which videos attain new views depends on the total view count using univariate linear regression (using log-transformed data). All three analysis approaches, namely regression analysis on individual clone sets, on the aggregate, and the aggregate considering content identity, were applied. Using hypothesis testing, we determine if the system is sub-linear ($\alpha < 1$), linear ($\alpha = 1$), or super-linear ($\alpha > 1$). Table 3.5 summarizes our results.

The first column in Table 3.5 shows the coefficient estimates and standard deviation resulting from the univariate regression analysis. The second and third columns show the corresponding confidence intervals. The results indicate that the slope estimates are often around one and that the rate at which videos acquire views is correlated with their current total view count, providing quantitative evidence for the existence of preferential selection. For the individual clone sets, and the extended content-based model, $\alpha$ is typically equal or slightly higher than one. The selection rate is linearly dependent on the current total view count, proposing that the popularity evolution is scale free, and strongly controlled by rich-get-richer behavior. For the aggregate model, $\alpha$ is less than one, indicating an exponential popularity evolution that could result in a much more even popularity distribution than that suggested by the pure (linear) rich-get-richer dynamics.

To validate these observations, we performed hypothesis testing to check whether each slope estimate is linear, super-linear, or sub-linear. The last three columns in Table 3.5 show the p-values for hypothesis tests that $\alpha$ is equal to 1 (H0), $\alpha$ is greater than 1 (H1), $\alpha$ is less than 1 (H3). Here, the p-values express the results of the hypothesis test as a significance level. If p-values are smaller than 0.05, the hypothesis is rejected. Using standard hypothesis testing, we remind the reader that

these tests can only determine if the null hypothesis can be rejected, in favor of the alternative hypothesis, not if the null hypothesis is true. However, note that if two out of three (mutually exclusive) hypothesis are rejected, this would suggests that the third hypothesis likely is true.

From the results in Table 3.5, we note that the hypothesis testing in fact validates many of the observations based on our discussion of the slope estimates. In particular, it is clear that a model that does not take the clone set identity into account (e.g., the aggregate model) may suggest a much weaker sub-linear relationship than if the clone identity is taken into consideration. Note that for the aggregate model, we can reject both the linear and super-linear hypotheses, which clearly suggests sub-linear preferential selection. Further, the extended model that takes into account the clone set identity does not allow us to reject any hypothesis, but the values suggest that a linear relationship is plausible. The results are interesting as they provide evidence for linear (and in some cases even super-linear) preferential selection once we control the individual heterogeneity in content.

### 3.4.2 First Mover Advantage

Rich-get-richer behavior may result in part from a "first mover" advantage. The first video to include particular content may have already achieved significant dissemination by the time that clones appear, causing it to acquire new views at a higher rate (for example, via recommendations from previous viewers, featuring, or bias in search algorithms). Using our clone-based methodology, we now evaluate the advantage of being the first to upload particular content.

To track video popularity over time, we use YouTube's insight data collected through HTML scraping. As a first step, we consider the success of the first mover in each cloneset, where a success event for a particular video is defined as when that video accumulates the larger number of total views compared to all other videos

| | 1st | 2nd | 3rd | 4th | 5th | later |
|---|---|---|---|---|---|---|
| Winner uploaded | 27.1 | 12.5 | 8.3 | 6.3 | 6.3 | 39.6 |
| Winner searched | 66.7 | 8.3 | 0.0 | 8.3 | 8.3 | 8.3 |

Table 3.6: The percentage of times a video clone that obtained the highest total view count was the first, second, third, fourth, fifth (or later) among the videos in the clone set with respect to being uploaded or searched. (Clone sets with relevant statistics considered.)

within the clone set. We first consider how often the most successful video within a clone set is the first to either be uploaded or discovered through search. Table 3.6 shows the number of times the video clone that obtained the highest total view count was first, second, third, fourth, or fifth, among the videos in the clone set, to be uploaded or found through search. Overall, the winner was uploaded first among the videos in the clone set in 27.1% of the observed cases, and was among the first five in 60.4% of the cases. Similarly, the winner is the first to be found through search in 66.7% of the cases for which we have (insight data) statistics, and among the first five to be found through search in 92% of the cases. Clearly, there is a significant advantage to the first mover.

While the first mover is not always the winner, it is often highly successful, even when it is not the winner. To illustrate this, we consider the view count of the first uploaded video relative to the winner. A ratio of one corresponds to the case where the first mover is the winner, and a small ratio indicates that the first mover was relatively unsuccessful. Figure 3.7 shows the complementary cumulative distribution function (CCDF) of this ratio for our dataset. In addition to the 27% of clones sets where the first mover was the winner, this ratio is above 0.5 for 35% and above 0.1 for 50% of the first-movers. Given the high skew in overall view counts, these values are high.

While these results suggest that the first mover typically is relatively successful, it is interesting to note that there are cases where other videos have been able to surpass the first mover in popularity. What is it that allows some other video to overtake the

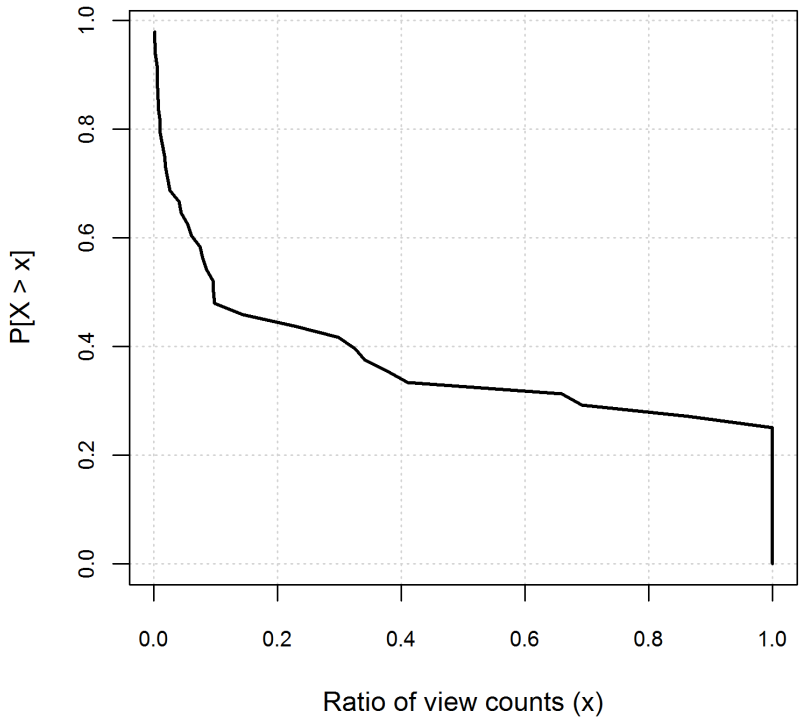



Figure 3.7: CCDF of the ratio of the view count of the first uploaded video in a clone set, relative to the view count of the video with the highest view count in the same set.

spot as the most popular clone? Section 3.5 takes a closer look at some influences that can cause such overtakings.

### 3.4.3 Video Discovery and Featuring

We now examine the roles that video discovery and featuring mechanisms may play in the observed rich-get-richer preferential selection behavior. Aspects such as featuring on YouTube, ranking of a video in YouTube search, and embedding of a video on external sites, are difficult to capture over time. Nonetheless, the "video referrers" part of the YouTube insight data provides (for some videos) additional information necessary for our analyses. The results presented here are based on analysis of clone sets that have multiple videos with insight data. We use YouTube's classification of registered referrers (Table 3.7).

| Referrer type | Description |
|---|---|
| Ad | The viewer was referred to the video through a paid Youtube promotion. Videos can be promoted on the YouTube website through paid advertisement. Such videos are labeled as "Promoted Videos" and appear next to related search results. |
| Featured | The viewer was referred to the video through an unpaid Youtube promotion such as the YouTube "Featured Videos" or "Spotlight Videos" sections. It's interesting to note that most of the videos in the "Featured Videos" list are selected from the ones uploaded by Youtube partners due to commercial advantages. |
| Mobile | The video views occurred on a mobile device through the mobile version of the Youtube website (m.youtube.com) or through Youtube apps. |
| Google Search | The viewer was referred to the video through keyword searches on the Google search engine. |
| YouTube Search | The viewer was referred to the video through keyword searches on YouTube. |
| Related | The viewer was referred to the video through related videos in YouTube. |
| Embedded | The viewer was referred to the video through an embed on an external website. |
| First embedded view | The video was embedded on another website when it was viewed. |
| External | The viewer was referred to the video through links on other websites. |
| Other/Viral | YouTube could not recognize a referrer for the views because the the user navigated directly to the video by copying and pasting the video's URL or by clicking on a link to the video from an email or instant message application. |
| Youtube other (Internal) | The viewer was referred to the video through a Youtube link other than a related video or search result. Other pages on YouTube could be the YouTube homepage, category pages, a user profile page, other peoples channel pages and a user generated playlist. |
| Subscriber | The video views occurred as a result of the uploader channel's subscribers clicking on it in one of the subscription notification modules. |
| Channel | The video views occurred on the uploader's channel page. |

Table 3.7: YouTube's classification of registered referrers

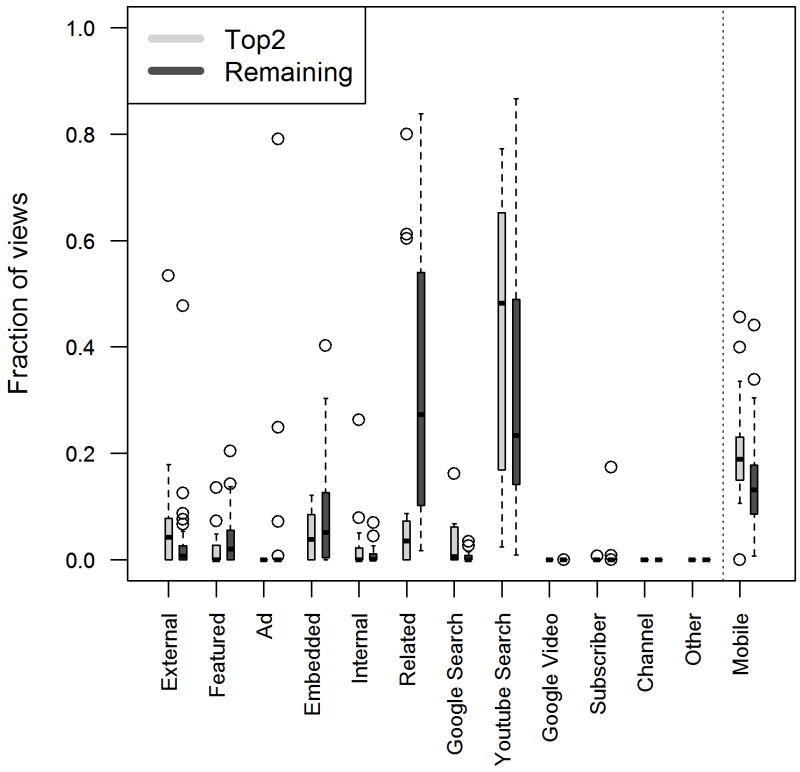



Figure 3.8: Boxplot of the average fraction of views (per cloneset) coming through different referrer categories.

We first consider how the most popular videos within each clone set have obtained their views, compared to their less popular counterparts. Figure 3.8 compares the average fraction of views coming through different referrer types for the most popular clones, with that of the remaining clones.[4]

The results are somewhat counter-intuitive. Notice that the "Top 2" most popular videos are not necessarily the videos that are prominently featured or externally linked. Instead, the search discovery method alone accounts for most of the difference. For example, for the search referrer category, the median of the top clones is almost equal to the 90th quartile of the remaining videos. On the other hand, the less successful clones get most of their views through related (video) referrals.

[4]We present the mobile referrers separately as it is not a source a discovery per se, but is nevertheless impacts discoverability, as more users are accessing videos exclusively through mobile devices

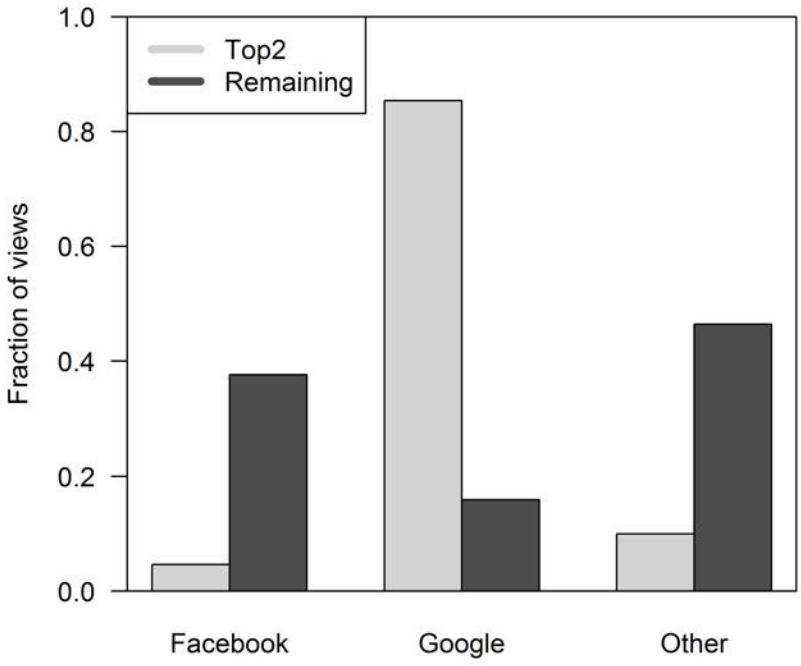



Figure 3.9: The fraction of views coming through external sources, for clones that are externally linked.

Figure 3.9 shows the fraction of views coming through external referrers only (not including embeds). Note that Google is shown as an external source of traffic and it is driving most of the external views of the popular clones. Google is considered an external referrer because views may come from a number of Google non-search services such as Google News, Google Reader, and Google Group posts.

Overall, the highest fraction of clicks to a video is coming through the search referrers. As all videos can potentially be found through search, but not all videos are featured or embedded on external websites, we take a closer look at these referrers. Figures 3.11(a) and 3.11(b) show the corresponding boxplots of the fraction of views coming through different referrer types for only the clones that are featured and externally linked, respectively. The same conclusion applies to these data subsets: search referrers are the most powerful in terms of the percentage of traffic they bring,

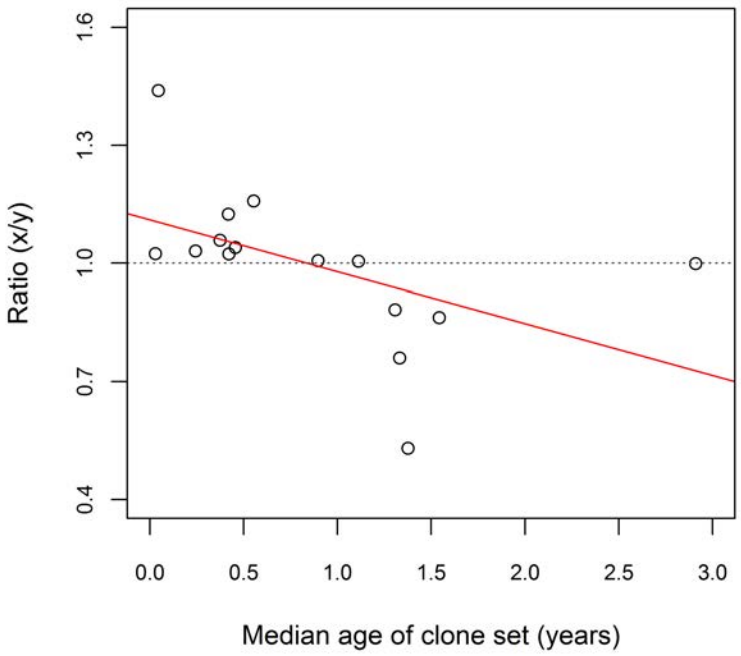
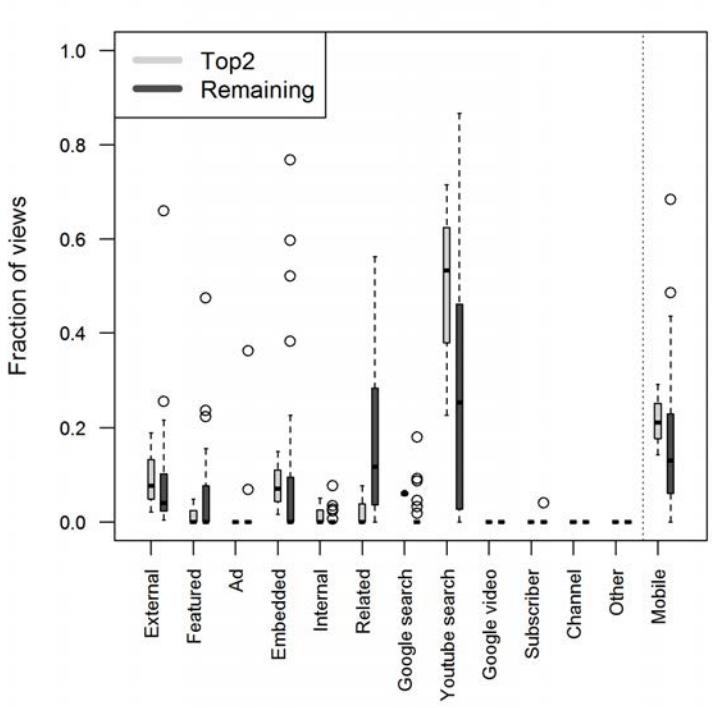


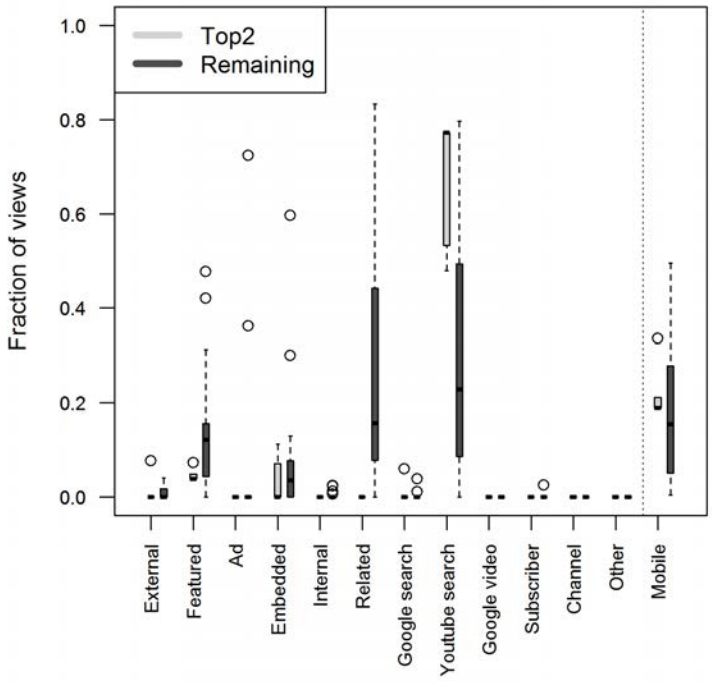

Figure 3.10: Search bias towards top 2 videos in each clone set versus the median age of the videos in each clone set.

whereas search and mobile referrers are still showing the biggest differences between successful and less successful videos.

Recall that we are considering multiple videos containing essentially the same content, and this allows us to remove biases introduced because of differences in content (e.g., popular content is more likely to be searched for than non-popular content). Our results suggest that successful videos are much more prominently selected through searches. This could potentially occur because of YouTube's internal search mechanism, the keywords associated with the videos, the keywords entered by the users, user biases when selecting among search results, or a combination thereof. For example, people may be more likely to pick the first search results than pick items lower down on the list, or to pick videos with higher view counts (visible to the user at




(a) Externally linked clones




(b) Featured clones

Figure 3.11: Boxplot of the fraction of views of clones externally linked and featured, coming through different referrer categories.

the time of selection) [58]. Again, we note that our unique dataset allow us to remove the importance of content. This is an important observation as we can eliminate the obvious bias that some videos have more views because more people searched for that content. Instead, videos with a higher fraction of views due to searches (within a clone set) must either have been more frequently returned by the search engine or more frequently selected by the user, than the other clones.

Next, we statistically test whether or not the search is providing additional bias towards more popular clones, and enabling the videos with larger view counts to attract even more views. For each clone set, we compute the fraction ($x_i$) of all search induced views that are to the top ranked videos in a clones set, and the fraction ($y_i$) of the views that are to the top ranked videos. Each ($x_i$,$y_i$) pair is an observation. If there is no additional (unproportional) bias in the search methods, the ratio $x_i/y_i$ should be bigger than one only half of the time (and less than one the other half). Figure 3.10 shows a scatter plot of this ratio against the median age of the videos in each clone set when considering the top two clones in each set. We note that 10 of the points are above the $x/y = 1$ line and only 5 are below the line. The probability for this or a more skewed observation in the case there was no bias is 0.171. This low probability suggests that the search may be biased. The figure suggests that any such bias likely is caused by the younger clone sets, for which search may be a more important factor.

We use hypothesis testing to systematically quantify the likelihood that the search mechanism is proportionally fair given the observed results. Similar to the scale-free rich-get-richer models, proportional fairness would ensure that views from search are proportional to the total view count. We formulate the null hypothesis, denoted $H_0$, as the "search mechanism is fair". Mathematically, we express the null and alternative hypothesis as:

$$H_0 : P(x_i > y_i) = 0.5$$

| | Top 1 | Top 2 | Top 3 | Top 4 |
|---|---|---|---|---|
| p-value | 0.075 | 0.171 | 0.035 | 0.285 |

Table 3.8: Hypothesis testing of whether or not the search mechanism is unproportionally biased towards the most popular clones.

$$H_a : P(x_i > y_i) \neq 0.5$$

where $p = P(x_i > y_i)$ is the probability of observations below the $x/y = 1$ line. We estimate the standard deviation and calculate the test statistic $z$, which measures how far the sample mean $p$ diverges from the expected mean of a fair search mechanism. We then compute the p-value of the test statistic and conclude whether or not the null hypothesis is rejected. We choose a 95% confidence interval, and so the level at which we reject the null hypothesis is 0.05. Table 3.8 shows the p-values for a different subsets of the most popular clones (top 1, top 2, top 3, and top 4 ranking videos). While the null hypothesis only is rejected for the top 3 case, we note that we get relatively small p-values for the other cases as well. We believe that a larger dataset would help further validate our observations.

We surmise that YouTube's search mechanism, at the time of these measurements, either was biased towards the most popular clones or in the case multiple clones are presented to the user, the users are biased towards picking the videos with higher view counts. This further strengthen the rich-get-richer phenomenon.

As previously discussed, the first mover advantage can be important for the success of a video. In addition to being uploaded, it is important that the video is discovered and/or made available through different paths. Using correlation analysis, we have observed that there is often a significant positive correlation between the total view count and the order in which clones are first referred, featured, or accessed through mobile devices. While omitted, these results suggest that there is also is a *first-discovery advantage*, where videos discovered earlier through internal search methods, featured earlier, or that is accessed through mobiles earlier, tend to be ranked higher.

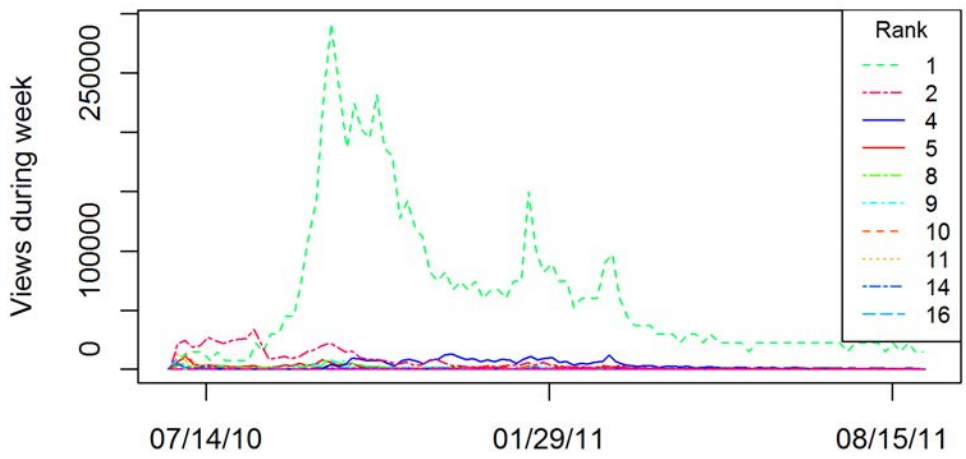



Figure 3.12: The weekly views for a number of example videos in clone set 14 (18 clones).

## 3.5 Factors Impacting Initial Popularity

This section considers factors impacting the view count early in a video's life, which in turns impacts the overall video popularity due to the rich-get-richer behavior, as shown in the previous section.

### 3.5.1 Uploader Characteristics

We analyzed the YouTube social network size of the uploaders' observed in our dataset. In general, uploaders of top-ranked videos have large social networks. Furthermore, manual examination of the top uploaders confirmed that they are often commercial entities promoting their official external websites. Usually commercial uploaders are the "first-movers". In fact, even when they were not, it appear that their videos often manage to move ahead in popularity. Figure 3.12 shows an example clone set where a commercial user (video with rank 1) catches up and surpasses a private uploader (video with rank 2) even though the former was not the first uploader.

| | Aggregate | | | | Content-based | | | |
|---|---|---|---|---|---|---|---|---|
| Predictor / Age | 1d | 3d | 7d | 14d | 1d | 3d | 7d | 14d |
| View count | 0.44 | 0.42 | 0.50 | 0.55 | 0.60 | 0.59 | 0.66 | 0.70 |
| Video quality | 0.08 | | | | 0.35 | | | |
| Video age | 0.00 | | | | 0.43 | | | |
| Number of keywords | 0.04 | | | | 0.36 | | | |
| Uploader view count | 0.41 | | | | 0.64 | | | |
| Uploader followers | 0.40 | | | | 0.58 | | | |
| Uploader contacts | 0.19 | | | | 0.42 | | | |
| Uploader video count | 0.08 | | | | 0.38 | | | |
| Uploader age | 0.02 | | | | 0.35 | | | |

Table 3.9: Age effect on $R^2$ values when taking into account the clone set identity (content-based) and when not (aggregate).

This is a typical example showing the importance of the uploader characteristics and its impact on a video's popularity.

### 3.5.2 Age-based Analysis

As seen previously in this paper, the uploader characteristics can be useful for prediction of future popularity. While the total view count can be an important factor when predicting a video's future popularity, it has the disadvantage that all videos start with a view count of zero. However, at the time of upload, some of the other predictors are known, including the uploader network and the prior success of the uploader. Indeed, based on our PCA analysis in Section 3.2 we have seen that the uploader characteristics can be a good predictor for young videos. We now perform an age-based regression analysis to determine how the relative importance of the total view count changes with time, relative to these more static factors.

Table 3.9 shows the coefficient of determination $R^2$ between the predictors in the first two weeks since a video's upload and the total view count at the half-year point since upload. We calculate the total view count of videos at 1 day, 3 days, 1 week, 2 weeks, and half a year using the historical view statistics. Linear interpolation

is needed to calculate the approximate total view count at specific time thresholds, as the data provides only 100 points, equally spaced through the video's lifetime. The file-related information and the uploader characteristics properties are assumed constant. The first four columns show results for the aggregate set of videos and the last four columns show results when clone set identity is accounted for.

These results show that the total view count quickly become the strongest predictor of the view count at the half-year point. The results also confirm that during the early stages of a video's lifetime, the uploader's social network is a more significant factor than the total view count. Indeed, already at upload, approximately 64% of the variation in views can be explained by the uploader view count alone, and it takes a week for the total view count to become a similar or better predictor than the uploader social network. The impact of the uploader characteristics are significant in the beginning, probably because an established social network is a source of initial views from subscribers, that could boost the video view count in a short duration, and let the video move forward in rankings even if it does not enjoy a first mover advantage.

Finally, we note that some factors have much more impact when the influence of the content is considered through the clone set identity factor. For example, factors such as the keyword number, the video quality and the video age, have a great impact in the early stages of a video's lifetime. The keyword metric, although appeared to be insignificant in the aggregate analysis, is an important factor when a video is first uploaded, explaining up to 36% of the variation in views. This may suggest that keywords in fact may be one of the main factors in helping find the video in the first place (when competing against videos with the same content). The more targeted keywords a video has, the greater the probability that it will be discovered after its upload.

## 3.6 Conclusion

With such a large user base, video sharing sites such as YouTube have the ability to impact opinions, thoughts, and cultures. On such sites, not all videos will reach the same popularity and have the same impact. Popularity evolution of such user-generated content is a complex process and has garnered increasing interest. Prior works have used datasets with videos that contains completely different content. These approaches fail to provide an unbiased view of the underlying factors and dynamics as they cannot distinguish the differences in popularity occurring because of content difference from other, non-content related factors.

In this chapter, we take a closer look at content-agnostic factors that impact YouTube video popularity. We design a content-aware methodology for studying, both qualitatively and quantitatively, the impact different factors have on video popularity. Using a systematic investigation of the content-agnostic factors that most influence a video's current popularity, we find that the most significant content-agnostic factors are the total number of previous views and the video age. When controlling for video content, we show that "rich-get-richer" preferential selection based on the current video popularity appears to provide a good model of popularity evolution, except for very young videos. However, when looking across different contents, the rich-get-richer behavior becomes inaccurate and significantly weaker. For young videos we find a variety of other significant factors, including uploader characteristics such as size of social network, and number of keywords.

# Chapter 4

# Popularity Dynamics Characterization and Modeling

In Chapter 3, we observed that the popularity evolution of user-generated content is goverened by a strong linear "rich-get-richer" behavior, with the total number of previous views as the most important factor. However, when not controlling for video content, we showed that the preferential selection behavior gets weaker and becomes inaccurate, and thus rich-get-richer type of models do not accurately capture the popularity dynamics. This finding motivates models that attempt to capture the popularity evolution of user-generated videos in time. In the second part of this thesis, we present a systematic study of the long term popularity evolution of videos. We perform a characterization and modeling of the popularity dynamics using only the total view count for analysis.

Examining the popularity dynamics of user-generated videos requires tracking a representative sample of videos over a period of time. Obtaining a random sample of user-generated videos from YouTube is, however, challenging because of the scale of the service, its continually-expanding catalogue of videos, and the service-specific limitations associated with discovering and tracking videos. Moreover, the possible

biases in the datasets introduced by common sampling approaches from services such as YouTube pose another challenge. In this work, we illustrate the biases resulting from the sampling approaches employed, and we use for analysis a seemingly unbiased sample of videos that we tracked on Youtube over a period of eight months.

This chapter is organized as follows. Section 4.1 describes how we collected our datasets, and presents some initial analyses concerning possible biases in the datasets owing to use of sampling techniques. Section 4.2 examines popularity dynamics and churn for our dataset of recently-uploaded videos. Section 4.3 presents our three-phase characterization of popularity evolution, and provides the underpinnings for the model proposed in this work. Section 4.4 presents the basic model, its validation, and also insights drawn from the model. An extension of the basic model is described in Section 4.5. Finally, Section 4.6 concludes the chapter.

## 4.1 Sampling Approaches and Bias

Studying the popularity dynamics of user-generated videos requires tracking a representative sample of videos over a period of time. Obtaining a random sample of user-generated videos from YouTube is, however, challenging because of the scale of the service, its continually-expanding catalogue of videos, and the service-specific limitations associated with discovering and tracking videos. Section 4.1.1 describes how we collected our datasets, including the two alternative sampling approaches used and the tracking of the sample videos over a period of eight months. Section 4.1.2 presents a high-level summary of our datasets. Section 4.1.3 describes the results of some initial analyses designed to identify possible biases in the datasets, as might result from the sampling approaches employed.

### 4.1.1 Data Collection Methods

We collected meta-data (such as number of views, ratings, and comments) on more than one million YouTube videos on a weekly basis for over eight months. The videos were selected by sampling, over a one-week period from 27 July to 2 August, 2008. A one-week sampling period was chosen to avoid potential day-of-the-week effects. We used two different sampling approaches, both based on functionality provided by the YouTube API[1], as described below:

- Sampling from the recently-uploaded videos: The API provides a call that returns details on 100 recently-uploaded videos. Using this API, we collected meta-data on approximately 29,500 videos during the one-week sampling period.

- Sampling using keyword search: The API also allows retrieval based on keyword searches; the API returns search results sorted by "relevance". We performed keyword searches using words chosen randomly from a dictionary. As search results for some words return a very large number of videos, for those returning more than 500 videos we selected only the first 500. We found approximately 1 million videos using this method during the one-week sampling period.

There are several other possible approaches to sampling videos. One approach is to sample from the "most-popular" lists. A closely-related variant is to start the sampling process from one or more videos in the "most-popular" list and subsequently follow "related videos". A detailed investigation of the biases introduced by other sampling approaches is left for future work.

During the remainder of our measurement period, specifically from 3 August 2008 to 29 March 2009, we collected meta-data for the videos identified in the sampling phase on a weekly basis. Using the timestamp at which the meta-data for a video was first captured, we ensured that subsequent measurements ("snapshots") were exactly

[1] http://code.google.com/apis/youtube/overview.html

Table 4.1: Summary of datasets.

| Dataset | Recently-uploaded | Keyword-search |
|---|---|---|
| Videos | 29,791 | 1,135,253 |
| Views (start) | 1,203,755 | 40,094,514,507 |
| Views (end) | 39,089,184 | 64,019,907,026 |

one week apart. For example, if a video was sampled on Tuesday evening, then each weekly measurement for this video was performed as close to the same time of day as possible, on Tuesday evenings, in the following weeks. This form of staggering allowed us to track a large number of videos without exceeding YouTube's query rate limitations, while enabling easier management of our own measurement resources.[2]

### 4.1.2 Summary of Datasets

A summary of our datasets is presented in Table 4.1. In total, we have 35 snapshots for each sampled video's meta-data (counting also the "seed" snapshot collected during the sampling phase), with one-week spacings between consecutive snapshots. From the total view count at each snapshot $i$ ($1 < i \leq 35$), we can determine how many times the video was viewed during the one-week period since snapshot $i-1$, which we term the "added views" at snapshot $i$. The total view count at snapshot 1 (the seed snapshot) tells us the total views acquired by the video from its upload time until the start of our data collection for that video.

During the measurement period, the 29,791 recently-uploaded videos acquired about 38 million additional views, and the 1,135,253 keyword-search videos received about 24 billion additional views. Note that the keyword-search videos acquired additional views at a higher average rate than the recently-uploaded videos. This suggests possible bias in the keyword-search dataset towards more popular videos, which is investigated further in the next section.

[2]Our datasets are available at http://www.cs.usask.ca/faculty/eager/Performance11.html.

### 4.1.3 Sampling Bias in the Datasets

One indicator of sampling bias is a skewed age distribution, where video age is defined as the time since upload of the video. Figure 4.1 shows histograms for the age at seed time (i.e., when meta-data for the video was first collected) for the recently-uploaded videos, using 6-hour bins (left plot), and for the keyword-search videos, using one-week bins (right plot). The age of the videos in the recently-uploaded dataset is approximately uniformly distributed within a week, which is consistent with the hypothesis that the YouTube API call used to obtain these videos returns randomly-selected videos at most one week old.[3] The age distribution of the keyword-search dataset videos, in contrast, shows that this dataset is far from being a random sample of (all ages of) YouTube videos. There is a strong skew towards younger videos, with a prominent spike in the distribution for the first bin corresponding to an age of at most one week. The age of the oldest video is about 38 months. One possible explanation of the observed skew is that the results returned from keyword searches are biased towards more popular videos. This hypothesis is supported by the popularity characteristics of the keyword-search and recently-uploaded videos, as described next.

Figure 4.2 shows the complementary cumulative distribution function (CCDF) of the added views at snapshots $i = 2, 8, 32$, using logarithmic scales on both axes, for both datasets. Comparing the added views of the keyword-search and the recently-uploaded videos at the same snapshot, note that the keyword-search videos receive substantially more views than the recently-uploaded videos. This is reflected, for example, by a heavier right tail for the keyword-search video curves. At each snapshot, the most (currently) popular keyword-search videos (i.e., those with the most added views) have an order of magnitude more new views than the most popular recently-

[3] We notice a dip in the histogram for videos that are approximately 96 to 108 hours old at time of collection. This dip is currently unexplained.

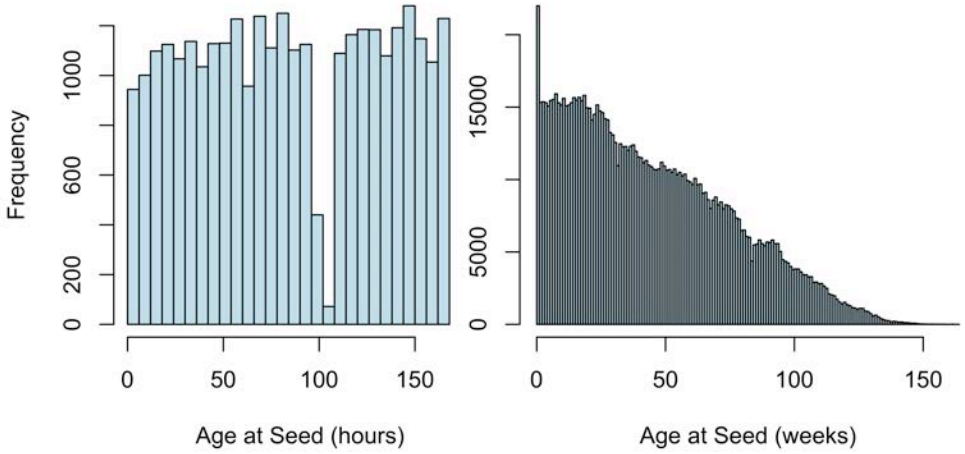



Figure 4.1: Age distribution of the videos (left: recently-uploaded; right: keyword-search).

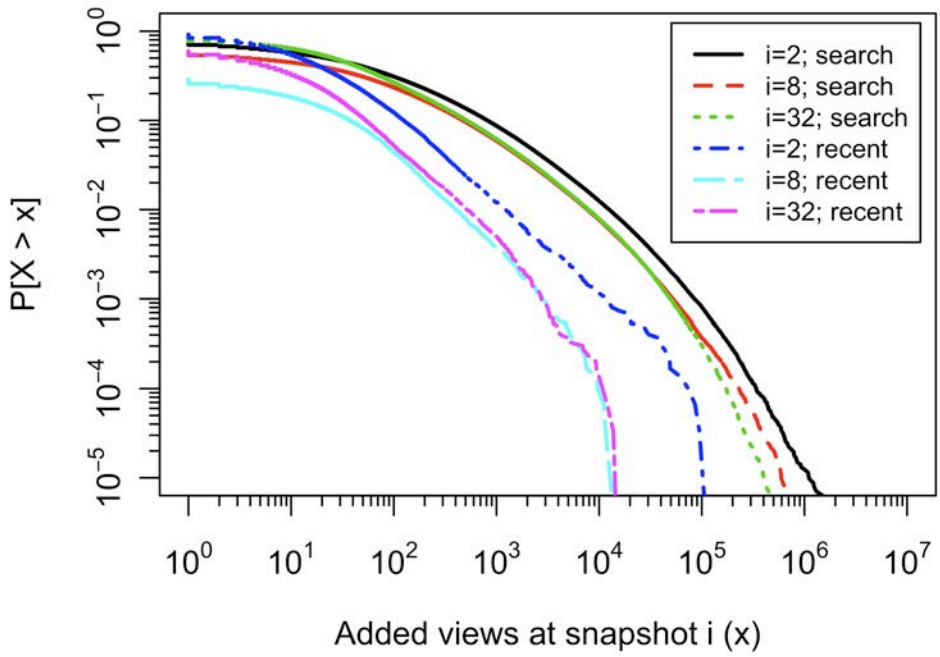



Figure 4.2: Distribution of added views at snapshot $i$, for recently-uploaded and keyword-search videos.

uploaded videos. Further, for the recently-uploaded video dataset, as we look further into the measurement period the curves shift to the left, owing to a decreasing fraction of these videos that are currently popular. The corresponding shift for the keyword-search videos is less pronounced.

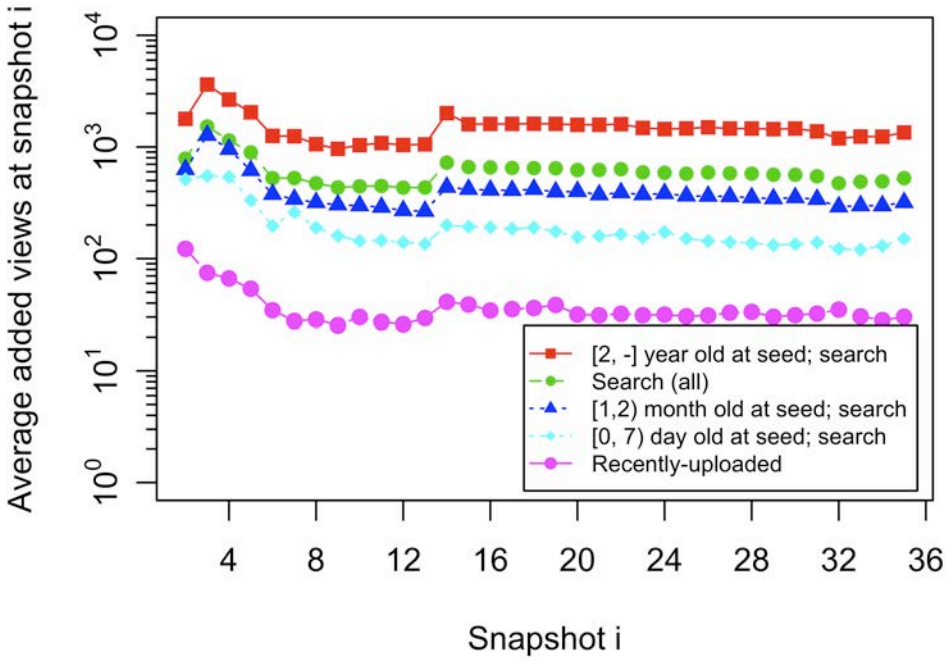



Figure 4.3: Average added views at each snapshot.

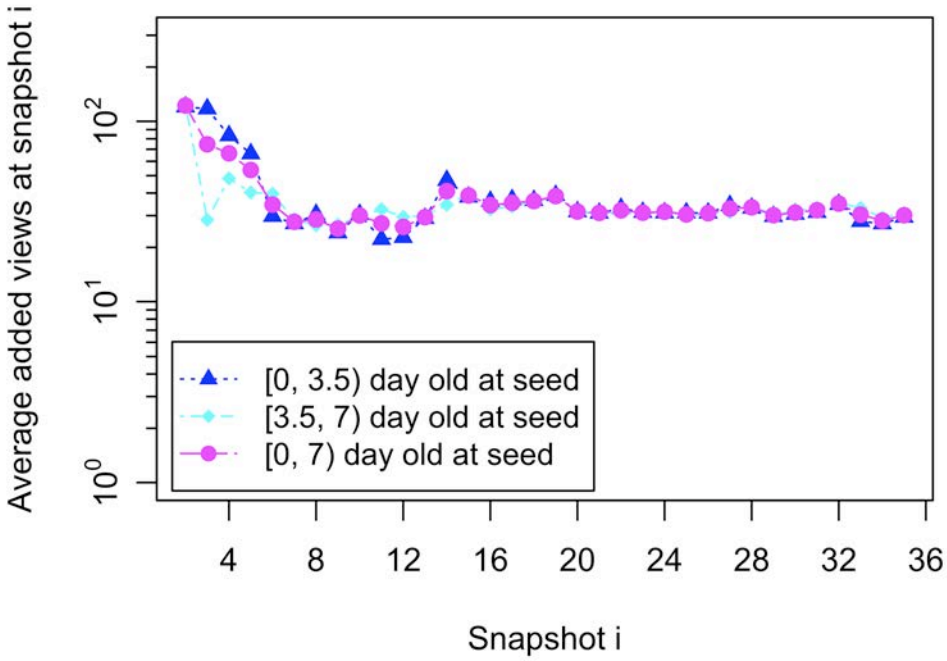



Figure 4.4: Average added views at each snapshot for subgroups of the recently-uploaded videos.

The popularity characteristics of the recently-uploaded and keyword-search videos are investigated further in Figure 4.3, which shows the average added views at each snapshot for both datasets. For the keyword-search videos, we also consider subgroups based on video age at the time of seeding. On average, the keyword-search videos (cf. the "Search (all)" line in the graph) attract more than 10 times as many views throughout the measurement period compared to the recently-uploaded videos, providing additional evidence that the keyword-search video dataset is biased towards more popular videos. (Note that the y-axis is on logarithmic scale.) The results for the keyword-search video subgroups based on age further support this conclusion. First, we note that the older videos in the keyword-search dataset appear to attract substantially more new views, on average, than their younger counterparts. Second, note that the week-or-less old keyword-search videos obtain new views at a higher rate than the recently-uploaded videos, which have the same age range, throughout the measurement period.

The fact that these popularity differences persist for the entire measurement period indicates that the keyword-search video dataset is biased towards videos with elevated long-term popularity. Figure 4.3 also suggests bias based on elevated short-term popularity. In particular, consider the results for keyword-search videos that are 2 years or older at the time of seeding. For a randomly-selected set of videos of this age, one would expect to see a fairly stable average viewing rate over periods of a few weeks, whereas for this subgroup of keyword-search videos, as seen in Figure 4.3 there is an initial period of significantly higher average viewing rate, reflecting elevated short-term popularity.

Next, we consider further the possibility of biases in the recently-uploaded video dataset. Recall from Figure 4.1 that the age at seed time for these videos is approximately uniformly distributed, up to a maximum of one week. One indicator of bias towards more popular videos would be a correlation between the age at seed time,

and the rate of accumulating new views, since it may be easier to predict (for the purposes of preferential selection) the future popularity of older videos. Figure 4.4 shows the average number of added views at each snapshot for those videos in the recently-uploaded video dataset whose age at seed time is less than 3.5 days, for those videos whose age at seed time is at least 3.5 days, and for the entire dataset. Note that there are no observable longer-term differences in the viewing rate behavior among these three groups of videos. There do exist some differences in the first few weeks of the measurement period, which is to be expected for recently-uploaded videos. We have also considered other properties such as the distribution of the total views at the end of the measurement period, and have similarly found no significant differences among these groups.

To summarize, our conclusions regarding sampling bias are:

- The keyword-search videos appear to be biased towards those videos that exhibit both higher short-term and long-term popularity.
- The recently-uploaded video dataset appears to exhibit no observable bias towards popular content.

Based on our analysis, we conjecture that the recently-uploaded video dataset is a random sample representative of the videos uploaded to the service. In the remainder of this paper, we characterize the popularity dynamics of these videos over the first eight months of their lifetime, and from this characterization develop a model for popularity evolution of newly-uploaded videos.

## 4.2 Popularity Dynamics and Churn

In this section, we dig deeper into the popularity dynamics of the videos in the recently-uploaded dataset, with the objective of developing insights for modeling the

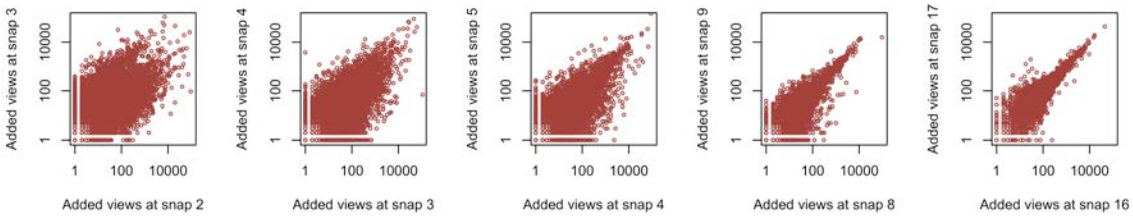
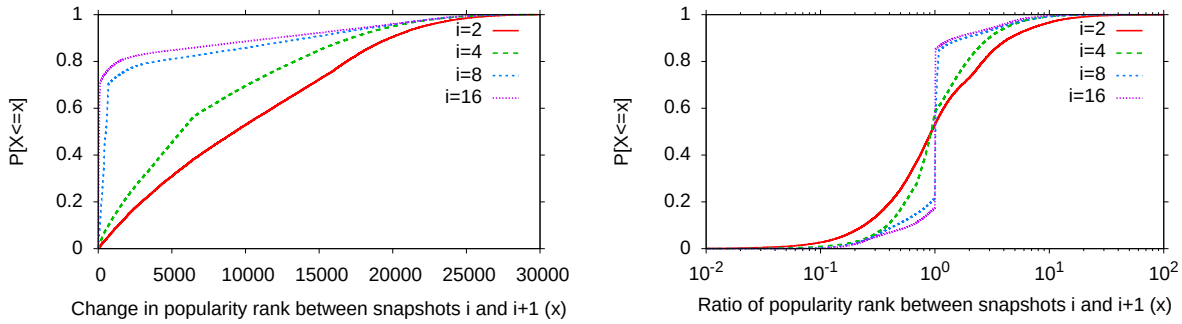



Figure 4.5: Scatter plot of the number of added views at snapshots i versus i+1.

popularity evolution of these videos. One goal is to understand whether or not current popularity is a good predictor of future popularity of user-generated videos. If current popularity is indeed a good indicator of future popularity, then modeling the popularity evolution of individual videos is certainly feasible [60]. However, the popularity of an object may be influenced by many exogenous and endogenous factors [59, 27], which may introduce some degree of inherent unpredictibility [59, 43]. In this section, we characterize the degree of (in)stability and (un)predictability of the popularity of individual videos and the extent of churn in the relative popularities of videos.

Figure 4.5 shows scatter plots for the number of added views received by a video at adjacent snapshots for some example early and later snapshots. With our notion of added views at a snapshot (or the weekly viewing rate, as determined by our measurement granularity), this figure illustrates the change in viewing rate between consecutive snapshots. The scatter plots, especially for the first few snapshots since a video is uploaded, show substantial point spreads which indicates that a large number of videos experience significant variation in viewing rate from one week to another. We observe that a video that is mildly popular in the week prior to one snapshot can become highly popular before the following snapshot, and vice versa. Videos that have about 1,000 views added at snapshot two, for example, could receive less than 100 additional views, or more than 10,000 additional views, at snapshot three. Overall, we observe substantial non-stationarity in the popularity of individual videos, especially within the first five to six weeks of their upload. Looking further in our measurement period, we see that there are fewer diverging points at the top right




(a) Absolute change in ranks (b) Ratio of new to old popularity ranks

Figure 4.6: Distribution of change in popularity ranks of videos.

quadrant of the scatter plots, as the videos become older. Note that the scatter plots have fewer points for later snapshots owing to videos that received no views in one or both of the weeks of interest (and hence are not shown on the log-log plots).

We also computed the Pearson's correlation coefficient between the added views at adjacent snapshots. A correlation coefficient value of 0.8 or more is considered to reflect strong positive linear correlation [18]. The correlation coefficient between the added views at snapshots two and three is close to zero (0.09). Until week eight of our measurement, as may be expected from visual inspection of Figure 4.5, at best, a weak positive linear correlation (less than 0.7) between added views at successive snapshots is observed. As videos become older, we observe a very strong positive linear correlation between the viewing rate of videos across adjacent snapshots. These observations, together with Figure 4.5, indicate that the initial or current popularity of a random young video is likely not a reliable indicator of its future popularity; on the other hand, it appears that the current popularity of an older video may be indicative of its immediate future popularity.

The non-stationarity in the weekly views to videos impacts the relative popularity of videos. For any snapshot of our measurement period, we can rank the videos according to the number of views added to each video's view count at the considered snapshot. Ties are broken using an assigned video id. Based on the assigned ranks

in any two snapshots, we can calculate how much each video's rank shifts. Figure 4.6 (a) shows the cumulative distribution of the absolute value of the rank shifts for some example snapshots. Early in the measurement period, and thus when the videos are young, videos experience significant rank changes. For example, between snapshots two and three more than 30% of the videos in our recently-uploaded dataset switch 10,000 or more rank positions. The changes in the relative popularities of videos stabilize after the initial weeks; however, there are still some videos that experience substantial rank shifts between consecutive weeks. This trend is consistent with the trend suggested by Figure 4.5.

In Figure 4.6 (b), we present the cumulative distribution of the ratio of ranks for some example snapshots. This analysis complements the results presented in Figure 4.6 (a) by considering each video's popularity rank increase/decrease relative to its current rank. Significant changes in the relative ranks of videos are observed when videos are young. Between snapshots two and three, for example, about 30% of the videos gain a factor of two or more in popularity rank, whereas less than 10% of the videos experience similar increases in popularity rank in later snapshots. In fact, when videos become eight weeks or older, approximately 75% of them retain their popularity rank across (weekly) snapshots.

Differences in how rapidly videos attain their peak popularity can be a major cause of churn in relative popularities. Figure 4.7 shows the cumulative distribution of time-to-peak for the videos in the recently-uploaded dataset, where we define time-to-peak for a video as its age (time since upload) at which its weekly viewing rate is the highest within our measurement period.[4] The time-to-peak distribution shows that a large fraction of the videos, approximately three-quarters of them, peak within the first six weeks since their upload. The remainder peak at times approximately

[4]Appendix A describes the details of how we determine this age given the fairly coarse granularity of our measurements. In general, we have found our results to be insensitive to alternative choices for these details.

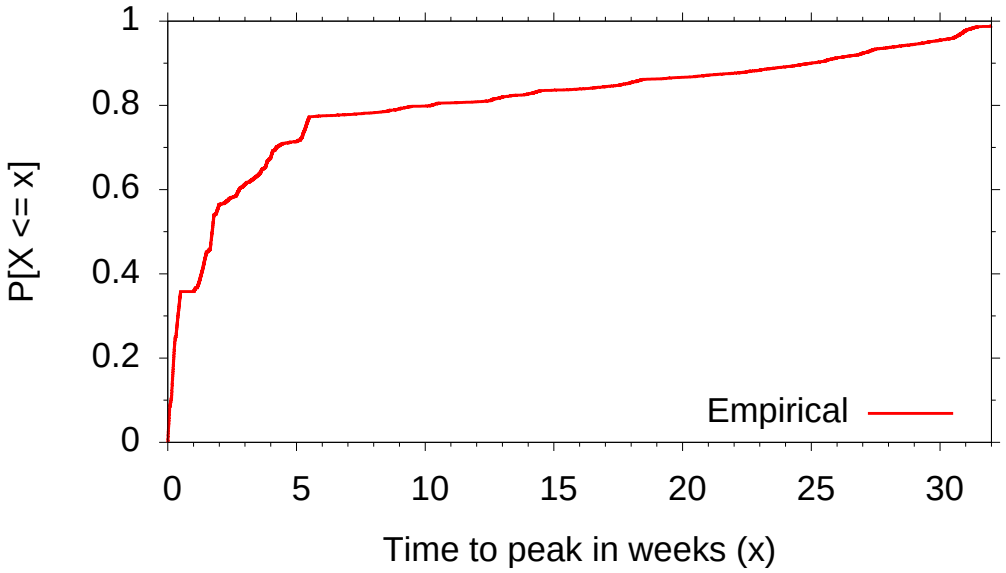



Figure 4.7: Time-to-peak distribution for videos.

uniformly distributed between week six and the end of our measurement period. For those videos that peak within the first six weeks after upload, we find the time-to-peak to be approximately exponentially distributed. As we show later in the paper, the fact that many videos reach their peak popularity quickly plays an important role in explaining the high churn observed in the relative popularity of the videos over the first few weeks of the measurement period.

## 4.3 Three Phase Characterization

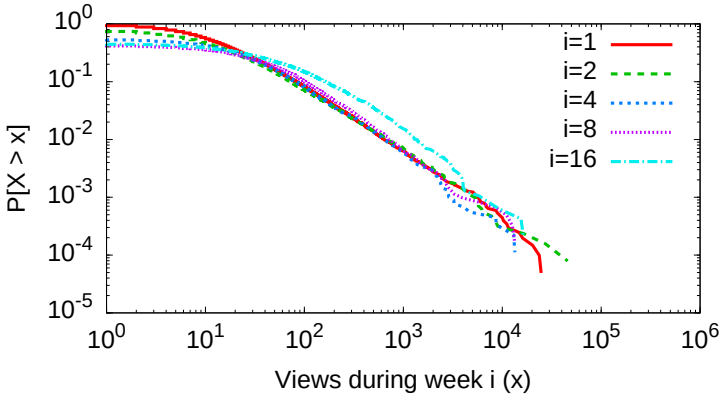
The results of Section 4.2 suggest the futility of attempting to reliably model the popularity evolution of individual videos. We can, however, attempt to model the popularity dynamics of a collection of videos. In this section, we develop a characterization of the popularity evolution observed for our dataset of recently-uploaded videos. This characterization is applied to develop a popularity evolution model in Section 4.4.

Perhaps the biggest challenge in developing such a characterization is that of capturing the churn in the relative popularities of videos that is observed in the empirical data. As noted in Section 4.2, variations in time-to-peak may be an important factor in this churn. This motivates us to develop a *three-phase* characterization of popularity evolution, in which videos are grouped according to whether they are *before*, *at*, or *after* the age at which they attain their peak popularity.

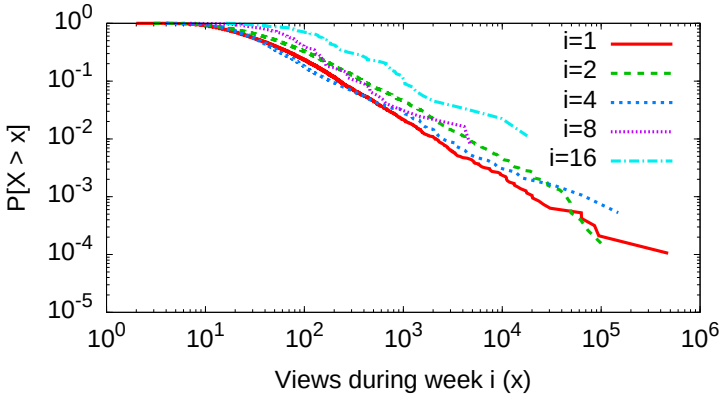
Of particular interest in this characterization are: (a) the movement of videos among these phases (i.e., the time-to-peak distribution, as examined in Section 4.2), (b) the distribution of the viewing rate for the videos belonging to each group, and (c) the dependence of these distributions on video age.

First, consider the distribution of weekly views to videos in each phase. Figure 4.8 shows the complementary cumulative distribution of views during a week within each phase, using a logarithmic scale on each axis. Note that, by definition, none of the videos are past their peak (i.e., in the after-peak phase) in the first week. Similar to the distribution of views during a week for all videos, the distribution of views for the videos within each phase is also heavy-tailed. It also appears that the skew towards larger view counts is the largest when the videos are at their peak, and the least when the videos are past their peak. By inspection of Figure 4.8, we notice that, within each phase the distribution of views each week is approximately similar and suggests the possibility of modeling the distributions as week-invariant.

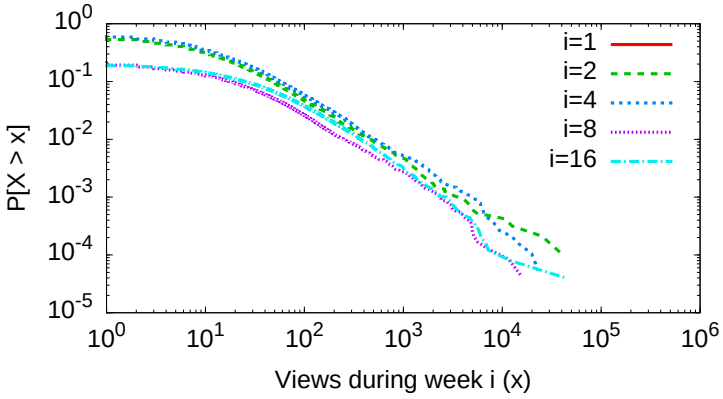
We now investigate the efficacy of assuming the weekly viewing rate within each phase to be approximately week-invariant. Figure 4.9 shows for each week and phase (of the lifetime of the videos) the average number of weekly views. The average viewing rate at peak exhibits a fair degree of variability. The observed variability may be expected as videos peak with varying (and occasionally very large) numbers of views during a week. An interesting observation is that there is no discernible trend in the average viewing rate in the at-peak phase; this provides additional evidence in




(a) Before-peak phase




(b) At-peak phase




(c) After-peak phase

Figure 4.8: Distribution of weekly views to videos in the before-peak, at-peak, and after-peak phases for example weeks $i$ ($i = 1, 2, 4, 8, 16$).

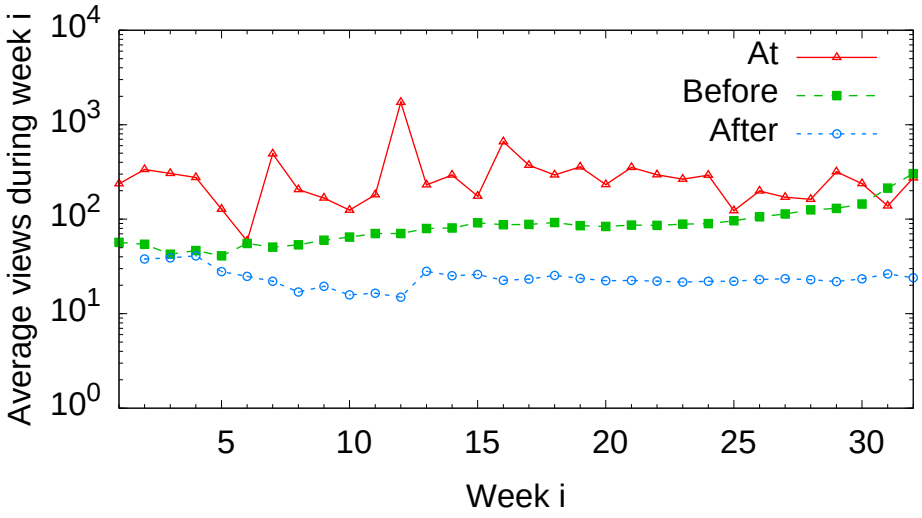



Figure 4.9: The average weekly viewing rate of videos in the before-peak, at-peak, and after-peak phases.

support of a modelling approach in which the at-peak views distribution is modelled as week-invariant.

Unlike the high variability observed for the average number of weekly views to videos that are in their at-peak phase, the average views for after-peak videos appears to be quite stable throughout the measurement period. The average views for before-peak videos also lacks the high variability that is observed for at-peak videos, but appears to exhibit an increasing trend. This increasing trend may be an artifact of the finite measurement period, however; note that as the end of the measurement period grows closer, the maximum time period until each of the before-peak videos peaks corresponding shrinks.

Working further with our week-invariant assumption, we take a closer look at the distribution of views during a week, or equivalently the distribution of the viewing rates, for each of the three phases. Our goal was to get an understanding of the distribution of the viewing rate when videos are grouped by phases, ignoring week-specific behavior. Figure 4.10 shows that the distribution of weekly views to videos

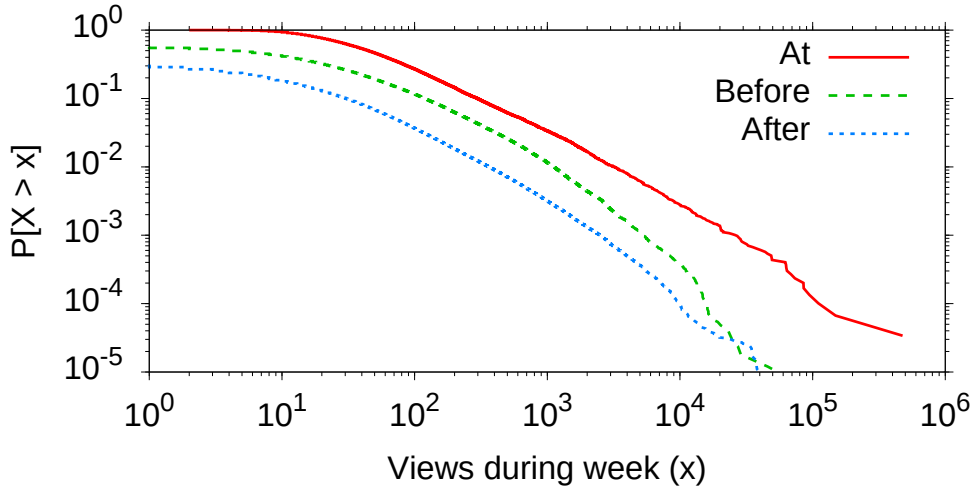



Figure 4.10: Distribution of views during a week for videos that are in their before-peak, at-peak, and after-peak phases.

within each phase is heavy-tailed. As expected from our earlier discussion, the skew towards larger views is greatest when videos are at their peak, and least when they are past their peak.

Because of the heavy-tailed nature of the views distribution in each phase, we took a closer look at the tail of each distribution where we define the tail to consist of only the largest ten percent of the views within each phase. Using our definition of the tail, we determine thresholds of 116, 296, and 31 weekly views for a video to be considered in the tail of the before-peak, at-peak, and after-peak view distributions, respectively. Figure 4.11 shows, for each week of our measurement period, the average number of weekly views for those videos that acquired greater than or equal to these threshold weekly views. The average viewing rate is quite steady for each phase, suggesting that the week-invariant assumption is a reasonable approximation for the distribution tails.

To summarize, our three-phase characterization suggests that the viewing rate distribution within each phase could be modelled as week-invariant. We find that the

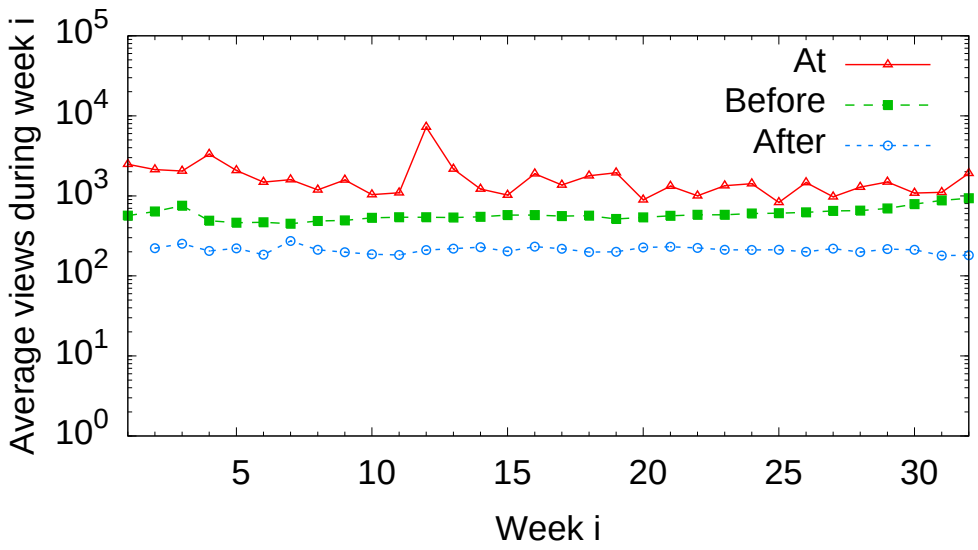



Figure 4.11: The average weekly viewing rate of videos in the *tail* of the before-peak, at-peak, and after-peak distributions.

tail of the viewing rate distribution can be modelled separately using heavy-tailed distributions. Appendix A.2 presents the specific distribution fits that are found to best capture the characteristics of the empirical data. Overall, we find that the distribution of weekly views, for each of the three phases, can be modelled using an appropriately parameterized lognormal distribution for the tail and a beta distribution for the views that are not in the tail.

## 4.4 Basic Model

Guided by the observations made in the foregoing sections, we develop a basic model for generating weekly views to *individual* videos in a *collection* of newly-uploaded videos. The model is developed using the observations pertaining to the before-peak, at-peak, and after-peak phases. The distribution of weekly views to videos within each phase is modelled to be week-invariant. From a modeling point of view, this is an attractive property as the distribution of weekly views can be succinctly represented

using only three distributions, one for each phase. Transitions of videos between phases, specifically from being in their before-peak phase, to their at-peak phase, and then to their after-peak phase, are modelled using a time-to-peak distribution (such as the one shown in Figure 4.7).

The basic approach consists of sampling views from the before-peak, at-peak, and after-peak distributions (cf. Figure 4.10), and assigning them to videos. For each modelled week, we sample views from the before-peak, at-peak, and after-peak distributions based on how many videos are in each of these phases. Note that the number of videos that peak in any week is determined using a time-to-peak distribution (cf. Figure 4.7). At the start of an arbitrary week, from among the videos that have not peaked thus far, some videos transition to being at their peak; subsequently, at the end of this week these videos will move into the after-peak phase.

For this approach to yield weekly views for individual videos, a framework for assigning the sampled views to individual videos is required. A straightforward approach for assigning weekly views to videos is based on an assumption that the relative popularities of videos in the same phase, or that were in the same phase during the previous week, are unchanged from the previous week, and precedes as follows. Assign the views sampled from the before-peak and at-peak distributions to those videos that were in their before-peak phase during the previous week; similarly, assign the sampled views from the after-peak distribution to those videos that were in their at-peak or after-peak phases during the previous week. In both cases, the assignment is made such that the relative popularities of the respective videos are preserved. Now, among those videos that were in the before-peak phase during the previous week, those that are assigned views from the at-peak distribution are assumed to peak in this week (and will be in their after-peak phase for all subsequent weeks). With this approach for assignment of views, churn with respect to the relative popularity of videos is introduced by videos moving between the three phases.

### 4.4.1 Views Generation Algorithm

Our algorithm requires the following input: the total number of newly-uploaded videos $N$, the total number of weeks $d$, a time-to-peak distribution, a distribution for the weekly views for videos in the before-peak phase, a distribution for the weekly views for videos in the at-peak phase, and a distribution for the weekly views for videos in the after-peak phase. The main steps of the algorithm are as follows:

1. **Determine the number of videos in the before-peak, at-peak, and after-peak phases.**

   Sample $N$ values from the time-to-peak distribution and determine the number of videos $n_i^{at}$ that peak at week $i$, for all $i \leq d$. Note that $n_i^{before} = n_{i-1}^{before} - n_i^{at}$, $n_i^{after} = n_{i-1}^{after} + n_{i-1}^{at}$, for $i > 1$. Also note that $N = n_i^{before} + n_i^{at} + n_i^{after}$ and $n_1^{after} = 0$. Therefore, following this step, we know the number of videos $n_i^{before}$, $n_i^{at}$, and $n_i^{after}$ that are in the at, before, and after-peak phases, respectively, during week $i$. In our experiments, as described in Appendix A.2 the time-to-peak distribution is chosen as a mix of an exponential and a uniform distribution.

**For** $i = 1, 2, \cdots d$**:**

2. **Sample views from the before-peak, at-peak, and after-peak distributions.**

   Sample $n_i^{before}$, $n_i^{at}$, and $n_i^{after}$ times from the before-peak, at-peak, and after-peak distributions. In our experiments, we use a mixture of beta and lognormal distributions for each of these three phases.

3. **Assign views to the videos.**

   (a) if $i = 1$: Note that $n_1^{after} = 0$, i.e., there are no videos in week one that are after their peak. Assign the $N$ $(= n_1^{before} + n_1^{at})$ sampled views to the videos.

   (b) if $i > 1$: Sort the sampled $n_i^{at}$ "at-peak" views and the $n_i^{before}$ "before-peak"

views and assign them to those videos that are in the "before-peak" phase during week $i-1$ such that the video with the highest view during week $i-1$ is assigned the highest view in week $i$, the video with the second highest view during week $i-1$ is assigned the second highest view during week $i$, and so on. Similarly, assign the sampled $n_i^{after}$ after-peak views to those videos that were either at or after their peak in week $i-1$.

4. **Determine the videos that peak in this week.**

   The videos that were assigned views sampled from the "at-peak" distribution are assumed to peak this week; for all subsequent weeks these videos will be in their "after-peak" phase.

### 4.4.2 Results and Discussion

This section presents results from our basic model. We used our implementation of the basic model to generate synthetic views for $N = 29,791$ videos for a total of $d = 32$ weeks. The parameterization of the distributions (i.e., time-to-peak, weekly views during each phase) was done as specified in Appendix A.2.

Simple tests of our model include comparisons of the time-to-peak distribution, and the viewing rate distributions for videos in each of the three phases, for the synthetic data versus the corresponding empirical distributions for the recently-uploaded video dataset. Good matches are obtained but this is not surprising since the model was parameterized from these empirical distributions. Such tests do not show that our simple three-phase characterization (on which our model is based) captures enough of the detail of popularity evolution, to ensure that our synthetic data matches the empirical data on the metrics of practical interest concerning popularity and its evolution.

For such an evaluation, we test whether the synthetic data matches the empirical data from the recently-uploaded video dataset with respect to: (a) the distribution

(over all videos) of views received during each week (e.g., the skewness in popularity among the videos, and the evolution of this skewness over time), (b) the distribution of the total views since upload at the end of each week (e.g., the skewness in total accumulated video views, or "long term average popularity", and the evolution of this skewness over time), and (c) hot set dynamics (e.g., how much churn is experienced in hot sets of various sizes from week to week). Note that the evaluation metrics considered above are not directly fitted from the empirical data. Instead, for the synthetic data, these evaluation metrics are consequences of the views generation algorithm and modeling parameters derived from the three-phase characterization of the recently-uploaded dataset.

### Distribution of Weekly Views

Figure 4.12 shows the CCDF of the views received during week $i$, for $i = 2, 8, 32$, for both the recently-uploaded dataset and the synthetic views generated by our basic model. We first observe that the weekly views distributions exhibit heavy tails, with videos receiving fewer large views in later weeks than earlier weeks. Further, we observe that the average weekly views to videos do not significantly change after the initial six weeks.

Overall, the match between the model generated views and the views in the recently-uploaded dataset is good, except for some differences for the least popular videos during a week (which is to be expected owing to our simplifying assumption of week-invariant distributions for the phases). Quantile-to-quantile (Q-Q) plots were used to evaluate the match for the body and tail of the distribution (cf. Appendix A.3). The observation pertaining to the average weekly views changing substantially only in the first six weeks is explained by the fact that a large majority, approximately 80% of the videos, peak within the first six weeks since their upload. Once a video is past its peak, it is in the after-peak phase where the viewing rate is

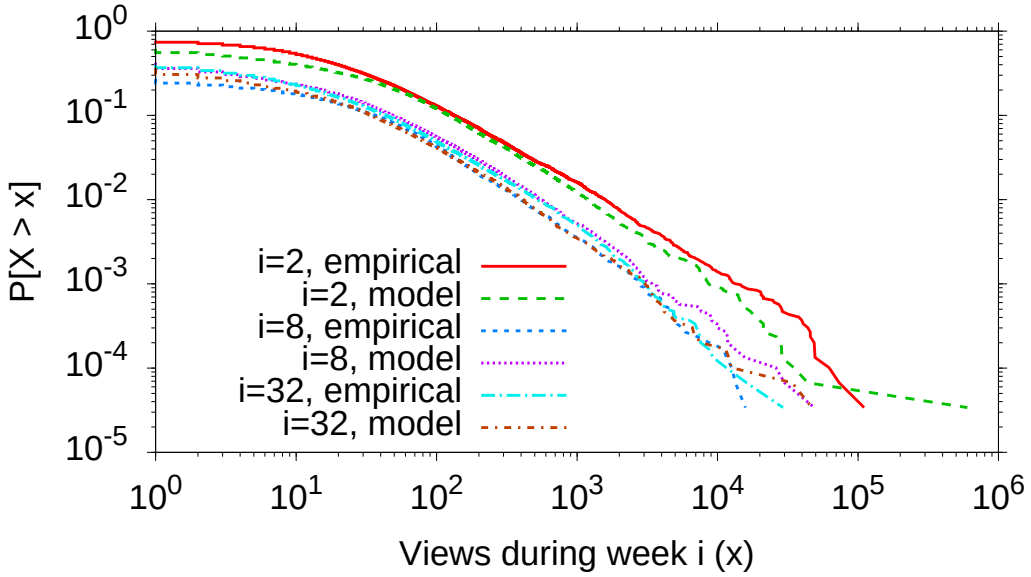



Figure 4.12: Distribution of the views during week i in the recently-uploaded dataset and the basic model (i = 2, 8, 32).

approximately week-invariant. With a majority of the videos in the after-peak phase, the average viewing rate remains approximately constant.

### Distribution of Total Views

The next property we consider is the distribution of total views as a function of weeks since upload (or equivalently the age of the videos). Figure 4.13 shows the CCDF of the total views received by week $i$, for $i = 2, 8, 32$, for both the recently-uploaded dataset and the synthetic views generated by our basic model. We observe an excellent match between the empirical and synthetic datasets. Again, Q-Q plots were used to evaluate the match for the body and tail of the distributions (cf. Appendix A.3).

The general shape of the total views distribution from the recently-uploaded dataset and the model provide insights into the popularity dynamics of the videos. Notice that the distribution of views during a single week, shown in Figure 4.12 for several representative weeks, appears to have a fairly "straight" tail. The total views distribution, for both the empirical dataset and the model, is fairly straight for the

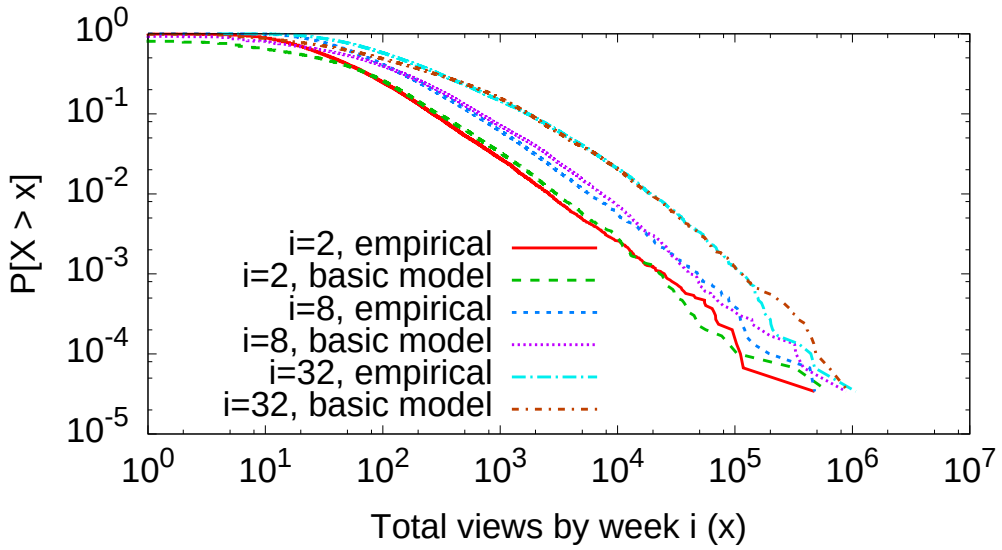


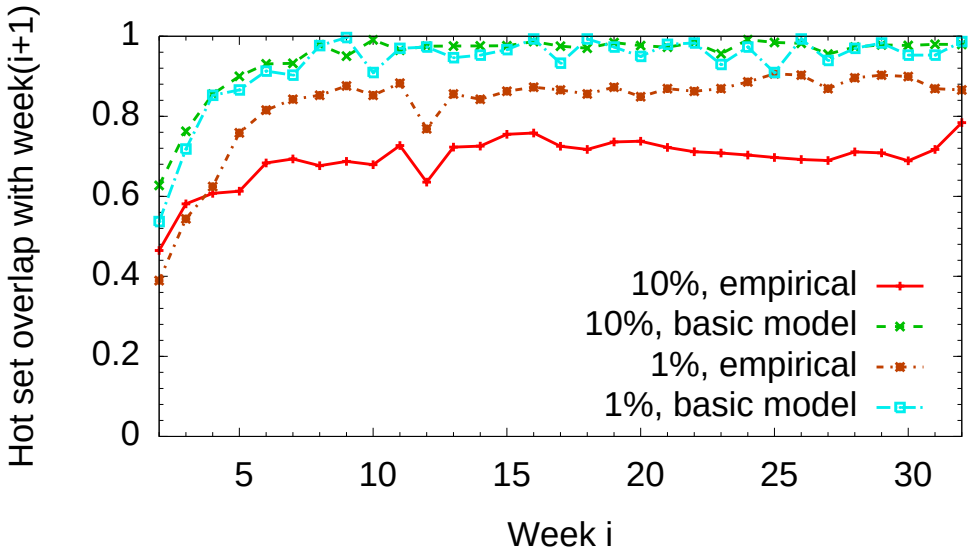

Figure 4.13: Distribution of the total views by week i in the recently-uploaded dataset and the basic model (i = 2, 8, 32).

first few weeks, but transitions to a more "curved" tail as the videos become older. In the figure presented, this change can be seen by comparing the curves for weeks two and eight. If videos that are currently popular continue to be popular in the future, as one expects in simple rich-get-richer models, then we expect the distribution of the sum of the views to also exhibit a "straight-ish" tail. We believe that this change in the characteristics of total views can be explained by the presence of churn in video popularity. Our basic model, which retains strong correlation between current viewing rate and future viewing rate except when videos change state (e.g., move from being in the before-peak phase to at-peak phase to after-peak phase), exhibits a very similar change in the shape of the distribution.

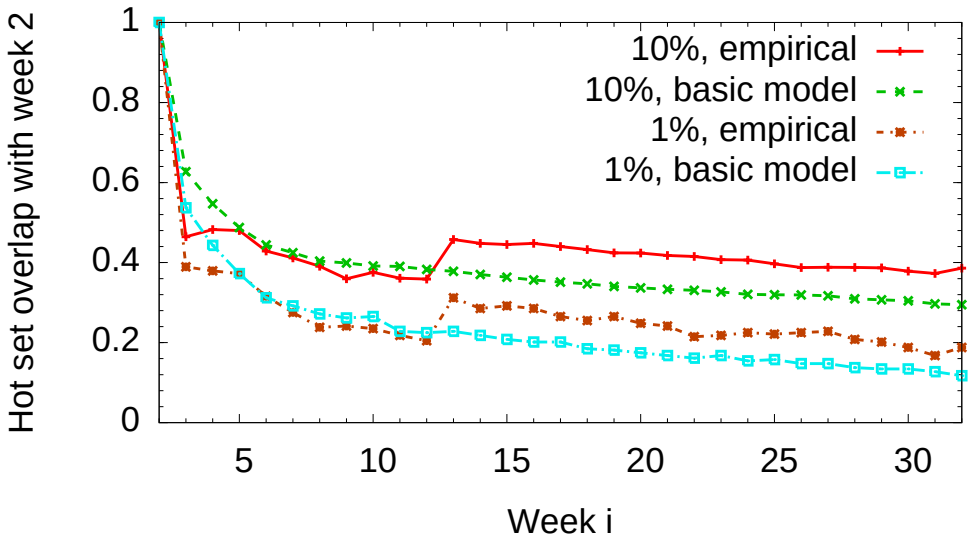
### Churn and Hot Set Dynamics

The skew observed in the popularity of videos can aid in caching [30, 18]. However, caching decisions become difficult with increase in churn among the videos [49]. This section quantifies the amount of churn among the most popular videos, its potential

impact on caching decisions, and studies how well our model captures the churn observed in the recently-uploaded dataset.




(a) Week i and i+1 hot sets




(b) Week 2 and week i hot sets

Figure 4.14: Churn in video popularity measured by changes to the hot set for the recently-uploaded dataset and the basic model.

For studying churn, we define the hot set at week $i$ to consist of the most popular $x\%$ of the videos with respect to the views received during week $i$. Figure 4.14 (a)

shows the overlap between hot sets of successive weeks for both the recently-uploaded dataset and our basic model, for hot sets of size $x = 1\%$ and $x = 10\%$. If caching decisions are made based on the hot sets of the week immediately preceeding the current week, then these graphs give us an indication of the amount of cache replacement traffic. Effectively, these results tell us how good an indicator the immediate past is with respect to the immediate future.

The results indicate presence of substantial churn. With a hot set of size $x = 10\%$, for example, we observe between 20-60% change in the constitution of the hot sets between two consecutive weeks, with significantly higher churn observed in the first eight weeks. Comparing the results for the smaller and larger hot sets, the percentage change is more for larger hot sets, because of replacement of videos in the hot set with videos of similar popularity. Our model captures the trend of increased churn early in the lifetime of videos. However, our model suggests relatively smaller week-to-week changes than the corresponding empirical dataset.

The results presented in Figure 4.14 (a) show the relative change in the hot set from week-to-week. Figure 4.14 (b) presents an example result for the absolute change in the hot set. Here, we measure the number of common videos between the actual hot set at week $i$ and the hot set of week two.[5] Both the model and the empirical hot set analyses show that there is substantial non-stationarity in the hot sets, with the model capturing the trend exhibited by the recently-uploaded dataset.

It is not surprising that our basic model does not exhibit as much churn as seen in the recently-uploaded dataset. Our basic model introduces churn by transitioning videos between phases. The model captures the large change in position caused by videos moving from being before their peak to being at their peak, and subsequently being after their peak. As we capture churn caused only by movement between phases,

[5]We present comparisons with the hot set for week two because the empirical dataset has only a partial first week (i.e., a snapshot at seed); refer to Section 3.1.1 and Appendix A.1 for details on data collection and sampling granularity.

we see a better match for the smaller hot set then the larger hot set. The next section extends the model to introduce *second-order churn* effects with respect to the relative popularity of videos within each phase of their lifetime.

## 4.5 Model Extension: Perturbations

Our basic model introduces churn in relative popularity of videos only owing to the videos moving between the three phases during their lifetime. We now extend the model to introduce *second-order churn* by shuffling the popularity of videos within each phase, while preserving each video's phase.

The model extension is as follows. We first generate weekly views for videos conforming to the basic model. Then, we introduce perturbations in the relative popularity of videos during a week by exchanging the views assigned to selected videos. The views are exchanged such that none of the key characteristics of the basic model, specifically the distribution of weekly views for the before-peak, at-peak, and after-peak phases, as well as the distribution of time-to-peak are affected.

The algorithm for introducing additional churn is as follows. First, we assign weekly views to $N$ videos for a period of $d$ weeks according to our basic model. Then, for each week $i$ and video $v$ that does not peak in that week, we define a window $W_i^v$ that specifies the bounds on views for a possible exchange:

$$W_i^v = [\frac{x_i^v}{g}, \min(x_i^v \times g, x_{max}^v)], g \in [1, \infty]$$

where $x_i^v$ is the view assigned to video $v$ during week $i$, $x_{max}^v$ is video $v$'s peak weekly viewing rate (i.e, $x_{max}^v = \max_j x_j^v$), and $g$ is a modeling parameter that controls the maximum distance a video would be shifted with respect to its view count in a week. Specifically, we repeat the following step a sufficiently large number of times:

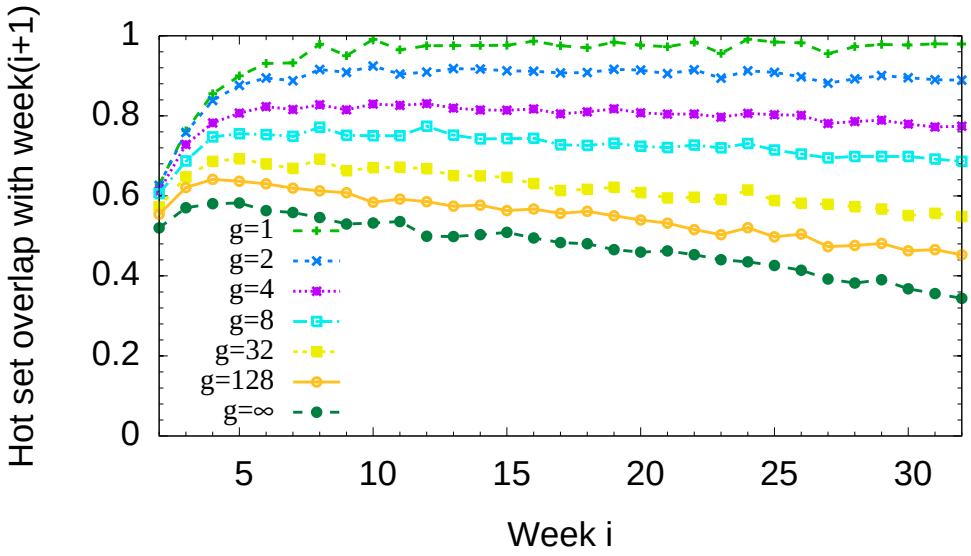



Figure 4.15: Impact of the churn modeling parameter, with respect to the weekly churn in video popularity, as measured by weekly changes to the hot set using the extended model.

- Randomly pick a week $i$ and two videos $u$ and $v$ such that either both videos peaked *before* week $i$, or both peak *after* week $i$. If the views during week $i$ to videos $u$ and $v$ can be exchanged without causing either video's views for week $i$ to move outside their respective windows $W_i^v$ and $W_i^u$, switch the views.

Following a large number of iterations, views would be "uniformly mixed" subject to the constraint specified by the tunable parameter $g$; $g = 1$ conforms to the basic model outlined in the preceeding section, and $g = \infty$ incorporates the maximum possible churn while still preserving the per-phase properties.

Figure 4.15 presents results for a hot set evolution for hot set of size $x = 10\%$ for various values of $g$. As expected, with increasing $g$ there is increased churn. This allows the model to capture a wide range of churn activity. With no additional churn, as described by the $g = 1$ (basic model), the constitution of hot sets between adjacent weeks changes by less than 10% once videos are seven weeks old. With $g = \infty$, the

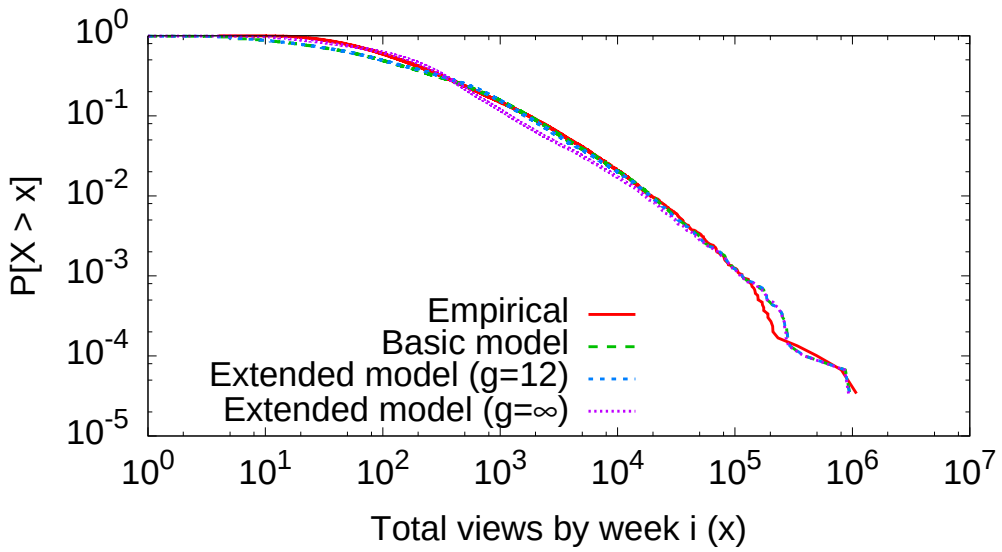
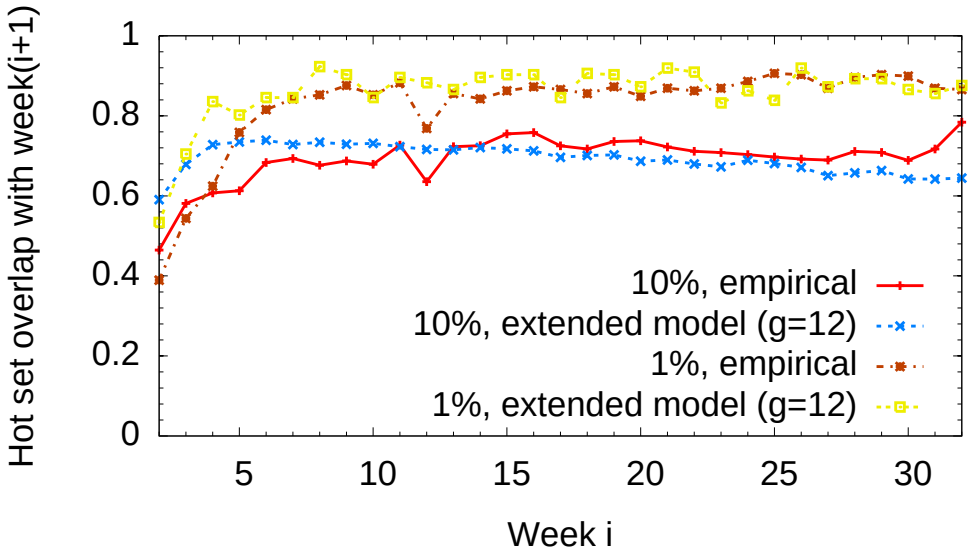



Figure 4.16: The total views distribution after 32 weeks, in the recently-uploaded dataset and the extended model.

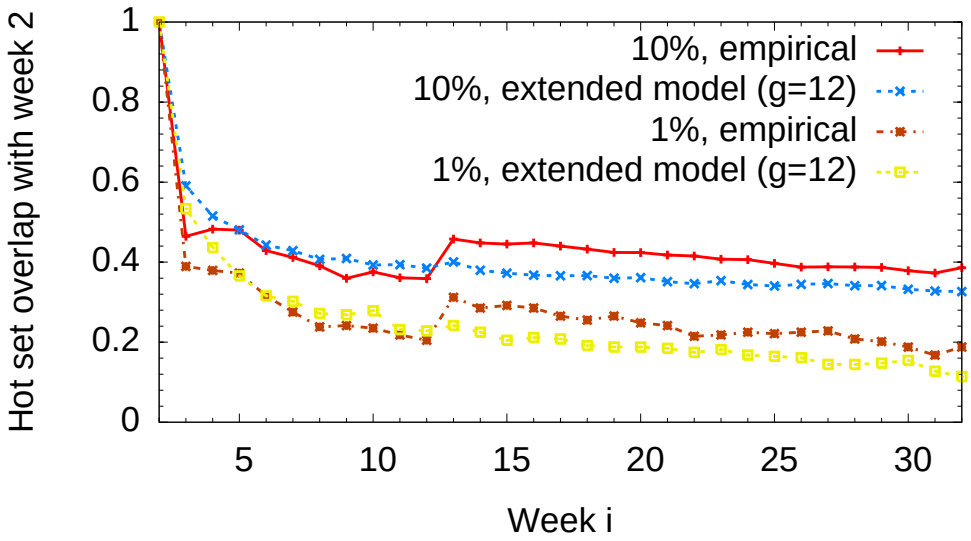
composition of videos in the hot sets change by as much as 50-60% during the course of the 32 weeks.

Through experimentation, we found that g in the range $8 \leq g \leq 16$ yeilds a close match to the churn observed in the recently-uploaded dataset. Figure 4.17 compares the hot set churn in the recently-uploaded data set with the hot set churn using our extended model, with $g = 12$. Results are shown for both hot set size $x = 1\%$ and $x = 10\%$. We obtain a much better match between the curves than when only using the basic model.

Note that the model extension is constructed such that the distribution of weekly views is not impacted by the introduction of additional churn. We close our discussion on model extensions by investigating how the additional churn influences the total views distribution of videos. The results are presented in Figure 4.16. Also here, we note that introduction of additional churn does not affect the tail of this distribution. Addition of further churn, by tuning the parameter $g$, however, has some (mostly negligible) impact at the head of the distribution.




(a) Week i and i+1 hot sets




(b) Week 2 and week i hot sets

Figure 4.17: Churn in video popularity measured by changes to the hot set for the recently-uploaded dataset and the extended model.

## 4.6 Conclusion

Content popularity dynamics can have a significant impact on the effectiveness of different designs of content distribution, content storage, and advertisement systems.

It is particularly important to understand the popularity dynamics of user-generated content, specifically user-generated video, given its widespread appeal. The volume of such content, as provided by popular services such as YouTube, however, makes it challenging to study these dynamics. In this chapter we make several contributions that address this challenge.

We first illustrate the biases that may be introduced in the analysis for some choices of the sampling technique used for collecting data; however, sampling from recently-uploaded videos provides a dataset that is seemingly unbiased. We then develop a framework for studying the popularity dynamics of user-generated videos. Using a dataset that tracks the views to a sample of recently-uploaded YouTube videos over the first eight months of their lifetime, we show that there is significant churn in the relative popularities of videos, mainly because of large differences in the required time since upload until peak popularity is finally achieved, and secondly to popularity oscillation. We find that the current popularity of a video is not a reliable predictor of its future popularity. This finding motivates models that attempt to capture the popularity dynamics of collections of videos, rather than attempting to predict the popularity evolution of individual videos. To this end, we develop a novel three-phase characterization of the popularity dynamics. Based on this characterization, we propose a model that can accurately capture the popularity dynamics of collections of recently-uploaded videos as they age, including key measures such as hot set churn statistics, and the evolution of the viewing rate and total views distributions over time.

# Chapter 5

# Conclusion

Video sharing services provide a convenient platform for widespread dissemination of content. Everyday, several million videos are uploaded and several hundred million videos are watched from YouTube alone. Over time, some videos reach iconic status, while many others are simply forgotten.

There are several factors that potentially impact the views to a video. Popularity evolution of such user-generated content is a complex process and have garnered increasing interest. The focus of this thesis is to study how content-related and content-agnostic factors impact future views to a video and to analyse the long-term temporal dynamics of user-generated videos popularity.

In the first part of this thesis we develop a methodology that is able to accurately assess the impact various content-agnostic factors have on popularity. We posit that the content itself plays a key role for a video's eventual popularity. Therefore, to study how content-agnostic factors such as the uploaders social network size, the current total view count, the current rating of the video, and so on, we need to factor out the impact of the content. Towards this goal, we identify and collect a large dataset that consists of multiple identical copies (called clones) of a range of different content; we make this dataset available to the research community. We then develop a rigorous

analysis framework, based on well-known statistical tools, that allows us to control bias introduced when studying videos that do not have the same content. Overall, our new analysis framework enables us to understand the influence the content and several content-agnostic factors have on a video's future popularity.

Using our clone-based methodology, we provide several findings. First, we show that inaccurate conclusions may be drawn when not controlling for video content. Second, controlling for video content, we observe scale-free rich-get-richer, with view count being the most important factor except for very young videos. However, when looking across different contents, we demonstrate that the rich-get-richer behavior gets weaker and becomes inaccurate, and thus rich-get-richer type of models do not accurately capture content popularity. Third, we find that while the total view count is the strongest predictor, other content-agnostic factors can help explain various other aspects of the popularity dynamics. For example, the uploader's social network can be a good predictor for newly uploaded videos, and the video age can help differentiate videos at similar view counts that are accumulating views at different rates. Finally, we present concrete evidence of the first-mover advantage where the early uploaders of a content have an edge over later uploaders of the same content.

The observations in the first part of our work motivate the need for a new model that captures the evolution of content popularity in time. The second part deals with the popularity evolution of user-generated video content and presents a systematic study of the long term popularity evolution of videos. The volume of such content, as provided by popular services such as YouTube, however, makes it challenging for researchers to study these dynamics. We make several contributions that address this challenge.

One main contribution concerns the use of sampling. Sampling is necessary given the huge volume of content available from popular services, but sampling may yield datasets biased towards content with elevated short-term and/or long-term popular-

ity. We find that sampling from recently-uploaded videos, as provided by the YouTube API, appears to yield a dataset that is seemingly unbiased, unlike sampling based on keyword searches.

We next show that there is substantial churn in the relative popularities of videos, particularly young videos, and that the current popularity of a video is not a reliable predictor of its future popularity. This finding motivates models that attempt to capture the popularity dynamics of collections of videos, rather than attempting to predict the popularity evolution of individual videos. To this end, we develop a novel three-phase characterization of the popularity evolution of a dataset of recently-uploaded YouTube videos. This characterization provides a basis for a model that, using a small number of distributions as input, is able to generate synthetic data matching empirically observed characteristics with respect to key metrics concerning popularity and its evolution, such as hot set churn statistics, and the evolution of the viewing rate and total views distributions over time.

## 5.1 Future Work

Our long-term research aim is to understand and accurately model the popularity evolution of user generated content, and ultimately design an accurate workload generator. To reach this goal, our future work consists of:

1. Conducting larger scale datasets collection, because some characteristics might be perceived only when the amount of data is fairly large.

2. Working with richer types of content from different user-generated content dissemination services that would provide scope for novel observations and models. We intend to study the suitability of our clone-based approach for other user-generated content platforms, and the applicability of our model for different UGC types.

3. Investigating how to enhance and extend our current workload generator to reflect the general characteristics and dynamics of UGC popularity. Our current model has its own limitations, such as generating views for a certain video set (newly uploaded videos), and for a specific duration (32 weeks). One of our short-term goals for example, is to explore whether or not it's possible to generate views for really old videos (beyond 32weeks).

4. Performing large-scale tests of the current model and its future extensions.

# Appendix A

## A.1 Sampling Granularity

The recently-uploaded dataset was obtained by sampling the video popularity at a weekly time granularity. By taking the difference of the total view counts between consecutive snapshots, we can measure the weekly viewing rate (i.e., the number of views in a week). For simplicity, and for the purpose of our analysis, we say that a video's peak viewing rate occurs at the midpoint between the two snapshots, between which the highest weekly viewing rate was observed. (As we sample videos of different ages at the time that we first start tracking them, we can still have videos peaking at an arbitrary age less than the total measurement period.)

To obtain a weekly viewing rate associated with the first snapshot (at which the videos may be of any age between 0 and 7 days), we inflate the view count at the first snapshot using a fraction of the added views during the following week to account for the missing days needed to get a weekly view count. Note that alternative ways of calculating a weekly viewing rate, such as dividing the views at the first measurement point by the time since upload, may result in extremely large viewing rates, if the time since upload is small, for example. Finally, in the case that the initial viewing rate is higher than for any other measurement point, we say that the video peaked at the halfway point between its upload time and the initial measurement point (as the

rate in this interval, in such a case, is higher than the average rate during following weeks).

## A.2 Model Parametrization

As discussed in Section 4.2, a large fraction of the videos, approximately three-quarter of our sample, peak within the first six weeks since their upload. The remaining peak at times uniformly distributed between week six and the end of our measurement period. To estimate the rate parameter $\lambda$ of the exponential part, we use the *Maximum Likelihood Estimation* (MLE) method. For the recently-uploaded dataset, we determine $\lambda$=0.598. Time-to-peak values greater than six weeks are then drawn from a uniform distribution $U(6, d)$, where $d$ is the duration of the measurement period. Figure A.1 shows the cumulative distribution function (CDF) of the empirical time-to-peak and the analytical fitting.

We parametrize the weekly views of videos belonging to the before-, at-, and after-peak phases. As our model assumes week-invariant distributions for the three phases, we only consider the aggregated before-peak, at-peak, and after-peak weekly views. The body and tail are modelled separately, with the tail assumed to consist of all videos with weekly views greater or equal to a threshold $x_{thresh}$ views, selected such that the tail of each phase contains the 10% videos with the largest weekly view counts.

The distribution of weekly views within each phase is heavy-tailed (cf. Figure 4.10). Using the approach developed by Clauset et al. [24], we investigated whether the tail of each phase could be modelled using a power-law distribution or a lognormal distribution.

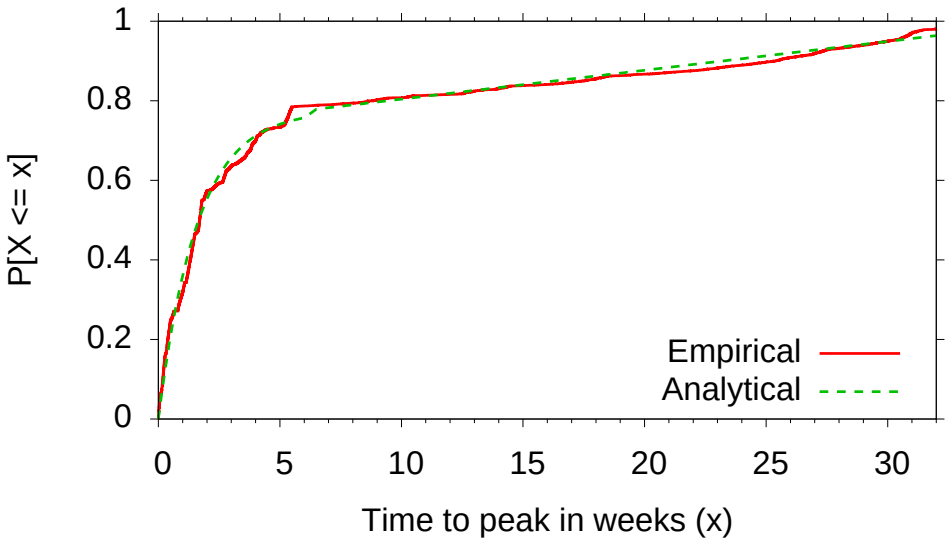



Figure A.1: Time-to-peak distribution of videos.

The MLE method is used to estimate the respective distribution parameters. The power law scaling parameter $\alpha$ (for the continuous case) can be estimated using:

$$\alpha = 1 + n \left[ \sum_{i=1}^{n} \ln \frac{x_i}{x_{thresh}} \right]^{-1},$$

where $n$ is the number of unique view count observations that fall into the tail distribution. To estimate the parameters $\mu$ and $\sigma$ of the lognormal model, direct maximization of tail-conditional log-likelihood was applied [24]. Note that to model the empirical data using a lognormal distribution above the specified lower threshold $x_{thresh}$, we use the *tail-method*, where we consider that the right tail exhibits the same shape as the right tail of a lognormal distribution, without essentially having an equal probability of being in the tail.

Table A.1 presents the parameter estimation results for the tail of the before-peak, at-peak, and after-peak distributions. Each group has $n_{tail}$ observations $x \in [x_{thresh}, x_{max}]$, $\alpha$ is the scaling parameter of the power law model, and $\mu$ and $\sigma$ are the parameters for the lognormal model.

Table A.1: Power law and lognormal fits for the tails of the distributions.

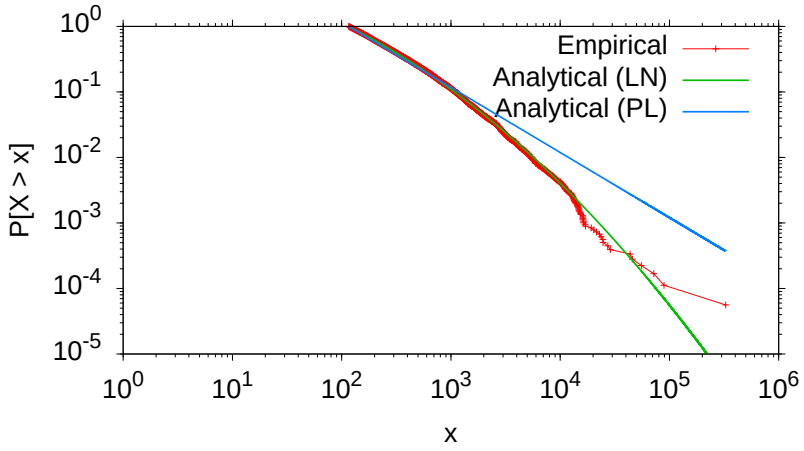
| Phase | Parameters | | | Power law Fits | Lognormal Fits | | R=$\frac{L_p}{L_{ln}}$ |
|---|---|---|---|---|---|---|---|
| | $n_{tail}$ | $x_{thresh}$ | $x_{max}$ | $\alpha$ | $\mu$ | $\sigma$ | |
| Before-peak | 16,829 | 119 | 89,090 | 1.996 | 2.000 | 2.135 | -186.947 |
| At-peak | 2,986 | 297 | 476,100 | 1.950 | -3.826 | 3.477 | -8.164 |
| After-peak | 83,148 | 30 | 94,930 | 1.895 | -0.356 | 2.533 | -762.172 |

In order to compare the power-law and lognormal models, we apply the *Log Likelihood Ratio* (LLR) test to determine which distribution best fit the empirical data [24]. This test computes the ratio of the logarithm of the the likelihood of our empirical data in the two candidate distributions, and find which distribution best fits the data depending on the sign of LLR. In our case, we calculate the log-likelihood ratio, $R = \log(\frac{L_{PL}}{L_{LN}})$ where, $L_{PL}$ is the likelihood of the power law distribution, and $L_{LN}$ is the likelihood of the lognormal distribution. The values of $R$ in Table A.1 suggest that the lognormal hypothesis is more suitable to model the distribution of weekly views for each of the three phases. To verify that the sign of $R$ can be reliably used to make a quantitative judgement about which model is a better fit, we computed the standard deviation of $R$ using the *Vuong* method [24]. Figure A.2 shows the CCDF of the weekly views in the tail of the before, at, and after-peak distributions, and the respective analytical lognormal (LN) and power law (PL) fittings.

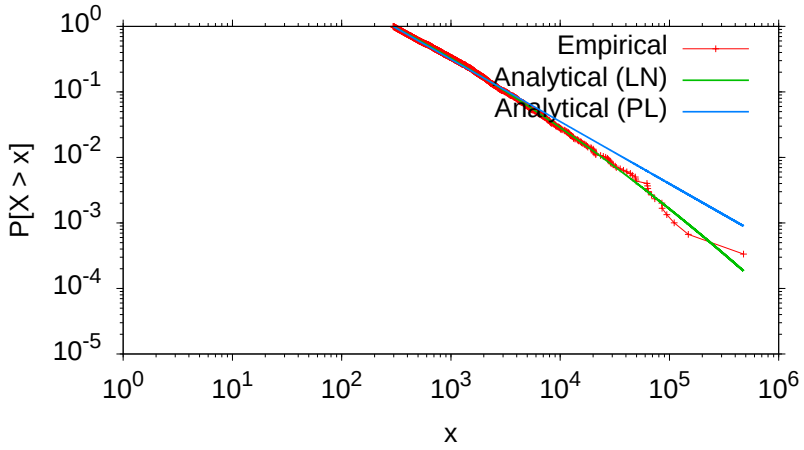
Table A.2: Beta fits for the body of the distributions.

| Phase | Parameters | | | Beta Fits | |
|---|---|---|---|---|---|
| | $n_{body}$ | $x_{min}$ | $x_{thresh}$ | $\alpha$ | $\beta$ |
| Before-peak | 151,051 | 0 | 119 | 0.191 | 1.330 |
| At-peak | 26,805 | 4 | 297 | 0.543 | 2.259 |
| After-peak | 732,075 | 0 | 30 | 0.077 | 0.968 |

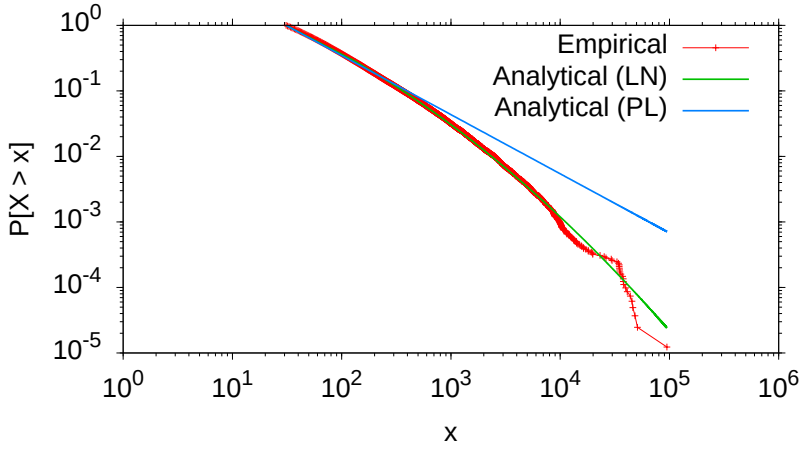
To determine a distribution that best models the body of the the weekly views distribution for each phase, we tried several probability distributions and found the beta distribution provides the best approximation to the empirical data. Since there is no closed-form of the maximum likelihood estimates for the parameters of the beta




(a) Before




(b) At




(c) After

Figure A.2: Power law and lognormal fits for the before, at, and after-peak phase.

distribution, we estimate the shape parameters $\alpha$ and $\beta$, over an interval $[x_{min}, x_{thresh}]$, using the method-of-moments, where:

$$\alpha = \tilde{x} \times (\frac{\tilde{x} \times (1 - \tilde{x})}{v} - 1),$$

$$\beta = (1 - \tilde{x}) \times (\frac{\tilde{x} \times (1 - \tilde{x})}{v}) - 1,$$

with

$$\tilde{x} = \frac{E[x] - x_{min}}{x_{thresh} - x_{min}}$$

and

$$v = \frac{V[x]}{(x_{thresh} - x_{min})^2}.$$

Here, the $E[x]$ is the sample mean and $V[x]$ is the sample variance. The estimated $\alpha$ and $\beta$ parameters results are shown in Table A.2. Each group has $n_{body}$ observations, $x \in [x_{min}, x_{thresh}]$, where $x_{min}$ is the smallest observation in the dataset and $x_{thresh}$ is equal to the threshold separating the body from the tail.

## A.3 Model Validation

To evaluate the goodness of fit of the synthetic data obtained from our models, in addition to the graphical illustrations presented in Sections 4.4 and 4.5, we present here quantile-quantile (Q-Q) plots. Figure A.3 shows the Q-Q plot for the views during a week. Recall that both the basic model and the extended model generate identical views during a week, and thus this plot is representative of both models. Figure A.4 shows the Q-Q plots for the total views by week 32, for three different values of $g$, including $g = 1$ (basic model), $g = 12$ (extended model), and $g = \infty$ (extended model).

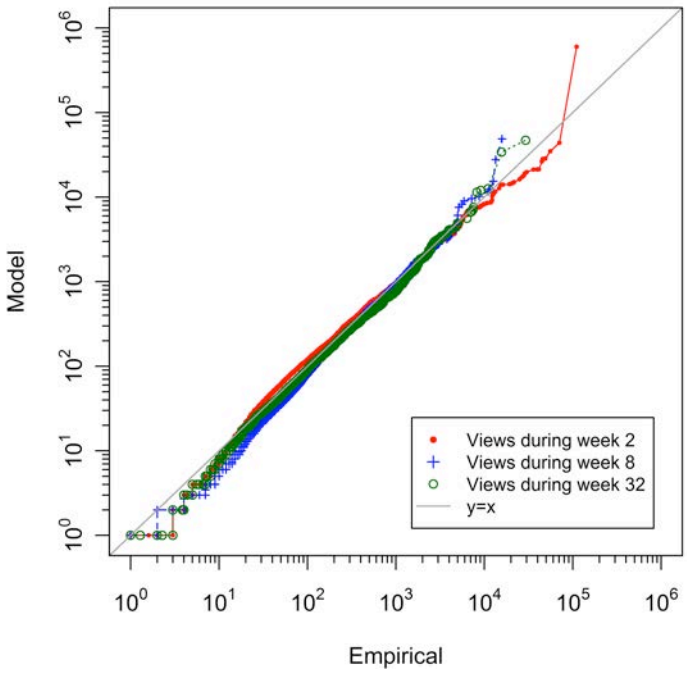



Figure A.3: Q-Q plot for the views during a week from the model and the recently-uploaded dataset.

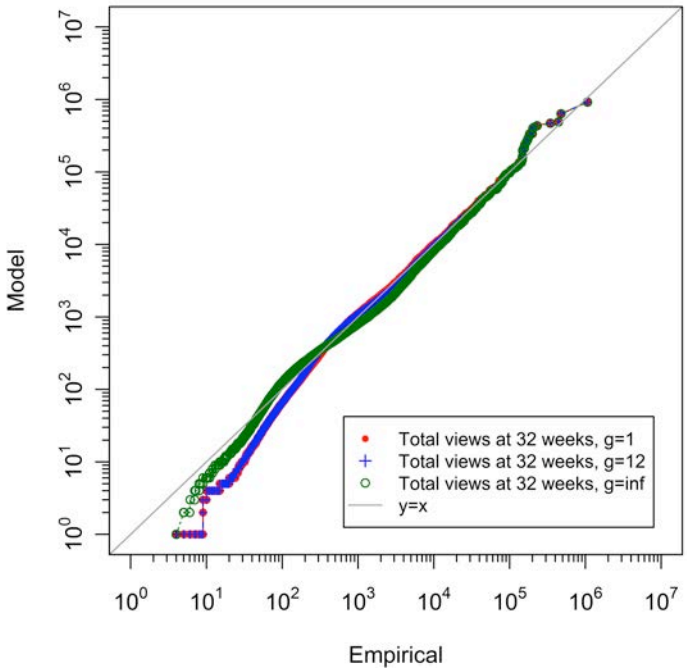



Figure A.4: Q-Q plot for the total views from the model and the recently-uploaded dataset.

In general, the Q-Q plots show that the models are able to to generate synthetic data matching the empirical views distributions. A good match is observed for the body and tail of the distributions. We observe some biases in the head of the distributions; however, these are for the less popular videos and we did not focus on accurately modeling these videos.

# Bibliography


[1] My Language Exchange, http://www.mylanguageexchange.com.

[2] Wikipedia Interactive Statistics, http://www.wikistatistics.net/, February 2012.

[3] Herv Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[4] Abdolreza Abhari and Mojgan Soraya. Workload generation for youtube. *Multimedia Tools and Applications*, 46(1):91–118, January 2010.

[5] A. Allen. *Probability, Statistics, and Queuing Theory with Computer Science Applications*. Academic Press, 1990.

[6] Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

[7] David Applegate, Aaron Archer, Vijay Gopalakrishnan, Seungjoon Lee, and K. K. Ramakrishnan. Optimal content placement for a large-scale vod system. In *Proc. of ACM Co-NEXT*, pages 4:1–4:12, Philadelphia, Pennsylvania, 2010.

[8] S. Asur, R. Bandari, and B. Huberman. The pulse of news in social media: Forecasting popularity. In *Association for the Advancement of Artificial Intelligence*, 2012.

[9] Zlatka Avramova, Sabine Wittevrongel, Herwig Bruneel, and Danny De Vleeschauwer. Analysis and modeling of video popularity evolution in various online video content systems: Power-law versus exponential decay. In *Proceedings of the 2009 First International Conference on Evolving Internet*, INTERNET '09, pages 95–100, Washington, DC, USA, 2009. IEEE Computer Society.

[10] A. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, October 1999.

[11] Fabrício Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, and Keith Ross. Video interactions in online video social networks. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5:30:1–30:25, November 2009.

[12] J.I. Biel. Please, subscribe to me! analysing the structure and dynamics of the youtube network. *Unpublished manuscript. Available at: http://lanoswww. epfl. ch/studinfo/courses/Dynamical_ Networks/Miniprojects_09/Joan_Isaac_Biel_Tresp/REP_biel. pdf (accessed 1 October 2010)*, 2009.

[13] Ignite Social Media Blog. 47 Outrageous Viral Marketing Examples over the Last Decade, http://www.ignitesocialmedia.com/social-media-examples/viral-marketing-examples/, June 2009.

[14] YouTube Official Blog. http://youtube-global.blogspot.com/2012/01/holy-nyans-60-hours-per-minute-and-4.html, January 2012.

[15] Youmna Borghol, Sebastien Ardon, Niklass Carlsson, Derek Eager, and Anirban Mahanti. The untold story of the clones: Content-agnostic factors that impact youtube video popularity. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2012.

[16] Borghol, Youmna and Ardon, Sebastien and Carlsson, Niklass and Eager, Derek, and Mahanti, Anirban. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*, 68(11):1037 – 1055, 2011.

[17] Tom Broxton, Yannet Interian, Jon Vaver, and Mirjam Wattenhofer. Catching a viral video. In *ICDM Workshops*, pages 296–304, Sydney, Australia, December 2010.

[18] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. ACM Internet Measurement Conference (IMC)*, pages 1–14, San Deigo, USA, October 2007.

[19] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and Sue Moon. Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems. *ACM/IEEE Transactions on Networking (ToN)*, 17(5):1357–1370, October 2009.

[20] Gloria Chatzopoulou, Cheng Sheng, and Michalis Faloutsos. A first step towards understanding popularity in youtube. In *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, pages 1–6. IEEE, March 2010.

[21] X. Cheng, C. Dale, and J. Lui. Statistics and Social Network of YouTube Videos. In *Proc. International Workshop on Quality of Service (IWQoS)*, pages 229 – 238, Enskede, The Netherlands, June 2008.

[22] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Understanding the characteristics of internet short video sharing: Youtube as a case study, 2007.

[23] Xu Cheng, Kunfeng Lai, Dan Wang, and Jiangchuan Liu. Ugc video sharing: Measurement and analysis. In Chang Chen, Zhu Li, and Shiguo Lian, editors,

*Intelligent Multimedia Communication: Techniques and Applications*, volume 280 of *Studies in Computational Intelligence*, pages 367–402. Springer Berlin / Heidelberg, 2010.

[24] A. Clauset, C. Shalizi, and M. Newman. Power-law Distributions in Empirical Data. *SIAM Rev.*, 51(4):661–703, November 2009.

[25] CNN. Digital fundraising still pushing Haiti relief, http://articles.cnn.com/2010-01-15/tech/online.donations.haiti_1_earthquake-haiti-haiti-relief-twitter-and-facebook?_s=PM:TECH, January 2010.

[26] comScore. More than 200 Billion Online Videos Viewed Globally in October, http://www.comscore.com/Press_Events/Press_Releases/2011/12/More_than_200_Billion_Online_Videos_Viewed_Globally_in_October, December 2011.

[27] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci.*, 105(41):15649–15653, October 2008.

[28] comScore DanPiech. Online Video by the Numbers, http://www.comscore.com, July 2011.

[29] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proc. of the fourth ACM international conference on Web search and data mining*, pages 745–754, Hong Kong, China, 2011.

[30] P. Gill, M Arlitt, Z Li, and A. Mahanti. YouTube Traffic Characterization: A View from the Edge. In *Proc. ACM Internet Measurement Conference (IMC)*, pages 15–28, San Deigo, CA, October 2007.

[31] Ulrike Gretzel and Kyung Hyan Yoo. Use and impact of online travel reviews. In Peter OConnor, Wolfram Hpken, and Ulrike Gretzel, editors, *Information and Communication Technologies in Tourism 2008*, pages 35–46. Springer Vienna, 2008.

[32] Lev Grossman. You Yes, You Are TIME's Person of the Year, http://www.time.com/time/magazine/article/0,9171,1570810,00.html, December 2006.

[33] Majed Haddad, Eitan Altman, Rachid El-Azouzi, and Tania Jimenez. A survey on youtube streaming service. In *Proceedings of VALUETOOLS 2011 - 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, Paris, France, May 16-20 2011.

[34] M. Halvey and M. Keane. Analysis of Online Video Search and Sharing. In *Proc. ACM Hypertext and Hypermedia Conference*, pages 217–226, Manchester, UK, September 2007.

[35] M. Halvey and M.Keane. Exploring social dynamics in online media sharing. In *Proc. International Conference on World Wide Web (WWW)*, pages 1273–1274, Banff, Canada, May 2007.

[36] The Sydney Morning Herald. Facebook sells shares for $US38 in landmark IPO, http://www.smh.com.au/business/world-business/facebook-sells-shares-for-us38-in-landmark-ipo-20120518-1yu7a.html, May 2012.

[37] Pew Internet. Health Topics: 80% of Internet Users Gather Health Information Online, http://pewinternet.org/~/media/files/reports/2011/pip_healthtopics.pdf, February 2011.

[38] Salman Jamali and Huzefa Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. *Web Information Systems and Mining, International Conference on*, 0:32–38, 2009.

[39] Andreas Kaltenbrunner, Vicenc Gomez, and Vicente Lopez. Description and prediction of slashdot activity. In *Proceedings of the 2007 Latin American Web Conference*, LA-WEB '07, pages 57–66, Washington, DC, USA, 2007. IEEE Computer Society.

[40] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, January 2010.

[41] Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho. Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In *Proceedings of the 2011 IEEE 11th International Conference on Computer and Information Technology*, CIT '11, pages 449–454, Washington, DC, USA, 2011. IEEE Computer Society.

[42] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, 85:4629–4632, Nov 2000.

[43] Jong Lee, Sue Moon, and Kave Salamatian. An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors. In *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, London, Canada, 2010.

[44] Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 621–630, New York, NY, USA, 2010. ACM.

[45] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. In *Proceedings of the 7th ACM conference on Electronic commerce*, EC '06, pages 228–237, New York, NY, USA, 2006. ACM.

[46] R. H. Lindeman, P. F. Merenda, and R. Z. Gold. *Introduction to Bivariate and Multivariate Analysis*. Scott Foresman & Co, 1980.

[47] J Miguns, R Baggio, and C Costa. Social media and tourism destinations: Tripadvisor case study. *Tourism*, 2008:1–6.

[48] Alan Mislove, Massimiliano Marcon, Krishna P. Gummandi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. ACM Internet Measurement Conference (IMC)*, pages 29–42, San Diego, CA, October 2007.

[49] Siddharth Mitra, Mayank Agrawal, Amit Yadav, Niklas Carlsson, Derek Eager, and Anirban Mahanti. Characterizing Web-based Video Sharing Workloads. *ACM Transactions on the Web*, (2):8:1–8:27, May 2011.

[50] Evgeny Morozov. *The net delusion : the dark side of Internet freedom*. PublicAffairs, Dec 2011.

[51] M. Newman. Power laws, Pareto distributions and Zipf's Law. *Contemporary Physics*, 46(5):323–351, September 2005.

[52] Bits Blog NYTimes.com. How Obamas Internet Campaign Changed Politics, http://bits.blogs.nytimes.com/2008/11/07/how-obamas-internet-campaign-changed-politics/, November 2008.

[53] Symeon Papadopoulos, Athena Vakali, and Ioannis Kompatsiaris. The dynamics of content popularity in social media. *IJDWM*, 6(1):20–37, 2010.

[54] Louis Plissonneau, Taoufik En-Najjary, and Guillaume Urvoy-Keller. Revisiting web traffic from a dsl provider perspective : the case of youtube. 2008.

[55] The Washington Post. YouTube: The future of entertainment is on the Web, http://www.washingtonpost.com/business/technology/youtube-the-future-of-entertainment-is-on-the-web/2012/01/12/gIQADpdBuP_story.html, January 2012.

[56] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. Characterizing and Modeling the Dynamics of Online Popularity. *Phys. Rev. Lett.*, 105(15):158701, Oct 2010.

[57] Daniel M. Romero, Brendan Meeder, and Jon M. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*, pages 695–704, 2011.

[58] M J. Salganik and D.J Watts. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly*, 71:338–355, 2008.

[59] Matthew J Salganik, Peter Sheridan Dodds, and Duncan Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(10):854–856, February 2006.

[60] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, 2010.

[61] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 67:1–67:8, New York, NY, USA, 2011. ACM.

[62] The Telegraph. Web could collapse as video demand soars, http://www.telegraph.co.uk/news/uknews/1584230/Web-could-collapse-as-video-demand-soars.html, April 2008.

[63] Social Times. Cisco Predicts That 90% Of All Internet Traffic Will Be Video In The Next Three Years, http://socialtimes.com/cisco-predicts-that-90-of-all-internet-traffic-will-be-video-in-the-next-thre b82819, November 2011.

[64] The New York Times. YouTube Plans to Make Big Bet on New Online Channels, http://www.nytimes.com/2011/10/29/business/media/youtube-plans-to-create-new-online-channels.html, October 2011.

[65] The New York Times. Googles Results Disappoint, Sending Shares Down, http://www.nytimes.com/2012/01/20/technology/googles-strong-results-less-than-expected.html, January 2012.

[66] USA Today. YouTube Serves up 100 million Videos a Day Online, July 2006.

[67] F. Wu and B. A. Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci. USA*, 104(45):17599–17601, 2007.

[68] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.

[69] YouTube. http://code.google.com/apis/youtube/overview.html.

[70] Jia Zhou, Yanhua Li, Vijay Kumar Adhikari, and Zhi-Li Zhang. Counting youtube videos via random prefix sampling. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 371–380. ACM, 2011.

[71] R. Zhou, S. Khemmarat, and L. Gao. The Impact of YouTube Recommendation System on Video Views. In *Proc. ACM Internet Measurement Conference (IMC)*, pages 404–410, Melbourne, Australia, November 2010.

[72] Feng Zhu and Xiaoquan (Michael) Zhang. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74(2):133–148, 2010.

[73] M. Zink, K. Suh, and J. Kurose. Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications. In *Proc. SPIE Multimedia Computing and Networking (MMCN) Conference*, San Jose, CA, January 2008.

# Publications

1. Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. The untold story of the clones: Content-agnostic factors that impact youtube video popularity. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2012.

2. Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager and A. Mahanti. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*, 68(11):1037 - 1055, 2011.

3. Y. Borghol, S. Ardon, N. Carlsson and A. Mahanti. Toward Efficient On-Demand Streaming with BitTorrent. In *Proceedings of the 9th IFIP TC 6 international conference on Networking*, August 2010.

4. F. Tan, S. Ardon, Y. Borghol, EMO: A Statistical Encounter-based Mobility MOdel for Simulating Delay Tolerant Networks. In *Proceedings of 9th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2008)*, June 2008.