CHARACTERIZATION OF SOCIAL VIDEO

by

JEFFREY R. OSTROWSKI

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2008

MAJOR: COMPUTER ENGINEERING

Approved By:

_____

Advisor Date

DEDICATION

To my family and friends who have supported me throughout this endeavor.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION

Interest in social media has grown dramatically across the Internet and continues to evolve. Instant messaging, online gaming, photo sharing, video sharing, social networking, and other online social interaction have all been ways that technology has connected people and increased avenues of communication. Photo sharing websites such as Flickr have recently gained popularity since their launch in the last several years [6], while massively multiplayer online role-playing games such as World of Warcraft have steadily increased revenues [9]. Blogging also continues to show growth [22], while social networking websites such as MySpace boast of more than 10 million users [16].

People have embraced video sharing most recently to entertain, inform, and teach. Webcams, cell phones, cameras, and a multitude of portable devices have enabled people to capture information in the form of video and share it with the world instantaneously. This is further underscored by the U.S. release of the iPhone, for example, in June 2007 which included YouTube as a built-in, marketed feature [3] - not just a potential capability. Instead of watching a favorite TV channel for a limited set of weekly programming, users of sites such as YouTube and Google Video are able to view as well as post thousands of videos for public view at no charge. These websites, and others like Metacafe, Yahoo! Video, and MySpace, have grown drastically in popularity. YouTube, for example, is now the third most-viewed website in the world according to Alexa Web Ranking [1].

What emerges as a result of social video programming is a medium much different from normal broadcast television. Compared to the cost of producing and airing programs on local television stations or cable networks, the price to create video content and have it published is significantly reduced. The distributor and the user are no longer restricted to vastly separate roles, as might be correlated to a large network like the American Broadcasting Company (ABC) and a stay-at-home mom. Now, the stay-at-home mom is able to shoot her own programming using a small digital recorder, such as a webcam or

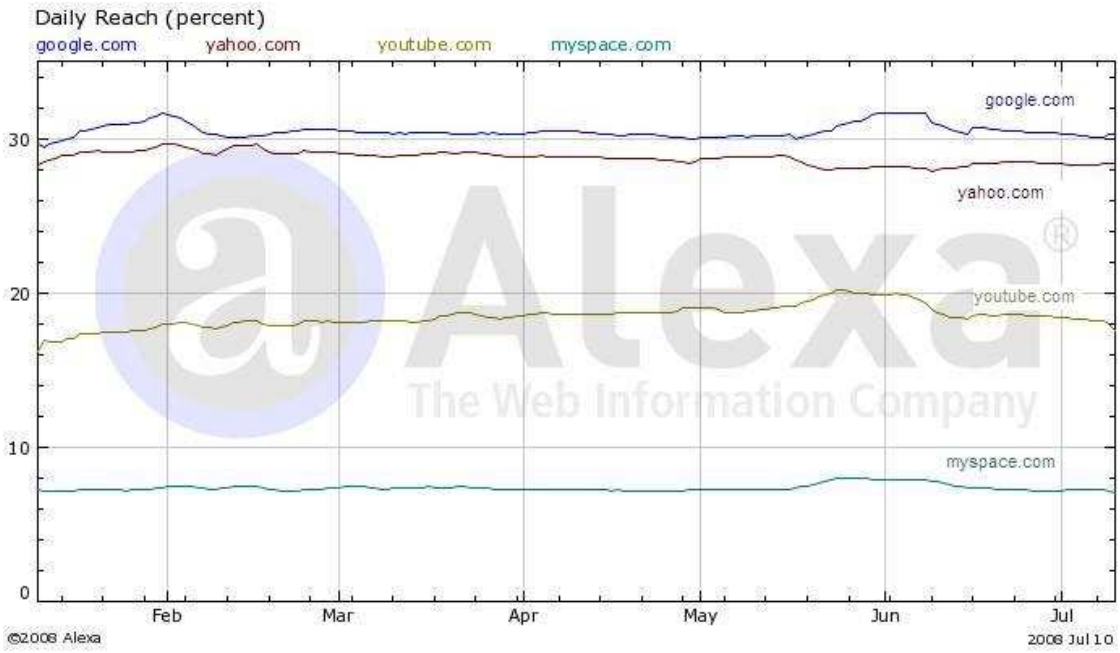


Figure 1: Daily Reach for Top Global Websites (Source: Alexa.com)

cell phone, and publish it instantaneously at minimal cost. She does not need to incur the syndication costs of a large broadcast network like ABC in order to reach a wide audience, as hosting is free through a personalized channel on YouTube. While ABC may have restricted the broadcast audience to only viewers in the U.S. at a particular time of the day, YouTube converts the content and publishes it for view to anyone who has an Internet connection, for view at any time of the day in any region of the world. Recent studies have concluded that, similar to normal broadcast television, the number of viewers of social video programming far outweigh the number of publishers [25].

This thesis investigates online social video content with a global perspective. In particular, it characterizes videos from YouTube, Google Video, and the Truveo search engine in terms of popularity distribution in daily, weekly, monthly, and yearly contexts over a ten month period, more than any other social video study of which we are aware. It also studies the correlation between video categories and popularity rankings and how the video access changes with the day of the week and the time of the day over an eight week pe-

riod. In addition, this thesis examines how the video access evolves over time. Moreover, it analyzes channel and video format encoding popularities. Furthermore, it analyzes the social aspects and how video category popularities vary depending on the users' world region. We approach these characterizations with the widest lens possible, not focusing only on a specific user group, category, or region, but rather including the entire compilation of users who participate in social video programming each day. The characterization of social media helps in understanding user behavior and the regional/social influence, as well as the design of delivery systems.

Very few studies have focused their study on social video. The study [28] focused on YouTube video clip popularity, payload, and data rates to determine that client-based caching can reduce network traffic and allow faster access to YouTube videos. It sought to optimize network traffic on a local network, which is different from our global approach, and did not analyze other social media sites. It also did not use popularity metrics directly from YouTube, which we do through the use of the available YouTube data services. The study [18] presented the popularity characteristics of YouTube and Daum, including the exhibition of power-law in the evolution of popularity. The data in that study were limited to "Entertainment" and "Science & Technology" videos over a period of several days. In contrast, we consider all categories and analyze YouTube popularity for a period of ten months. A different user-level approach to analysis of YouTube was taken in [23]. That study compared YouTube data with traditional website user sessions using duration, inter-transaction times, and content types, among other methods. User data were taken from users on a local area network as well, not characterizing all YouTube users as done in our study. In [21], a geographical approach to YouTube characterization was carried out in Latin America to identify the locality space for that region in comparison to the United States and the rest of the world. In our approach, we take a global view by presenting Google Video data from each world region. We further study the types of videos that are watched, not only in Latin America but all regions, to gain insight into categorical trends.

The main findings can be summarized as follows.

- The popularity of social videos on YouTube data is characterized by a Zipf-like distribution with $t = 0.30$ for daily popularity distributions and $t \approx 0.50$ for weekly, monthly, and yearly popularity distributions. This is contrast to [24], which showed that the stretched exponential distribution is a better fit for streaming media.

- Music, comedic, and entertainment videos are the the most popular among all categories.

- Most social videos published are less than five minutes in length, with the most popular videos averaging 3.5 to 4 minutes in length.

- Video access frequency over time is characterized by an oscillating curve as popularity varies per day of the week.

- Fridays and Saturdays from 9pm - 1pm UTC are the most popular times for users to access social videos.

- YouTube is, by far, the most popular channel for accessing online social video, and the Flash video format is now available for over 80% of all videos published.

- The category of music dominates social video popularity in Latin America and the Carribean, Northern America, and Western Europe. Geographical data we uncover is a starting point for understanding regional influence on social video selection.

The rest of the thesis is organized as follows. In Chapter 2, we provide an overview of social media. In Chapter 3, we review previous related work. Chapter 4 provides our evaluation methodology. In Chapter 5, we present and analyze the data in terms of popularity distribution, categories, days of the week and evolution, length, category comparisons, and social aspects.

CHAPTER 2
OVERVIEW OF SOCIAL MEDIA

Social media is a type of social interaction in which a group of people uses shared technology to create words, pictures, audio, video, or other means to communicate ideas, facts, understanding, and perspective. There are several types of social media today including multimedia sharing, social networking, social gaming, and social collaboration. An explanation of each of these types is now further discussed.

2.1 Multimedia Sharing

Multimedia sharing is a very popular type of social media. It is a method by which users are able to publish photos, videos, audio clips, and music for view and feedback by other users. The most popular website devoted to video sharing today is YouTube, yet other websites such as Google Video, the French Dailymotion.com, and Korean Daum.net are popular in different regions of the world [1]. Flickr remains the most popular website devoted to photo sharing, while imeem.com, more commonly referred to as a social networking website, is very well known for its music sharing [1]. Most recently, multimedia sharing has extended itself to live video sharing on websites including Justin.tv and Ustream.tv. These websites have brought forward a very distributed form of broadcasting, and also a method by which individuals can continously stream portions of their life in real time.

2.2 Social Networking

Social networking is a type of social media in which people are connected through common interests and explore the interests of others. On social networking websites, users are often required to create an online account, publish a profile, and approve access for others to view their content. While social networking websites like MySpace originally were limited to email and web page customization, social networking websites like MyS-

pace and Facebook now allow users to communicate through email, blogging, short status messages, multimedia sharing, and even small games and applications. Social networking websites serve communities of friends, professionals [8], social action groups [10], and other common bonds that bring people together. As of September 2008, Facebook is now the most popular social networking website, overtaking the popularity dominance of MySpace earlier this year [1].

2.3 Social Gaming

Social gaming means playing games while interacting with people rather than in solitude. It has actually been around for a long time in the form of multiplayer games such as Monopoly, Trivial Pursuit, and other board games and the like which involve several people. This social gaming phenomenon, however, has further made its way to the Internet and online gaming. World of Warcraft, for example, is known as a massively multiplayer online game in which a virtual environment is simultaneously occupied by thousands of players [27]. There are now many games released for video game systems which have an online multiplayer mode, such as Halo 3. Further social gaming is being implemented by social networking websites, such as Facebook, which allow users with common interests another method of social interaction.

2.4 Social Collaboration

A widespread type of social media, social collaboration is a broad term which refers to the exchange of information between people which often leads to common understandings and meanings. Wikis are one form of social collaboration in which many people may author and publish content to using a web browser [26]. Sometimes reviewed before final publication, the content is published as common knowledge. This aggregation of content and reuse of knowledge often becomes very popular, such as with Wikipedia [13] which is now the seventh most popular website in the world according to Alexa [1]. Blogs, a very

popular method of social collaboration, are another way in which people publish content, engage in conversations, and produce knowledge. With blogs dedicated to thousands of different topics - from technical support issues on technology websites to comments in response to articles on news websites - individuals publish articles and other individuals may respond to the said articles inline. Blogger, as the most popular blog website today [1], also allows photo posting and mobile access [2]. Social bookmarking and social news websites, such as Delicious [4] and Digg [5], are different derivatives of social collaboration and allow users to discover websites and content submissions through tagging mechanisms and voting. While no social news or social bookmarking website is currently in the top 100, Digg is the most popular at rank 124 [1].

CHAPTER 3
RELATIONSHIP TO PRIOR WORK

Many studies have characterized online and streaming video trends; however, few have focused on social video. We compare our work to both social video studies, as well as online video studies in general, in the following section.

3.1 Social Video Studies

The frequency of user access to videos is one of the main characteristics of social videos. In general, studies have either supported or opposed the notion that user access to videos follows a Zipf-like distribution. In Zipf-like distribution, the probability of accessing video with rank v_r is

$$P(v_r) \propto 1/(v_r^{1-t}). \tag{1}$$

where t is a constant less than one.

When analyzing social videos, study [21] presented data showing that the number of views when compared to rank did not exhibit Zipf's law for the data subset collected, in contrast with our findings. Further, the difference between that study and ours is that it took an aggregated snapshot of YouTube data over a period of eleven days. In our study, we collected data for the most popular videos each day for over ten months, allowing us to take a consistent view count average over a much longer period of time. Study [18] claimed adherence to the Zipf-like distribution; however, we believe its inclusion of only two categories ("Entertainment" and "Science & Technology") over several days is limited. In our work, we show how videos of each available category vary in popularity within a ten month period.

One characteristic of social video that has not been widely studied is category. As we mentioned previously, study [18] limited its scope in comparison to a full categorical analysis of social video. Even though study [21] stated that it had collected video length

data, it did not present that data for analysis. In our study, we take categorical analysis further by presenting results for all social video categories provided by YouTube, Truveo Search, and Google Video data.

The access evolution and length of a video are also important characteristics. Social video access evolution has recently been studied in [23] over a period of a day and in [18] over a period of a few days, which are much shorter than our eight-week analysis of access evolution. Length is an important metric to understand when determining cache and bandwidth for serving social videos. The study [18] only mentioned two average lengths from its analysis: 30 seconds for the Daum Commerical Film category and 203 seconds for the Daum Music Video category. It did not correlate this data to any further analysis and did not show trends. Study [21] stated that 80% of the videos recorded in its analysis were less than five minutes in length, again showing no further trends as we do in our analysis.

We further contend that a geographical approach is necessary in modern social video characterization. In paper [21], a geographical approach to YouTube characterization was carried out in Latin America to identify the locality space for that region in comparison to the United States and the rest of the world. That study is an attempt to understand geographical factors in video services and how regional differences influences social video behavior. It is the first to recognize that social video trends may indeed be influenced by social aspects practiced within a given region. We expand this approach in our study, as we present social video data from each world region. We further study the types of videos that are watched in each of these regions to gain insight into categorical trends.

3.2 General Online Video Studies

Let us now compare our study with other online video studies. Study [15] attempted to disprove that user accesses to online video followed the Zipf-like law. The environment for the study was a university network, with videos types characterized by both general

entertainment videos (e.g., movies) and class lectures. The study distributed videos over a high-bandwidth network, removing network latency as a major factor for user behavior. Another study [20] characterized streaming media workload of streaming servers, including stream merging, server popularity, and session characteristics. That study analyzed client-based streaming media at the university and compared its characteristics with more traditional Web pages and HTTP protocols. It determined that user access frequency did follow a Zipf-like distribution, with $t = 0.47$, for media access from the university to the public Internet. Additional work [17] contributed to the debate when extensive client workloads were examined on two media servers at the University of Wisconsin-Madison and the University of Saskatchewan for classroom lecture videos and other course content. That study showed that user access frequencies on each server were modeled by the concatenation of two Zipf-like distributions. Taking the analysis to the corporate workplace, paper [19] analyzed enterprise media server workloads for Hewlett-Packard Corporation in 2002 and correlated Zipf-like distribution with monthly video popularity metrics. More recently, the study [24] suggests that a stretched exponential distribution is best to characterize video popularity. In our study, we show, however, that the Zipf-like distribution characterizes the popularity of the top YouTube and Truveo Search videos.

There has been minimal general online study regarding video categories. The study [15] analyzed online video distinctions between educational lectures and movies. It found that access to movies, for example, tends to be evenly distributed over time, while access to educational videos exhibits very high access rates over small periods. In our study, we present results for all categories provided by YouTube, Truveo Search, and Google Video data.

Comparing our study to general online video evolution studies, the study [15] showed a cyclic pattern of access for videos which increases over time. Evolution analysis in [19] defined the life duration of a media file to be the time between the first and last accesses of the file in a given workload. That study further showed that for an enterprise media

server, more than 50% of media file accesses occur in the first week of the file's existence. The paper [24] stated that 50% of requests were for objects older than 150 days. Our study focuses on social video data and takes a different approach, first identifying the most popular videos on social video websites and then tracking their progress over time by looking for cyclic patterns and life duration characteristics.

Looking toward a more in-depth online video length analysis, study [20] showed that the most common media streams were 3.5 to 4.5 minutes in length, leading to the conclusion that clients have a strong preference for viewing short multimedia streams. The study [19] further concluded that even though a proportional number of videos are accessed in an enterprise over the spectrum of video lengths, the duration for a majority of the accesses ($77\% - 79\%$) is less than 10 minutes in length. In our further study, we extend length distribution studies by presenting data for YouTube and videos indexed by Truveo Search, including video length distribution data by category.

Building on the above work, we aim to characterize the video popularity distribution for social video websites, the types of videos which are accessed, how videos are accessed over a period of time, length metrics, and geographical influences. The summary of these aspects will help to provide insight into social video patterns.

CHAPTER 4
EVALUATION METHODOLOGY

We collected data from two social video websites and one video search engine: YouTube, Google Video, and the AOL-owned Truveo Video Search [11]. We created scripts using the Ruby programming language, including scripts for the Google and YouTube developer interfaces [7, 14]. Data were closely monitored and changes to the scripts were made to accommodate Google and YouTube website modifications over the collection period.

YouTube data was collected for a period of forty-two weeks. Daily, weekly, monthly, and yearly Top 100 video feeds from YouTube were successfully downloaded each day, including video ID, title, duration, view count, category, and URL. We then took a snapshot of the top 100 most viewed YouTube videos of all time and tracked their popularity evolution over an eight-week period. We collected video IDs, titles, durations, view counts, category, and URL information for these videos as well twice a day at 9 PM Coordinated Universal Time (UTC) and 1 PM UTC. Once the data collection was complete, we calculated the daily view count for each video in our top 100 list, as well as the view count statistics for the periods of 1PM to 9PM UTC and 9PM to 1PM UTC each day.

Expanding our data collection, we then used data from the Truveo Video Search engine to understand the effect of a much larger set of online social videos. The top 1000 videos of all time were collected for a period of forty-two weeks. Video popularity, category, length, and format metadata were extracted for these videos and normalized.

Lastly, we took a snapshot of the top one hundred most popular daily videos from Google Video. Not only did we collect the top 100 videos for the world (“All Countries & Regions”) but also we captured the top 100 videos for each country and region for which Google Video provided a top 100 list (Australia, Finland, Peru, etc.). After collecting the top 100 videos and their associated rank in the top one hundred list for each country and region, we then proceeded to find the associated category of those videos. As Google Video incorporates YouTube, Google, and other online video websites into its lists, we had

to come up with an intelligent strategy to get the category data. First, we parsed the top hundred list to determine if each video listed was a YouTube video. We did this by making a call directly to the URL of the Google Video ID, which soon informed us after some HTML parsing whether or not this was a YouTube video. We then were able to search and cross-reference the listed Google Video identified with the YouTube identifier that was parsed. From this YouTube identifier, we were able to appropriately search YouTube for the category information per our previous interface addition. Since the category information for Google videos was not readily posted on Google's website, we needed to find a way to get this information.

The Google Advanced Video Search engine allowed us to query videos for a particular category using the Google identifier. In order to determine whether a video was in one of the thirty-nine Google Video categories, however, we would have to search up to thirty-nine categories using the search engine until we found the category to which the video belonged. This could feasibly equate to more than 100,000 Google searches for all of the downloaded videos, over 175,000 searches for the worst case scenario. At the end of the automated exercise, we had only a handful of videos for which we needed to look up the category manually.

The retrieved video lists now needed to be related to one another based on country. In order to account for these geographic differences, we acquired the regional groupings of countries from the U.S. Census Bureau (Asia excluding Near East, Commonwealth of Independent States, Eastern Europe, Latin America and the Caribbean, Near East, Northern America, Oceania, Sub-Saharan Africa) and weighted the results we collected for each country based on population for the region from the Total Midyear Population from U.S. Census Bureau International Data Base [12]. A sample of the data collected from the U.S. Census Bureau is shown in Table 1. If Country X had double the population of Country Y, for example, that category information for a particular rank within the Country X data would then be weighted double in comparison to the category information for

Country or Area	Population
Afghanistan	31,889,923
Albania	3,600,523
Algeria	33,362,742
American Samoa	64,025
Andorra	80,757
Angola	12,263,596
Anguilla	13,779
Antigua and Barbuda	83,425
Argentina	40,048,816
Armenia	2,971,650
Aruba	100,018
Australia	20,749,625

Table 1: Sample of 2007 Total Midyear Population Data - U.S. Census Bureau

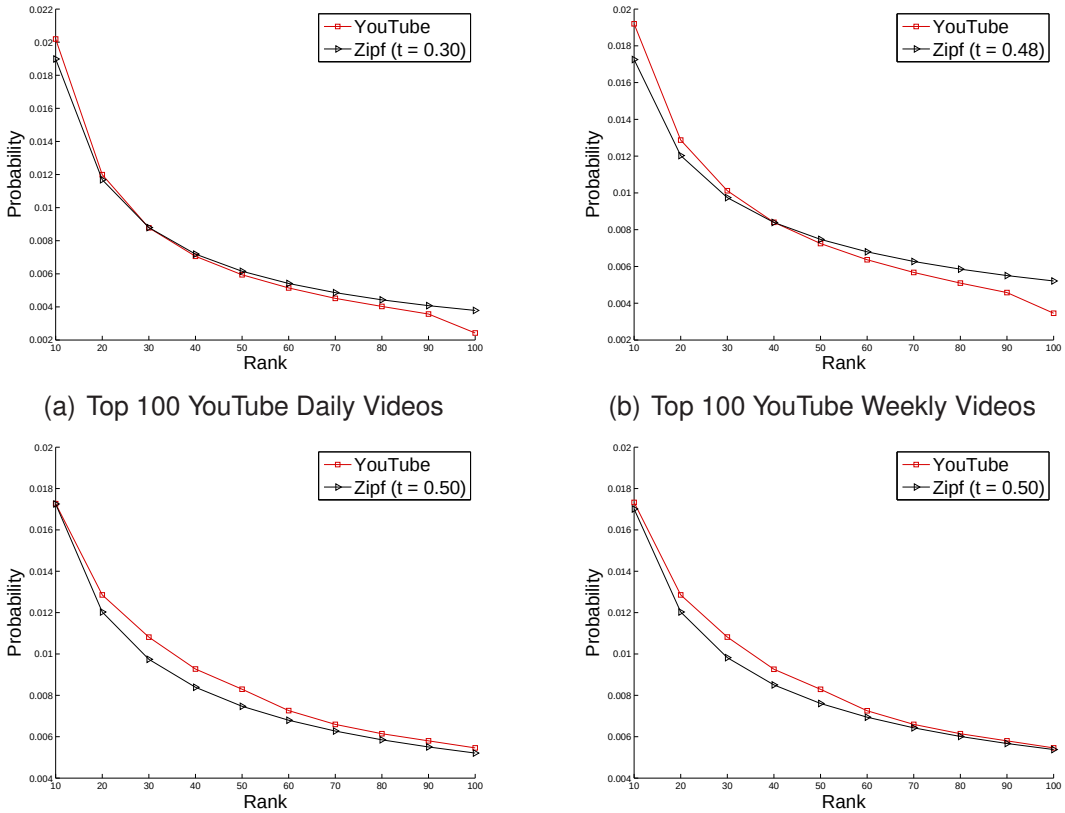
a particular rank for Country Y. This normalized the results between countries so that the population of the region was better represented. Our next issue, however, was to figure out how to ensure rank data was accurately represented when merged. In order to maintain the rank contribution, we also weighted each value further by the probability density function for each particular rank.

CHAPTER 5

DATA PRESENTATION AND ANALYSIS

5.1 Popularity Distribution

In accordance with previous studies, we seek to characterize video popularity with the YouTube data that we have collected. Figure 2 shows the probability density function of video popularity in terms of access frequency (number of views) for the Top 100 YouTube Videos lists (daily, weekly, monthly, and yearly). As can be seen by the figures, the distribution for all popularity periods follow Zipf-like curves.



(a) Top 100 YouTube Daily Videos

(b) Top 100 YouTube Weekly Videos

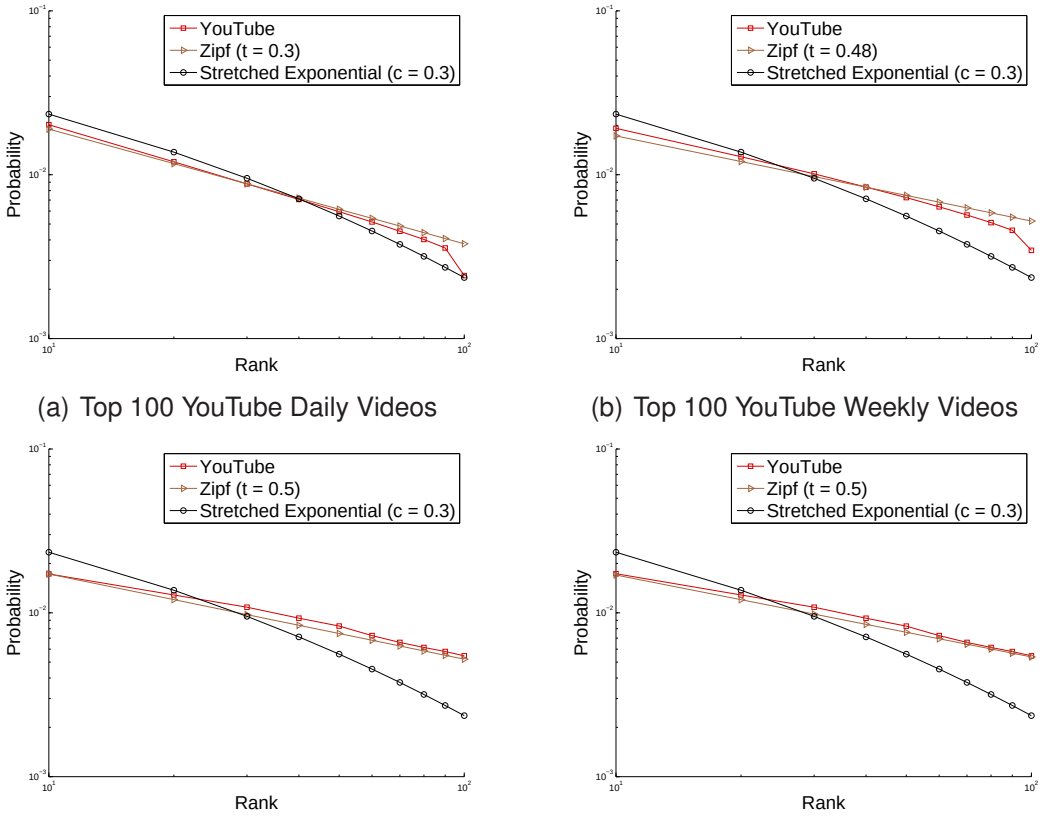
(c) Top 100 YouTube Monthly Videos

(d) Top 100 YouTube Yearly Videos

Figure 2: Video Probability Distributions [Average over 42-Week Period]

Daily videos show very close similarity to the Zipf distribution with $t = 0.30$, while all other popularity periods (weekly, monthly, and yearly) show t near or equivalent to 0.50. This

difference in t has not been previously been uncovered by any other study. In study [20], a week's worth of access frequency data showed that media requests over the public Internet produced Zipf-like results as well with $t = 0.47$, very close to our findings for weekly, monthly, and yearly distributions. Our findings show, however, that Zipf-like distributions for YouTube access frequency may be different based on the specified reference period. Daily access frequency for the top 100 videos must be characterized differently than weekly, monthly, and yearly videos. This characterization of daily YouTube videos is very powerful, as it allows us to determine the number of views a top video might receive during a day for capacity planning as YouTube site traffic is accumulated. Further, we confirm that weekly, monthly, and yearly access frequency on YouTube can be approximated to calculate capacity planning over a longer period of time.



(a) Top 100 YouTube Daily Videos

(b) Top 100 YouTube Weekly Videos

(c) Top 100 YouTube Monthly Videos

(d) Top 100 YouTube Yearly Videos

Figure 3: Video Probability Distributions [Average over 42-Week Period]

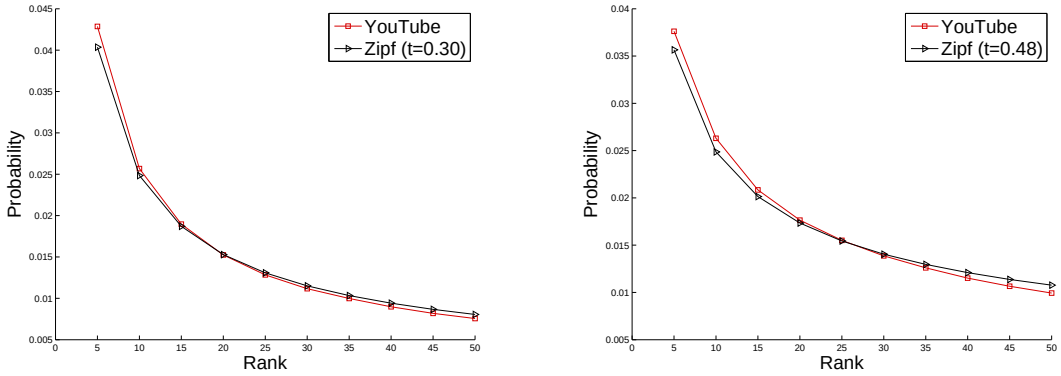
One recent study [24] attempted to show that video access frequency is instead characterized by a stretched exponential distribution. That paper analyzed a host of different web and P2P data sources between 1998 and 2006 and demonstrated that sixteen analyzed workloads can be characterized by the equation

$$P(v_r) \propto e^{-t/{t_0}^c}, \tag{2}$$

where t_0 is the characteristic relaxation time (a constant), and c is a number between 0 and 1. As can be seen from the data plotted on a log scale in Figure 3, the recent YouTube data does not follow stretched exponential tendencies. The closest margin of error between the data and the stretched exponential form is exhibited in Figure 3(a), yet still a Zipf-like curve with $t = 0.30$ is a better fit for the data. We conclude that YouTube access frequency over daily, weekly, monthly, and yearly periods is more readily characterized by a Zipf-like curve and does not follow the stretched exponential distribution suggested in study [24].

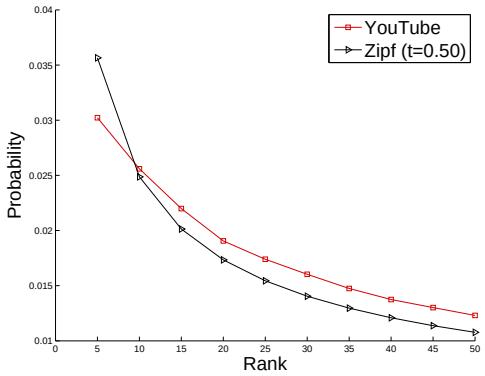
Our further analysis reveals that smaller video populations (such as the top 50 videos) follow a similar trend. In Figure 4, we determine that the Zipf-like distribution ($t = 0.30$ for daily probabilities and $t = 0.50$ for weekly, monthly, and yearly probabilities) is maintained. Daily and weekly access probability is closely consistent with a Zipf distribution. Monthly and yearly popularity averages slightly stray from the Zipf distribution, both more popular for rank 10 through 50, yet being less popular at rank 10 and higher. While variation from the Zipf-like distribution occurs for these videos, the difference between the curves do not support the findings in [24].

In order to further prove our findings with the YouTube data, we look toward a larger video population to understand if Zipf-like behavior is maintained. Figure 5 shows the popularity distribution of the Top 1000 Videos (All Time). The videos follow a Zipf-like distribution with $t = 0.20$, different than the $t = 0.50$ determined with the YouTube data. The reason between the discrepancy between these two values of t may be explained by

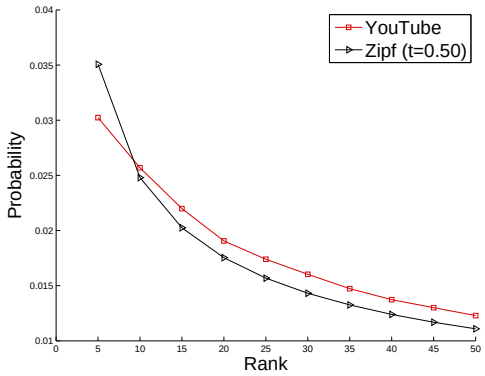


(a) Top 50 YouTube Daily Videos

(b) Top 50 YouTube Weekly Videos



(c) Top 50 YouTube Monthly Videos



(d) Top 50 YouTube Yearly Videos

Figure 4: Video Probability Distributions [Average over 42-Week Period]

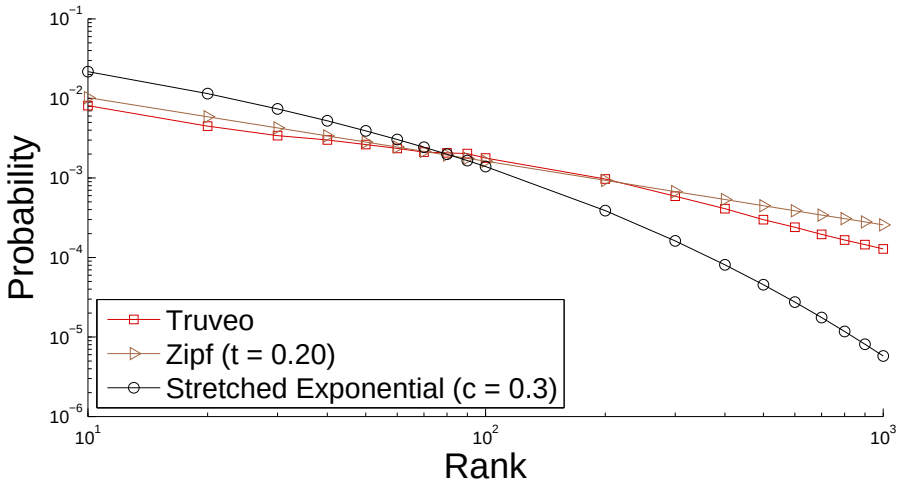


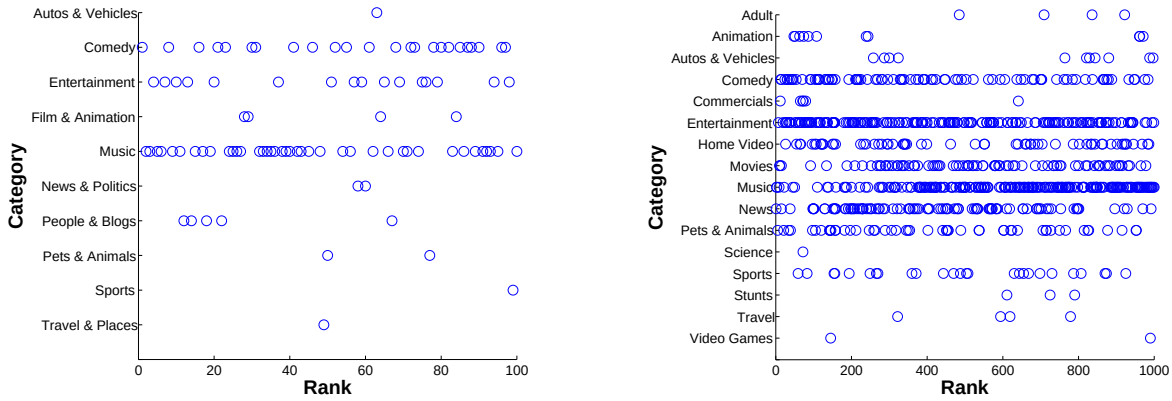
Figure 5: Truveo Video Search Probability Distribution

the contributions of the Truveo Search algorithm and the additional online videos it may index. As this data was collected from the Truveo Search engine, it is possible that the

Truveo Search engine results generated by individuals using the service could affect the value of t. While further test methods and data analysis would be required to prove this, it is clear that the Zipf-like distribution still explains the behavior for the access frequency of videos acquired through Truveo Search engine data. In contrast to study [24], Figure 5 illustrates that the data video access frequency does not follow the stretched exponential distribution.

5.2 Categories

The popularity of a video on social video websites may be caused by many factors. One of the first factors we examine is the type - or category - of the video that is published. Previous studies have showed that this video characteristic affects access frequency [15, 18], yet no study thus far has surveyed all categories on a public social video website. We do so in this subsection and also analyze the relationship between video category and length. In Subsection 5.6, we compare these categories by region to study social aspects.



(a) YouTube (b) Truveo

Figure 6: Category vs. Rank Distribution

The video popularity distribution by category for the Top 100 YouTube Yearly Videos list is shown in Figure 6(a). Note that the majority of videos in this list are categorized as music videos, comedic videos, or entertainment videos. While a few comedic and

entertainment videos have some of the top popularity, music videos are the most popular and have the most frequency of occurrence across the entire rank range on YouTube. Comedic videos have the second most frequent occurrence, yet most of the videos in this category have rank of 50 or lower. The third most frequent category, entertainment, has a similar trend. Autos & Vehicles, Travel & Places, and Sports each have the lowest access frequency, with Sports trailing all categories in terms of rank as well.

A similar trend can be observed for the Top 1000 Truveo Video Search list as shown in Figure 6(b). Comedy, entertainment and music videos again dominate the majority of videos published and accessed by users. Entertainment videos have the most frequency of occurrence across the entire rank range. Most comedic videos tend to have high ranking, whereas most music have low ranking. News is much more frequently accessed in Figure 6(b) in comparison to Figure 6(a) - most notably in the 200 through 1000 rank range. As new categories are introduced with this dataset, Autos & Vehicles, Sports, and Travel & Places are no longer last in terms of categorical frequency - although their access remains sparse in the Truveo Search dataset. Science and Video Games are the two categories with the least number of videos in the Top 1000 list, even though both categories have a video with a rank better than 200.

It is interesting to note the relationship between the video length and category. As shown in Figure 7, Film & Animation, Pets & Animals, People & Blogs, and Sports categories all have video length averages over four minutes. As these video categories have very low frequency (see Figure 6(a)), it is evident that their larger video length averages do not have a significant impact on the overall average video length, compared with the Music category, which has an average of approximately 3 minutes and 45 seconds. Likewise, Figure 8 shows that entertainment and music videos in the Truveo Search data have an average length between 200 and 250 seconds. The overall average video length, however, is more dominated by the more popular comedic videos (approx. 100 and 115 seconds).

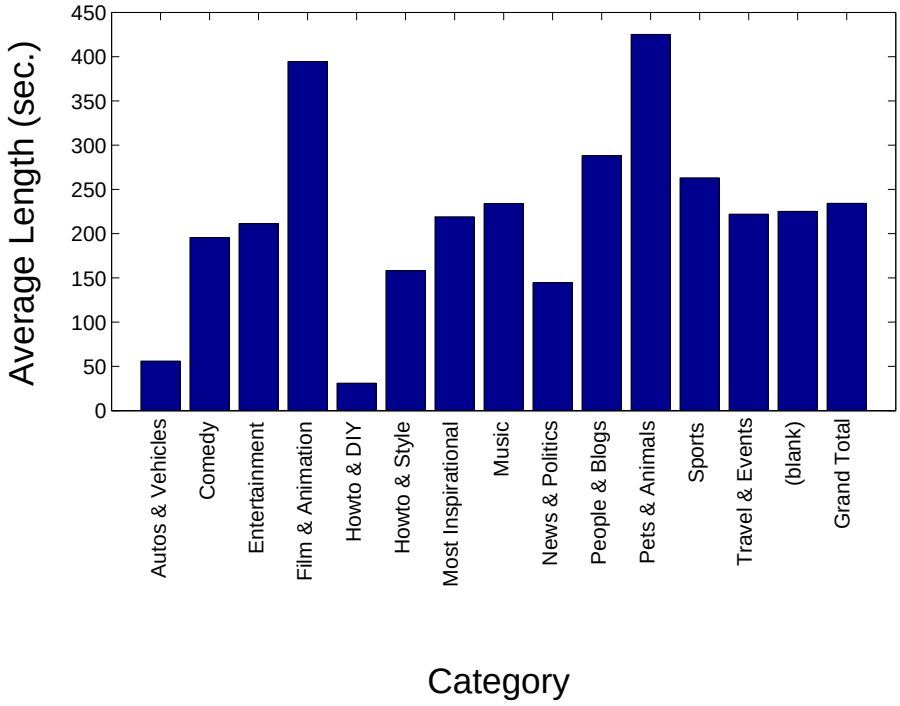


Figure 7: Average Length of Top 100 YouTube Yearly Videos

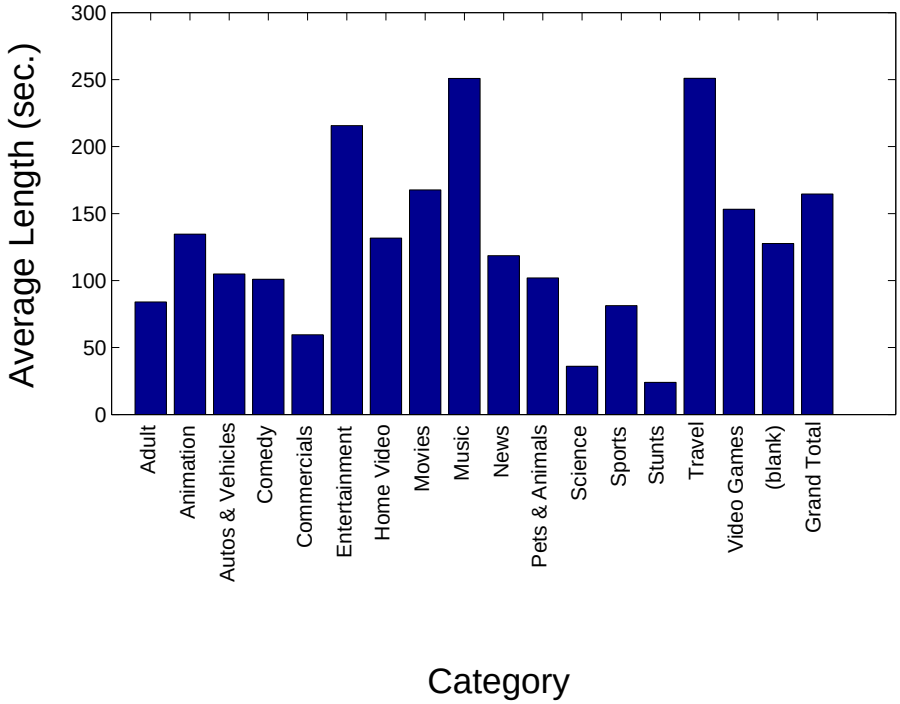


Figure 8: Average Video Length for Top 1000 Truveo Videos

We analyze video length further later in this thesis to identify more general trends across the rank distribution.

5.3 Days of the Week and Evolution

Let us now discuss how a video's popularity changes over time, as well as how the video access changes based on the day of the week and the time during the day. On August 23, 2007, we selected the Top 100 YouTube videos and froze the list for analysis. When analyzing the total daily view count for this list over an eight-week period, we observe interesting trends. First, Figure 9 shows that video popularity is nearly triple when

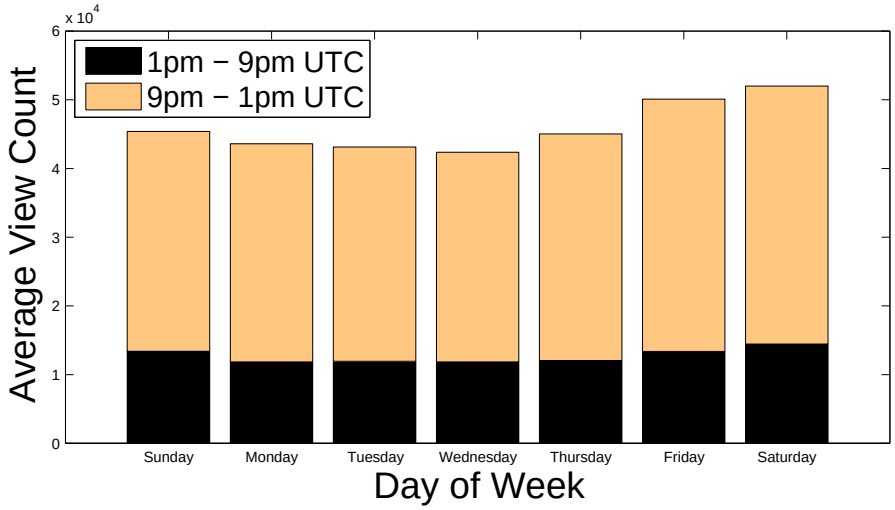


Figure 9: Total View Count Per Day of Week (Top 100 YouTube Videos, Averaged for Each Day over 8-Week Period)

analyzing accesses between 9pm - 1pm coordinated universal time (UTC) in comparison to the remaining twelve hours of the day. Further, Friday and Saturdays in general appear to be the most popular times for video access frequency, while Wednesdays have the least access frequency contribution. This is the first set of data known to show social video access frequency by day of the week and period of the day over such a long period.

Furthermore, our analysis shows that the access frequency of the Top 100 Videos decreases over time as illustrated in Figure 10. The total accesses for the Top 100 Videos

decreased by nearly one million views per day from August 23rd to October 17th. The trends observed in the average view counts per day of the week are observed in similar fashion over the eight week period, where Fridays and Saturdays have the most views and dropping to a minimum for the week once Tuesday and Wednesday are encountered. This oscillating nature of video access frequency is very important to consider when planning for necessary bandwidth and load balance on social media websites such as YouTube.

Understanding how the Top 100 videos evolve based on rank provides answers regarding the video access frequency declines observed over the eight-week period. The rate of change in the accumulative view count is show in Figure 11. Approximately 10% of the videos in the Top 50 continue to have frequent views by users; however, with a correlation coefficient of -0.465 (medium negative correlation), the general trend is less frequent access. In particular, the videos from rank 50 to 100 show a low rate of change across the period. It is further interesting to note that 6 out of the 10 videos with the greatest slope change in Figure 11 are music videos. Popular music videos appear to maintain their access frequency over longer periods of time compared to other video categories.

5.4 Length

The length of a video is very important to understand when planning client sessions with multimedia systems. As YouTube does not allow videos more than ten minutes in length, user sessions on YouTube are limited in time. This is much different from the lengthy videos studied in [15]. Figure 12 shows the average video length profile for daily videos accessed. While rank does not appear to have a direct correlation to video length, it is interesting to note that 5 of the top 6 rankings have videos of two and a half minutes or less.

Videos greater than six minutes do not appear to be the norm in Figure 12, with the majority of videos having a length of five minutes or less. The correlation coefficient is

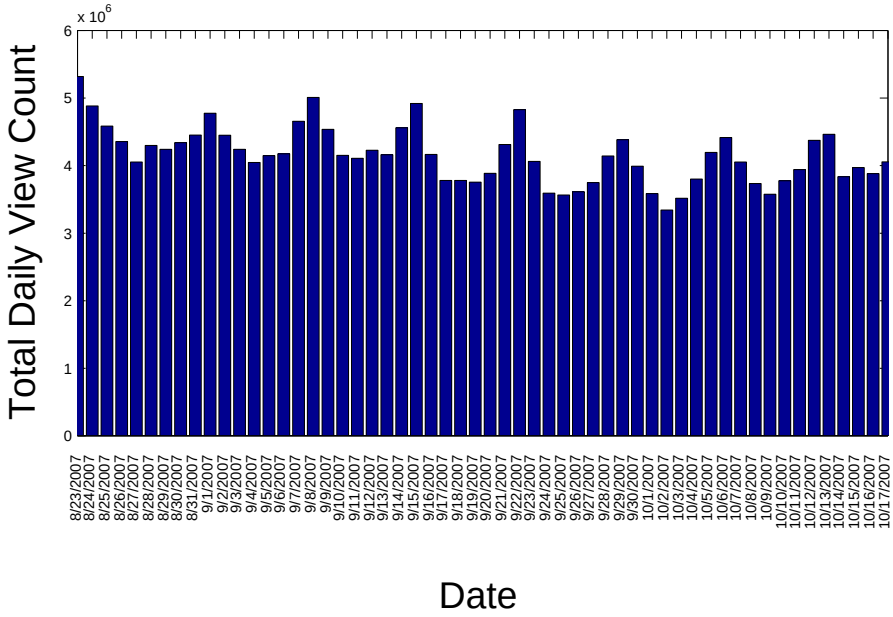


Figure 10: Total Daily View Count (8-Week Period) - Top 100 YouTube Videos

0.340 in Figure 12, which shows a tendency for videos of lower rank to be more lengthy in comparison to videos of higher rank. As you can see in Figure 14(a), a large portion of the most popular videos are in the three to six minute range.

When looking at the average length distribution in Figure 14, it is interesting to note the

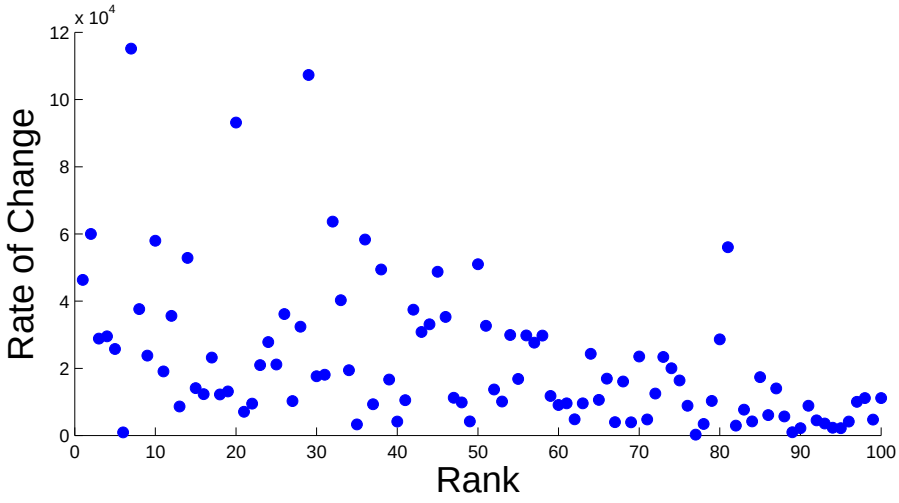


Figure 11: Rate of Change in the Accumulative View Count - Top 100 YouTube Videos

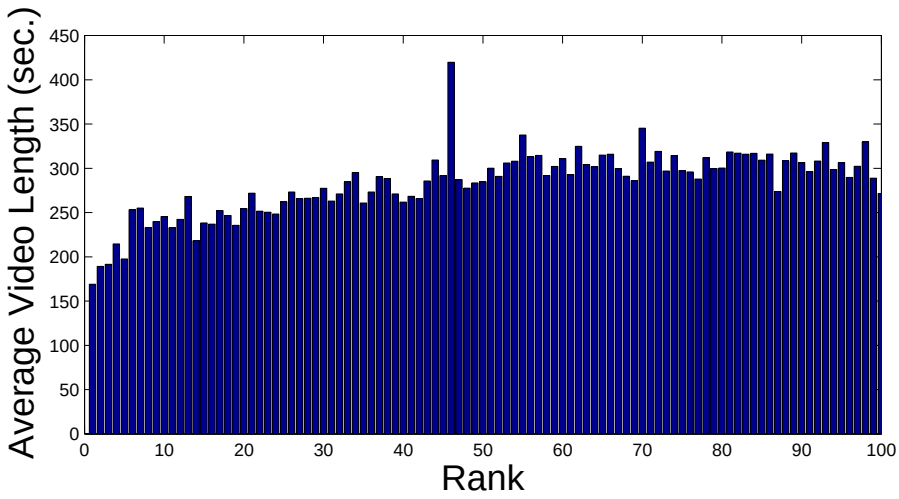


Figure 12: Average Length vs. Rank - Top 100 Daily YouTube Videos

progression in average length data for each YouTube population - top 100 daily, weekly, monthly, and yearly videos. When looking at daily accesses, most videos fall within the 225 to 350 second range. Over 50% of videos accessed are near 300 seconds in length, while videos less than 225 seconds and greater than 350 seconds only accounting for 6% of the video accesses on average within the daily population. As we move to the top 100 most popular weekly videos, we see a significant shift in the frequency of videos accessed for videos of approximately four minutes (250 seconds) in length, with nearly 50% of videos near 250 seconds and 75% of videos now within the 225 to 325 second range. Frequency of access across the average length spectrum is not as distributed when analyzing the weekly population, and this trend continues when continuing to analyze the monthly and yearly populations further. Approximately 70% of the popular videos are concentrated in the three to four minute range in Figure 14(d). This is very important information, as it has direct implications on user access time, necessary bandwidth, and storage maintenance for the video's lifecycle. Videos are stored and available to be accessed on YouTube until the publisher removes them (or until a terms of use violation is realized). When analyzing the average video length for a larger set of videos, we observe a difference in the video population. Figure 13 shows that the Truveo Video Search engine indexes

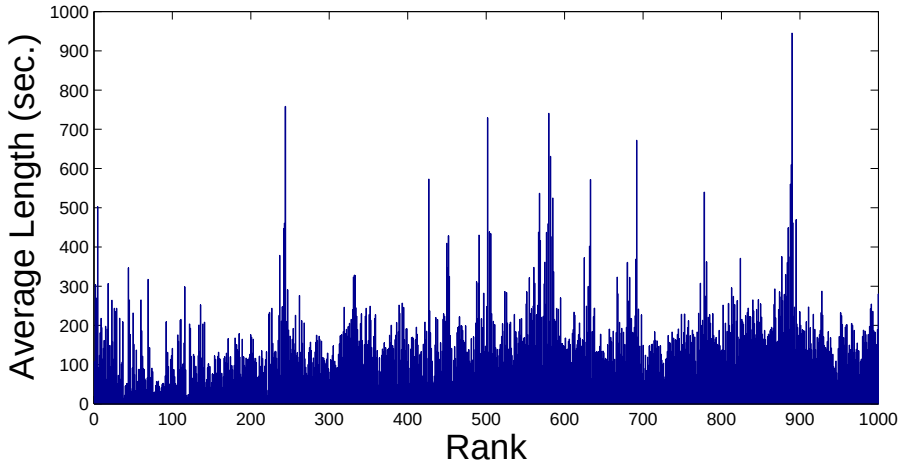


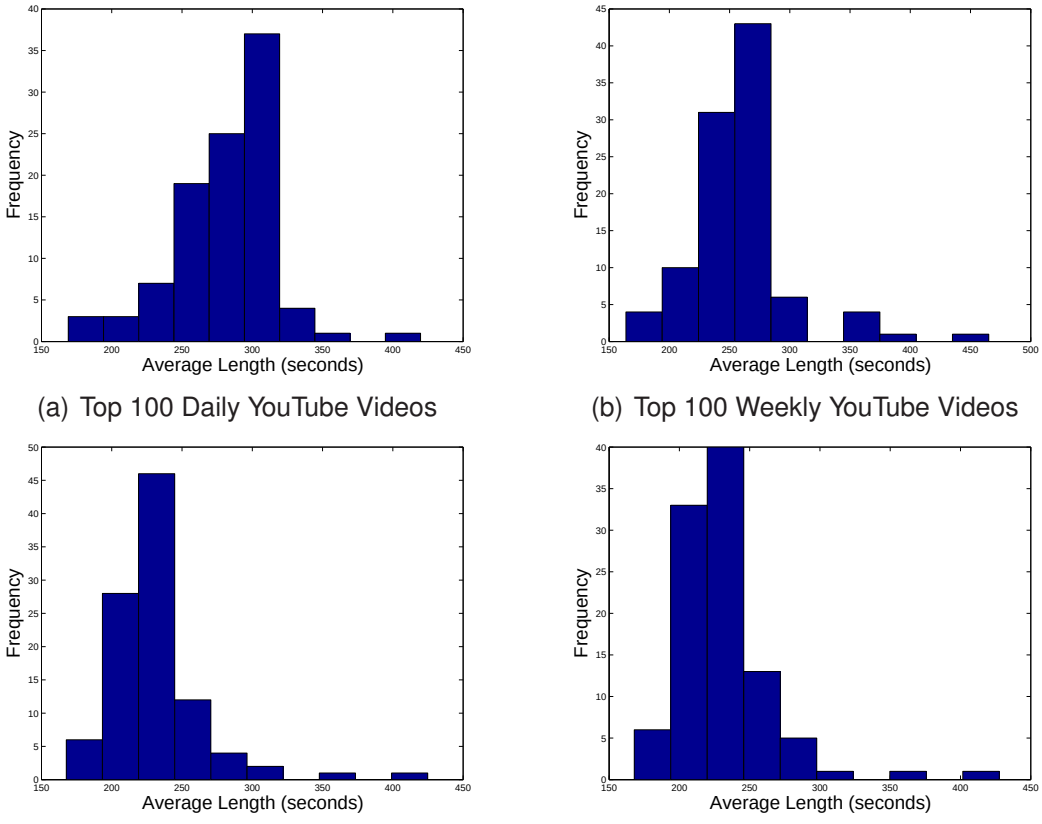
Figure 13: Average Length vs. Rank - Top 1000 Videos Indexed by Truveo

videos greater than ten minutes in length, different than the YouTube restriction. There are actually several videos - notably less than rank 200 - that are present in Figure 13. Second, a majority of videos across the distribution are below 250 seconds in length. Further, a majority of the videos with a rank of 100 or less appear to be less than 200 seconds in length in Figure 13. The correlation coefficient is 0.274, which shows small positive correlation and a slight tendency for videos of lower rank to be more lengthy than those of higher rank.

Figure 15 shows a histogram of the Top 1000 videos from the Truveo Video Search engine. The frequency of videos in the 100 to 200 second range accounts for more than 50% of the videos over the distribution. There are several videos over five minutes in length; however, the frequency of this occurrence is much less in comparison to the norm.

5.5 Channels and Video Formats

We now further analyze two additional video characteristics: channel and video format frequency. As discussed earlier, YouTube is now the largest social video website and the third largest website accessed in the world. This is further confirmed by the channel



(a) Top 100 Daily YouTube Videos

(b) Top 100 Weekly YouTube Videos

(c) Top 100 Monthly YouTube Videos

(d) Top 100 Yearly YouTube Videos

Figure 14: Average Length Distribution

distribution from the Truveo Video Search engine data. As we see in Figure 16, the YouTube channel (or, source of videos) accounts for 35% of video access alone, while the AOL Music, Movies, and News video channels together account for nearly 30% of the videos indexed by Truveo. MySpace, Metacafe, iFilm, Google Video, and Glumbert are some of the other channels that make up the top 90% of all videos accessed on the web. This channel distribution is significant, as it shows us that a large majority of users rely much more on YouTube than other websites for their online videos. The channel frequency distribution, therefore, from this data not only might show the potential necessity for resource allocation per channel, but also where opportunity for channel growth and consolidation of resources might lie.

Further analysis of the Truveo Video Search data shows a startling imbalance in video formats accessed by users (see Figure 17). The Flash video format is the sole format for

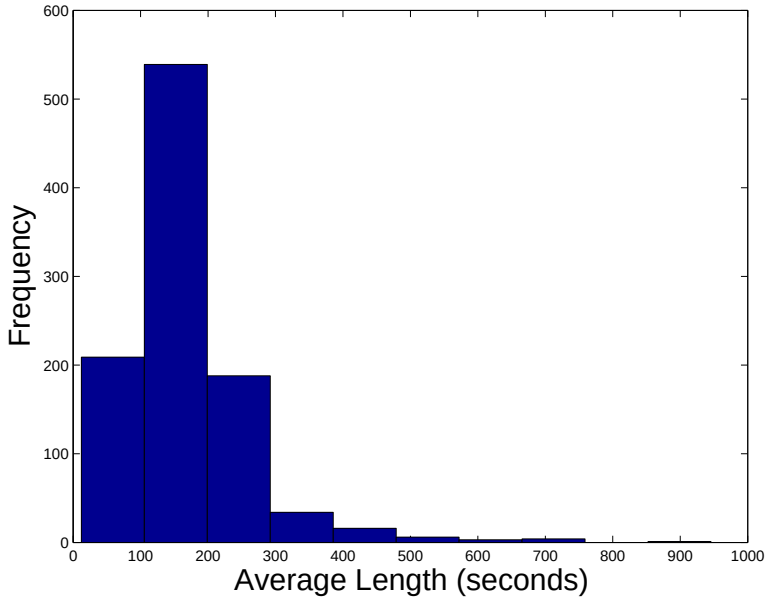


Figure 15: Average Length Distribution - Top 1000 Truveo Search Videos

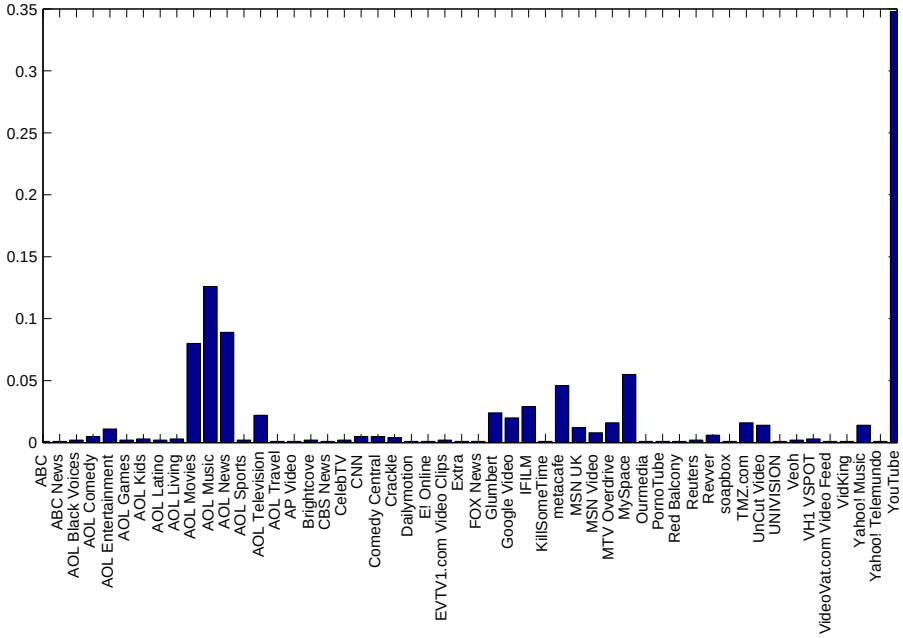


Figure 16: Channel Distribution - Top 1000 Videos Indexed by Truveo

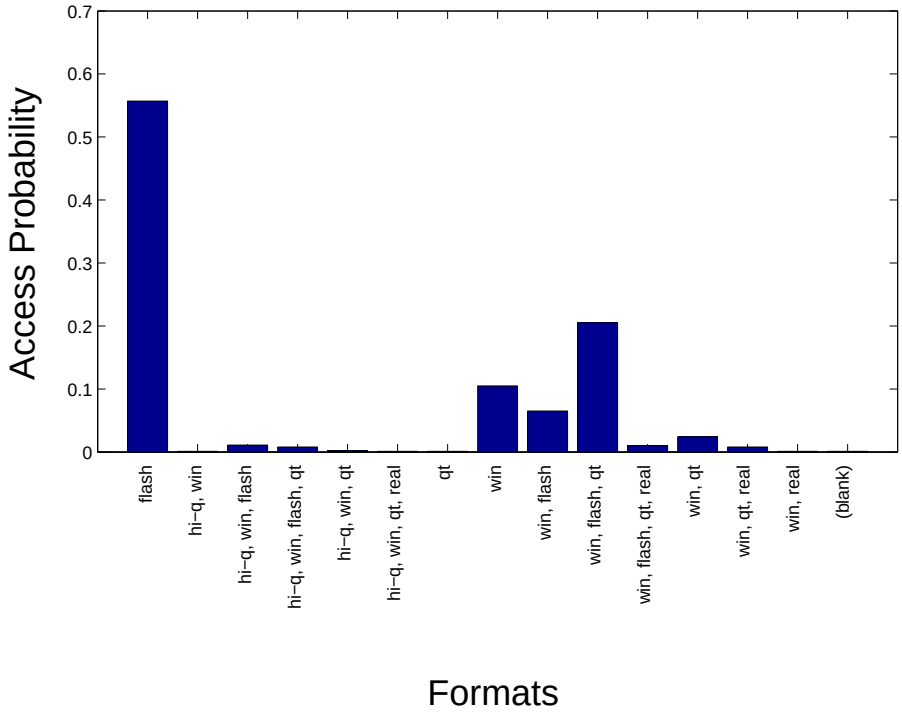


Figure 17: Video Format Distribution - Top 1000 Videos Indexed by Truveo

55% of the videos published online. Further, more than 80% of videos published online can be provided to users in this format. Flash dominates Real Media, Windows Media, and QuickTime formats in Figure 17. This is significant data, as it allows for network engineers to anticipate necessary bandwidth loads, for consumers of the content to ensure they have the latest updated codecs and players, and ultimately content providers to move toward the most relevant media type to maintain and attract consumers.

5.6 Geographical Aspects

Video popularity and rank up until this point have been aggregated and presented without specific details regarding the users that are participating in social video websites. Understanding the background of people who participate in social video websites adds an entirely new dimension to understanding a particular culture. The analysis of Google Video data provides this added dimension and helps to understand which types of videos

users from different regions around the world like to watch.

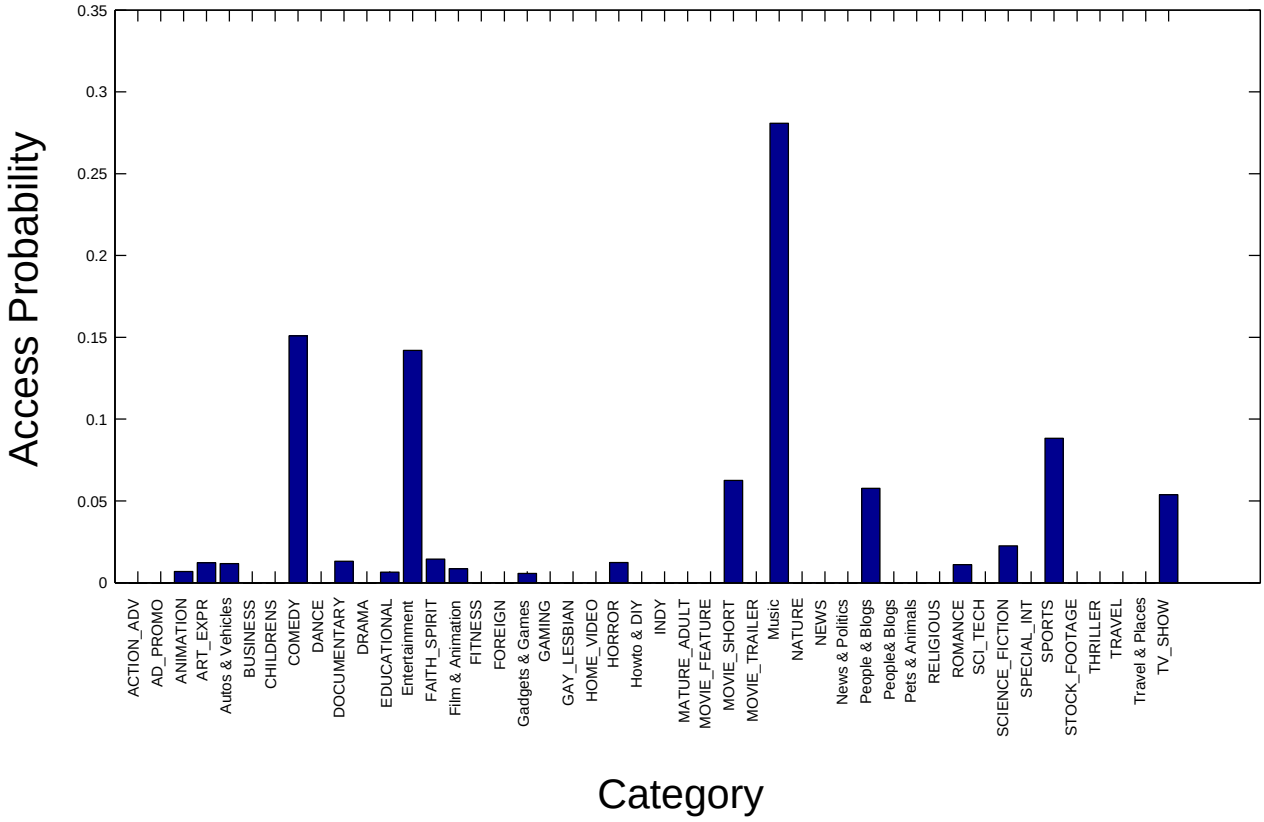


Figure 18: Video Category Distribution for Top 100 Google Videos

The distribution of video popularity for the Top 100 Google Videos for all countries and regions is shown in Figure 18. This Google Video data confirms again that music, comedic, and entertainment videos are the most popular, similar to our analysis of YouTube and Truveo results in Figure 6. Here, music is nearly twice as popular as the comedy and entertainment categories. Different from our previous YouTube category analysis in this thesis, sports, short movies, blogs, and TV shows are the next most significant categories that are watched, each accounting for at least 5% of the Top 100 videos.

As the data from Google Video is based on videos from Google, YouTube, and a few other video websites, we get a significant picture of what people around the world like to watch in Figure 18. The data in Figure 18 is biased, however, by the total number of

view counts and does not give any information on what people in the region of Oceania like to watch in comparison to people in the Near East and how video popularity among categories might differ based on geographical region.

In order to get a much more detailed analysis, we analyzed data for the forty-four countries for which Google reports and grouped the data into the regions identified by the U.S. Census Bureau: Asia (excluding Near East), Commonwealth of Independent States, Eastern Europe, Latin America and the Carribean, Near East, Northern America, Oceania, Sub-Saharan Africa, and Western Europe. We weighted the data collected for each region by the country from which it was attributed. We then further weighted the ranking by the Zipf distribution to ensure each category achieved appropriate importance in the analysis.

As can be seen in Table 2, music, comedy, and entertainment have video categories which are watched in all regions. Their favor in each region varies considerably, however. Music in Latin America and the Carribean, Northern America, and Western Europe each account for more than 30% of the videos watched. This is in comparison with more than 20% influence in Sub-Saharan Africa and approximately 15% influence in the Commonwealth of Independent States and Eastern Europe, respectively. The only other video category that comes close to such an influence within a particular region is comedy, which accounts for 30% of the videos watched in Eastern Europe and 28% in Oceania, with further considerable influence in nearly all other regions. Entertainment videos have a wide array of influence. Most notably, 18% of videos watched in Asia (excluding Near East) are entertainment videos, while nearly 15% in Northern America and 13% in Sub-Saharan Africa respectively. We now take a look at each region in detail and look for trends.

CATEGORY	REGION								
	Asia (excl. Near East)	Common. of Indep. States	Eastern Europe	Latin America & Carribean	Near East	Northern America	Oceania	Sub-Saharan Africa	Western Europe
Action & Adventure	0.30%	3.20%	1.02%	0.05%	8.21%	0.00%	0.38%	1.06%	0.23%
Anime & Animation	1.54%	0.00%	7.21%	1.63%	0.82%	0.66%	1.91%	0.72%	3.57%
Art & Experimental	0.90%	1.15%	0.61%	2.22%	0.04%	1.59%	0.24%	2.04%	2.71%
Autos & Vehicles	0.48%	1.32%	0.67%	1.15%	0.47%	0.89%	0.12%	2.98%	1.13%
Children & Family	0.16%	0.83%	1.39%	0.18%	0.05%	0.00%	0.00%	0.00%	0.16%
Comedy	9.24%	15.58%	30.04%	12.15%	13.83%	15.09%	28.14%	14.49%	13.89%
Documentary	1.74%	13.73%	1.55%	1.50%	1.80%	0.87%	2.60%	1.17%	1.88%
Drama	0.40%	1.57%	0.00%	0.18%	2.40%	0.00%	0.00%	0.00%	0.20%
Educational	1.62%	0.53%	0.00%	1.25%	2.15%	0.66%	1.64%	0.00%	1.17%
Entertainment	18.16%	5.29%	1.53%	8.63%	8.03%	14.64%	10.93%	12.63%	8.12%
Faith & Spirituality	0.54%	0.44%	0.63%	0.21%	0.06%	2.03%	3.14%	1.51%	0.52%
Film & Animation	2.17%	0.37%	2.54%	1.82%	0.51%	1.17%	0.62%	0.00%	0.57%
Foreign	0.71%	1.39%	0.00%	0.03%	0.19%	0.00%	0.46%	0.00%	0.13%
Gadgets & Games	0.40%	1.31%	0.00%	0.08%	0.00%	0.84%	0.60%	0.00%	0.05%
Gaming	3.04%	7.00%	0.58%	2.24%	0.37%	0.18%	0.14%	0.75%	1.86%
Home Video	0.49%	1.09%	2.62%	0.37%	3.51%	0.65%	0.37%	1.00%	0.64%
Horror	0.86%	0.00%	0.75%	0.08%	0.00%	1.22%	2.41%	0.45%	0.33%
Howto & DIY	0.61%	0.69%	0.00%	0.76%	0.06%	0.00%	0.11%	1.97%	0.26%
Movie (feature)	2.61%	1.86%	1.05%	1.62%	16.86%	0.06%	0.69%	0.91%	1.76%
Movie (short)	7.83%	4.42%	11.08%	4.38%	9.11%	4.29%	6.38%	3.24%	4.87%
Movie Trailer	0.60%	0.00%	0.00%	0.27%	2.87%	0.00%	0.19%	0.63%	0.57%
Music & Musical	12.10%	15.25%	15.45%	34.17%	3.91%	31.90%	8.43%	22.24%	31.90%
News	0.47%	0.00%	0.00%	1.39%	0.02%	0.55%	0.00%	0.60%	0.03%
News & Politics	2.88%	1.46%	0.00%	0.68%	0.45%	0.67%	0.00%	0.00%	0.11%
People & Blogs	5.32%	7.68%	2.28%	3.86%	4.19%	7.21%	3.56%	1.78%	3.81%
Pets & Animals	0.06%	0.00%	0.00%	0.00%	0.21%	0.13%	2.67%	1.70%	0.42%
Romance	4.07%	0.87%	0.00%	0.35%	1.36%	0.95%	0.68%	0.53%	0.23%
Science & Technology	0.50%	0.42%	0.00%	0.11%	0.00%	0.06%	2.16%	0.35%	0.32%
Sci-Fi & Fantasy	2.87%	1.18%	0.00%	0.30%	0.27%	1.59%	3.20%	0.00%	0.60%
Special Interest	0.32%	1.82%	0.00%	0.60%	0.35%	0.00%	0.56%	0.00%	0.51%
Sports	7.40%	4.35%	9.70%	8.50%	10.76%	2.34%	6.79%	18.31%	10.06%
TV Show	3.83%	4.01%	5.52%	4.69%	3.22%	5.23%	6.33%	3.88%	3.50%

Table 2: Regional Video Category Comparison

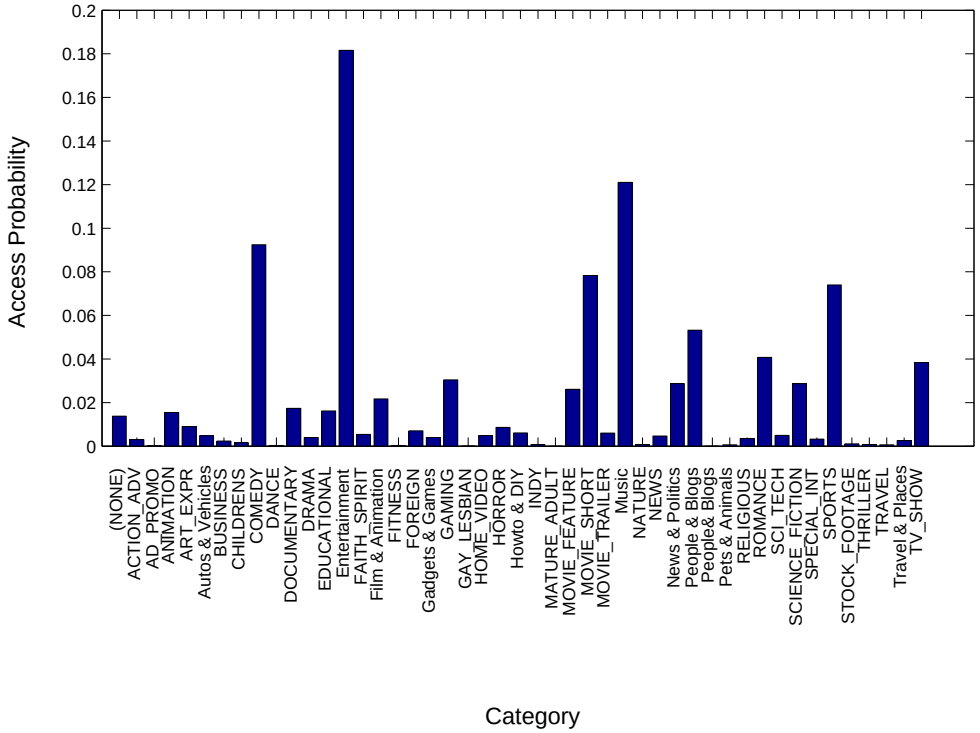


Figure 19: Asia (Excluding Near East)

5.6.1 Asia (Excluding the Near East)

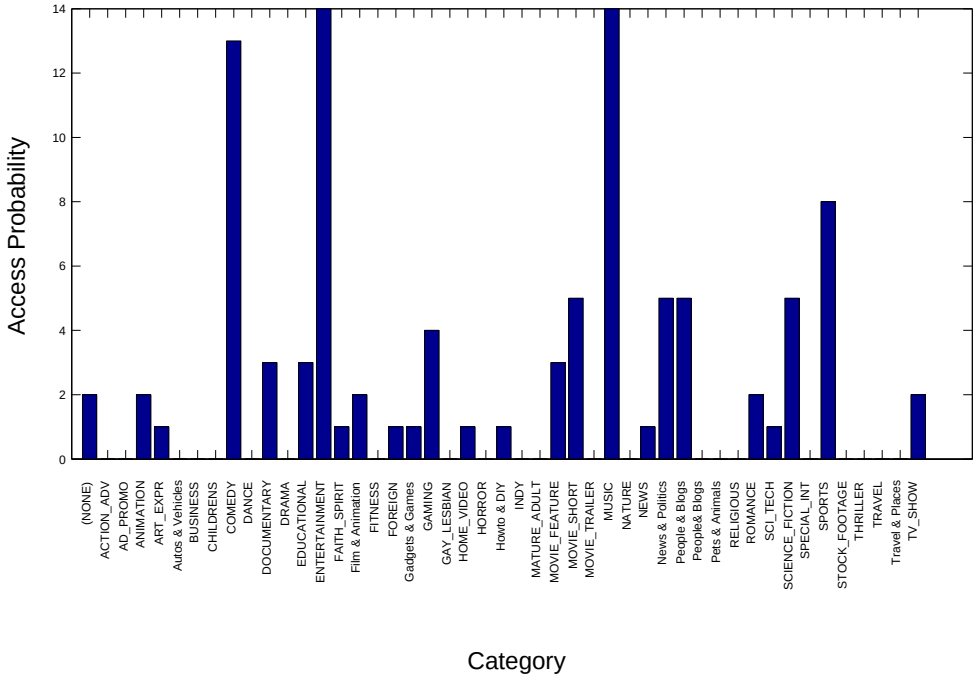
When we look at Asia (excluding the Near East) in Figure 19, we note that the top three categories are slightly different than the world trend. In this region, entertainment videos are the most popular; however, their access probability is only 18% in comparison to the most popular category (music) in the world trend. What is notable as well is that entertainment videos are more popular in Asia in comparison to entertainment videos within the world trend (29%). Music (14%) is much less popular in Asia in comparison to the world (29%), and comedy (9%), short movies (8%), and sports (7%) all hold the next most significant popularity. While other categories hold less popularity in the region, Asia does show significant diversity with thirty-six different categories watched when compared to all of the other regions in the world, only second to Western Europe in the number of categories in its access probability profile. This is compares to eighteen, or double the amount of categories, in the access probability for the world. To summarize the region,

Asia (excluding the Near East) appreciates watching many different types of social videos, yet the most popular are consistent with the top three of the world - entertainment, music, and comedy - though in a different order.

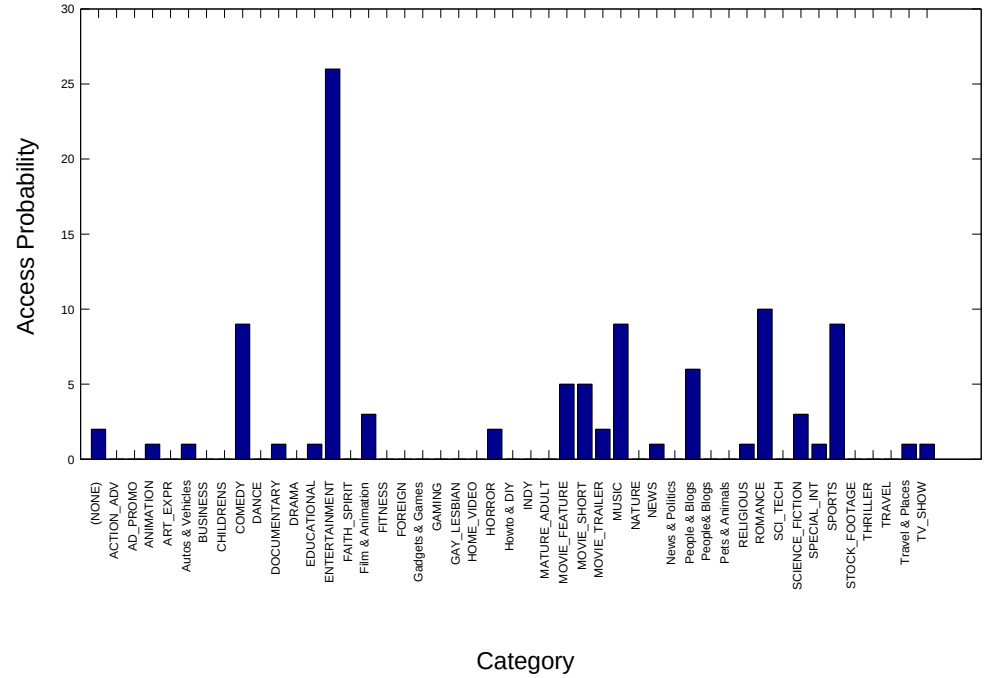
The two largest countries - China and India - in Asia (excluding the near East) each boast of over 1 billion people and together account for over 75% of the population when compared with other countries reported by Google. When looking at China in Figure 20(a), we see that music and entertainment each constitute 14% of users' access probability in the region, with comedy closely behind with 13% access probability. The Chinese show additional interest in sports social videos with 8% access probability, while gaming, short movies, news and politics, people and blogs, and science fiction each account for more than 4% in the country. Transferring our focus to India, we quickly see in Figure 20(b) that entertainment videos are the most popular with more than 25% of the access probability, while comedy, music, romance, people and blogs, and sports each account for 10% or less of the access probability among viewers in India. We note that interest in romantic social videos at 10% is the largest in India in comparison to any other country.

5.6.2 Commonwealth of Independent States

The Commonwealth of Independent States has a different access probability profile in Figure 21. Comedy is the most popular category, compared to music which is the most popular from a world view and entertainment in Asia (excluding the Near East). What's interesting to note in the Commonwealth, however, is that while music holds the second most popular place in terms of popularity, documentaries are the third most popular category at 14%. This is a unique characteristic of social video watched in the Commonwealth of Independent States, as no other region shows an access probability of more than 3% for documentaries. Why are documentaries so high in popularity in comparison to other categories watched in the region, as well as so prominent in the Commonwealth in comparison to other regions? We believe this data may be a very good starting point for



(a) China



(b) India

Figure 20: Asia (excluding Near East): Country-Level Category Comparison

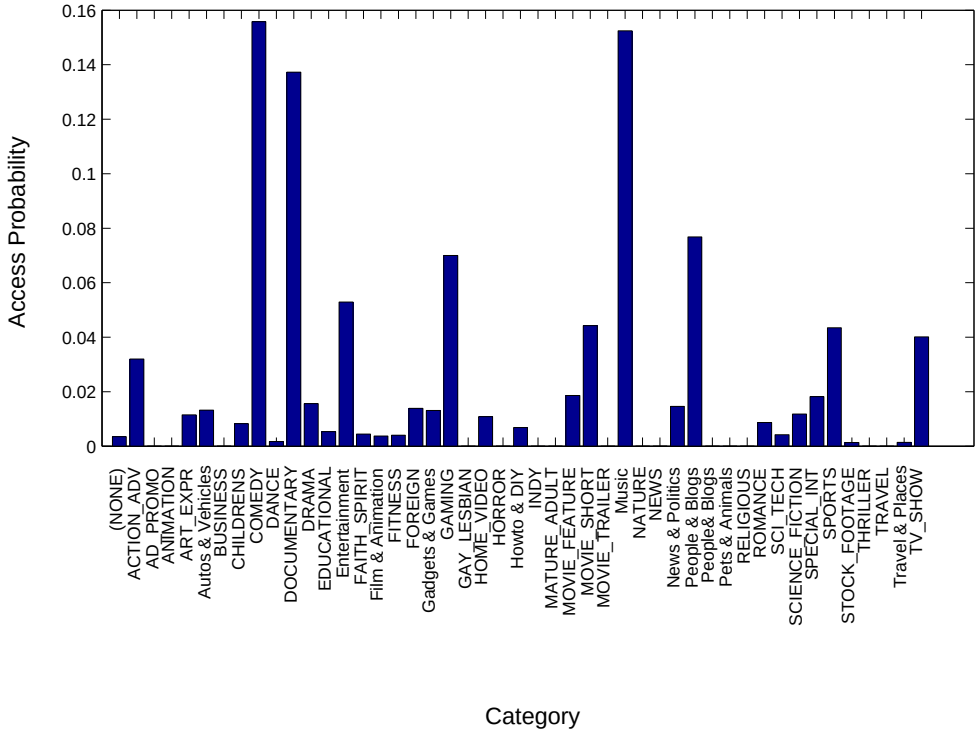


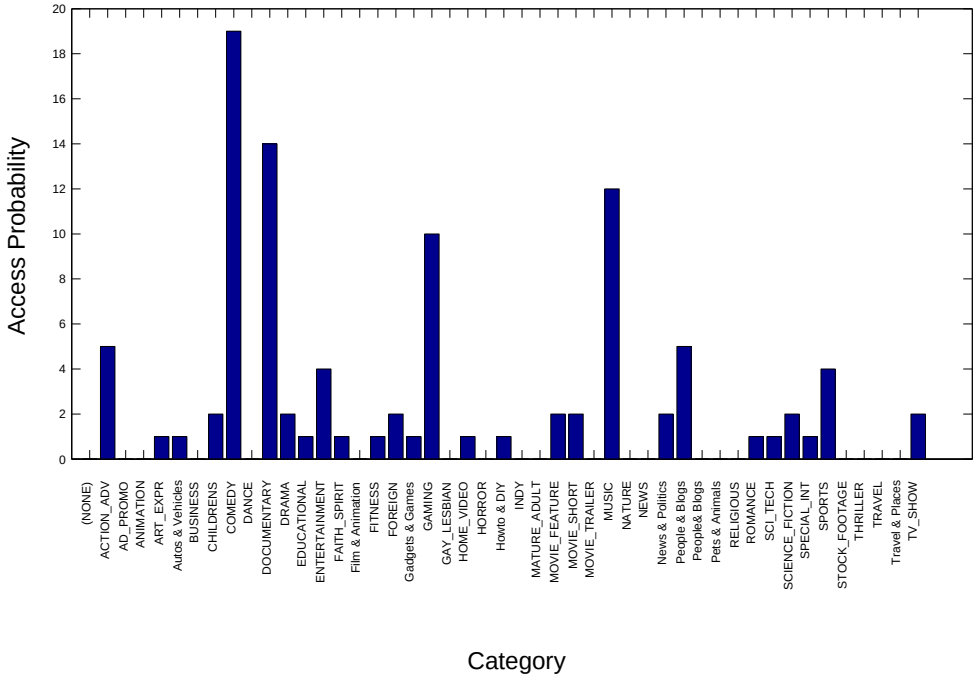
Figure 21: Commonwealth of Independent States: Top 100 Google Videos

further anthropological and social study.

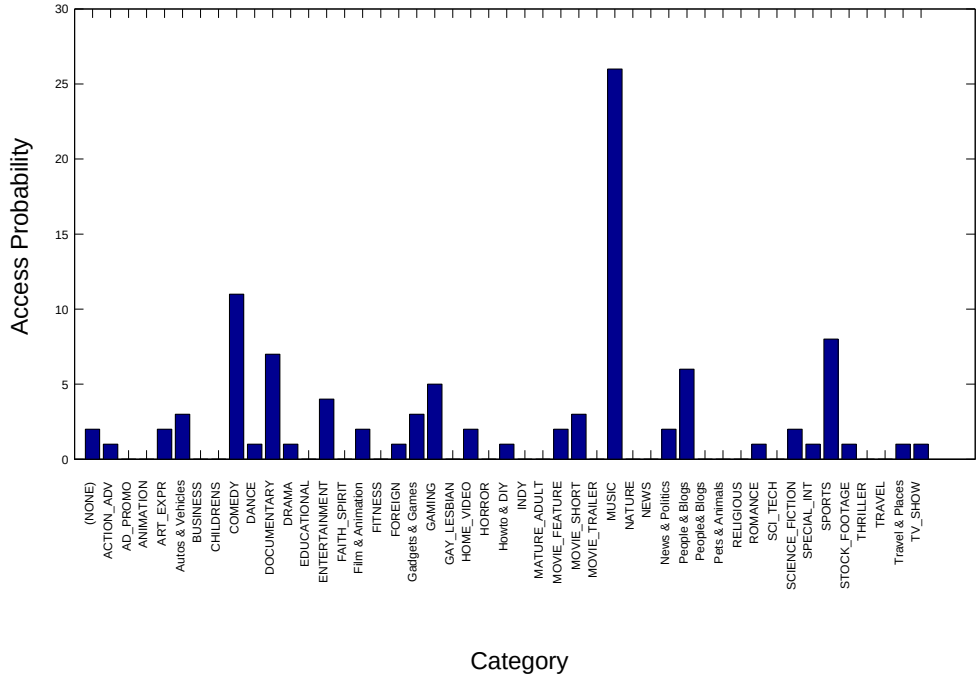
Looking at the country level category comparison for the Commonwealth of Independent States in Figure 22, we see that Russia is more interested in comedic social videos overall at 18%, while the Ukraine is more interested in music social videos at 26%. While comedy and music appear to be significant in both countries, we see that entertainment social videos are not as popular in these countries as they have been in other countries and regions. Documentaries are viewed quite frequently in comparison with other countries, with Russia at 14% access probability and the Ukraine at 7% access probability for documentaries, respectively. Gaming and sports both show significant views in both countries as well, with Russia attributing 10% of its video accesses to gaming.

5.6.3 Eastern Europe

In Figure 23, the social video category distribution in Eastern Europe is considerably



(a) Russia



(b) Ukraine

Figure 22: Commonwealth of Independent States: Country-Level Category Comparison

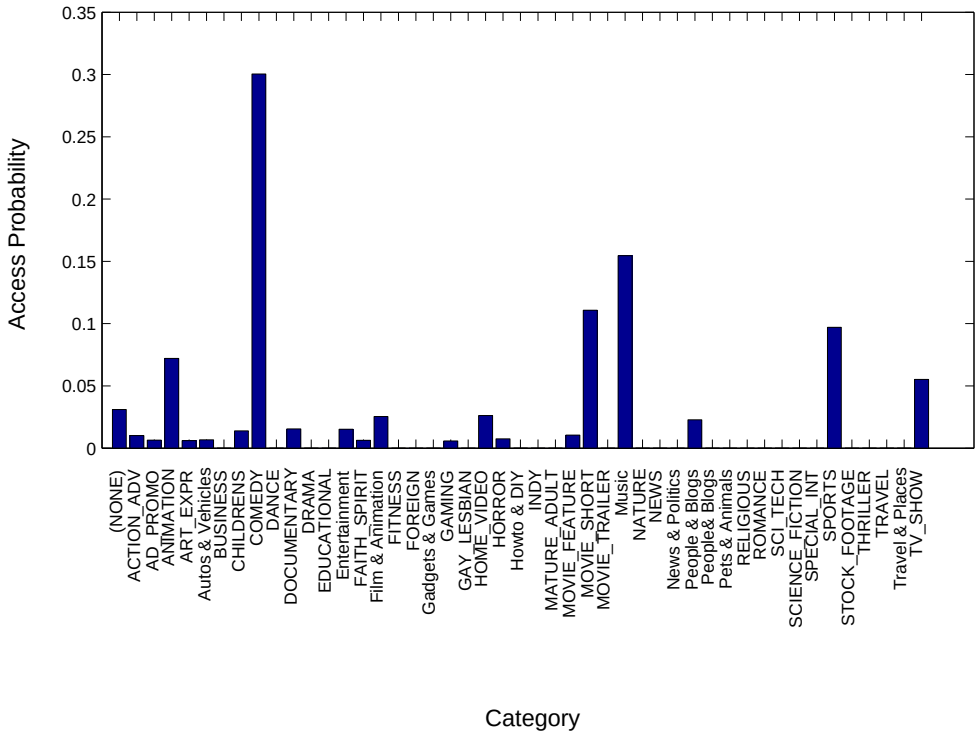


Figure 23: Eastern Europe: Top 100 Google Videos

different. (We note in this instance that this data solely comes from the country of Poland, as Poland is the only Eastern European country reported by Google.) Comedic videos account for 30% of the accesses in the region, compared to music at 16%, short movies at 13%, and sports at 10%. What is interesting to note in Eastern Europe is that entertainment videos do not have a significant access in the region - only 2% in comparison to an access probability of 14% for the world. Music videos, still with significant access at 16%, are not as popular in this region in comparison to world statistics. We leave the reasons for these differences to further investigation.

5.6.4 Latin America and the Carribean

The access profile in Latin America & the Carribean is somewhat similar to the world trend in Figure 24, yet shows more frequent music video access. Music videos account for 34% of access within the region, with comedy and entertainment videos accounting for 12% and 9% of accesses respectively. Sports shows 9% of regional access as well,

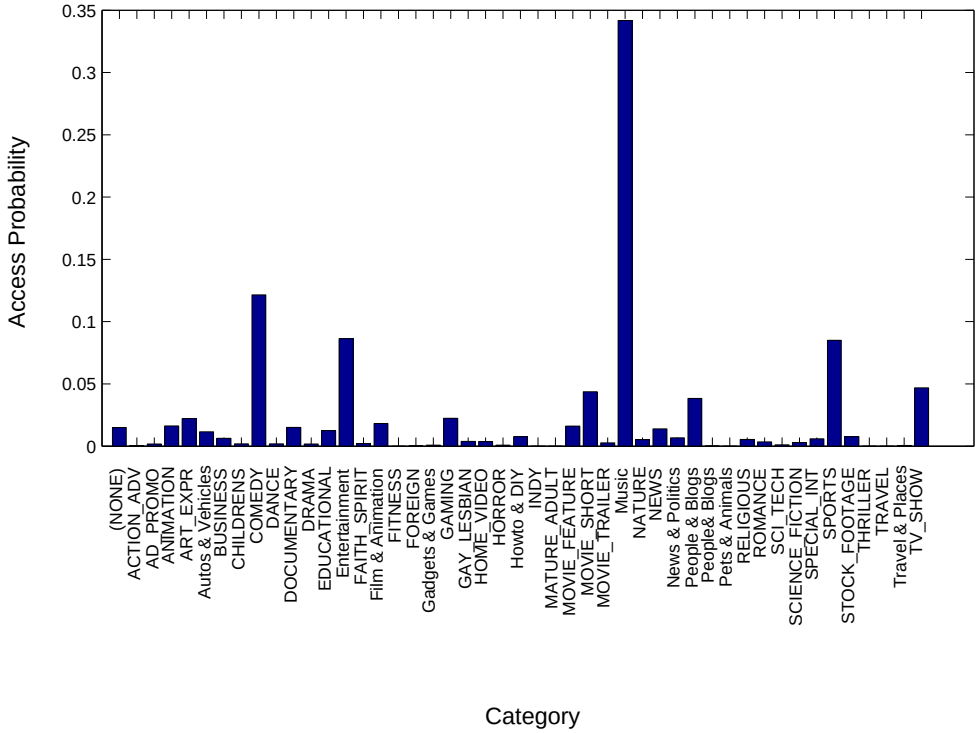
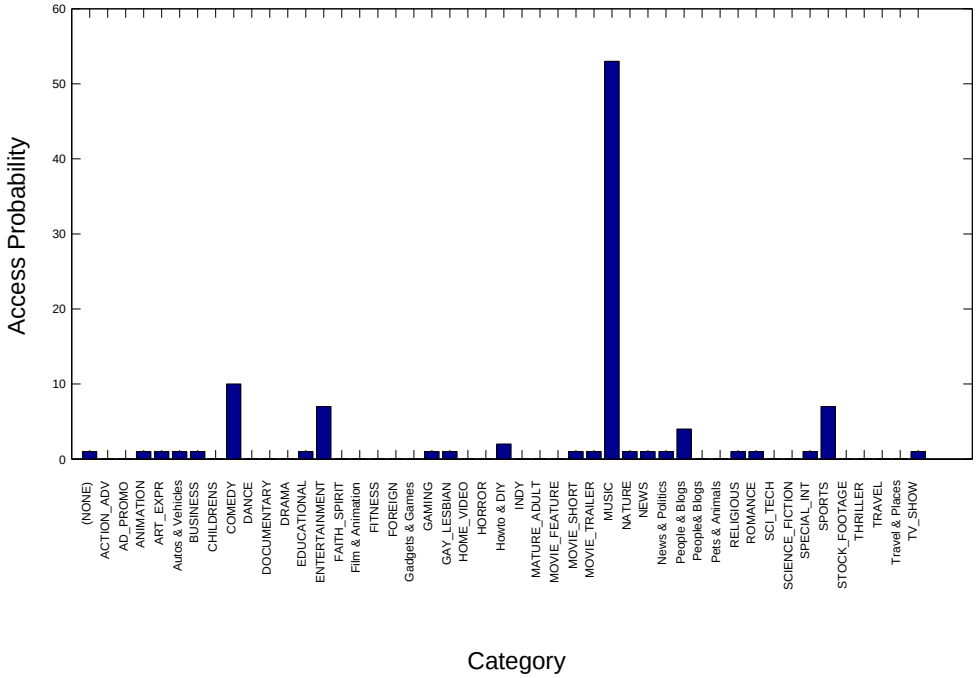


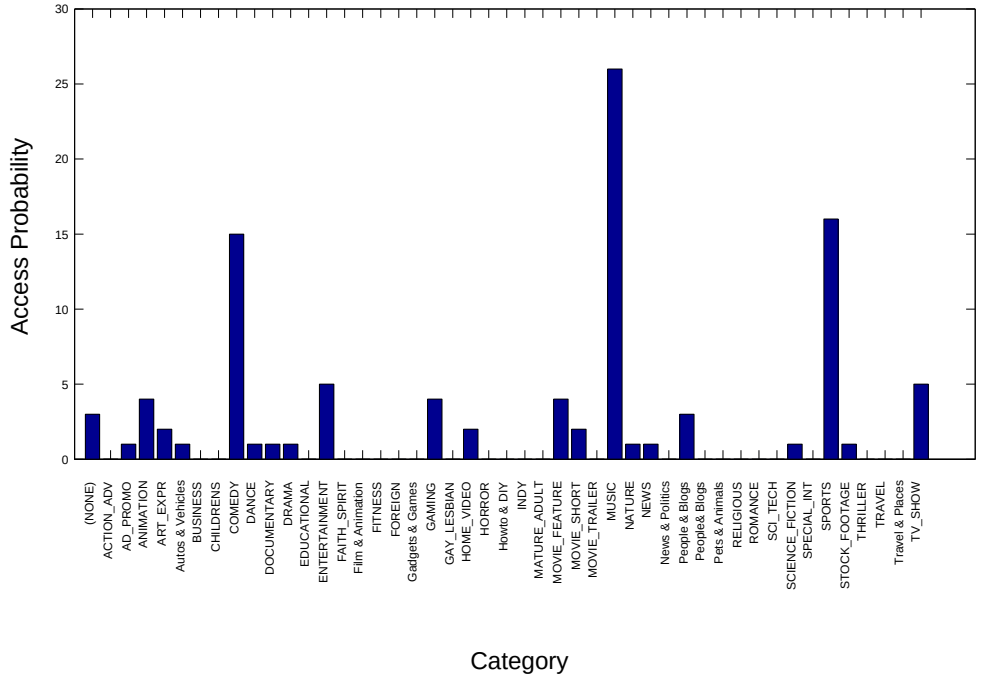
Figure 24: Latin America & the Carribean: Top 100 Google Videos

with TV shows showing 5% access in the region. To summarize this region, Latin America & the Carribean appreciate watching social music videos the most, more than any other world region. Comedy, entertainment, and sports categories each are watched approximately a third as often as music videos.

As we look at the two most populous countries in this region, we see why the music category is so popular in this region. In Figure 25(a), we see that over 50% of the social video accesses in Brazil are to videos with a category of music. This is the most significant percentage for a specific social video category in the world, much less a specific region. Comedy, entertainment, and sports are the only other categories in Brazil that have more than 5% access probability, and no category but music has more than a 10% access probability. As we look to Mexico in Figure 25(b), we see that music, comedy, and sports social videos each have more than 10% access probability, with music having the most significant attention at more than 25%. Again here, entertainment and TV show



(a) Brazil



(b) Mexico

Figure 25: Latin America and the Carribean: Country-Level Category Comparison

videos each only capture approximately 5% of the social videos watched in the country. We see deviation here in both countries from the world trends; however, we do see the regional and country-level affirmation for the social video category of music.

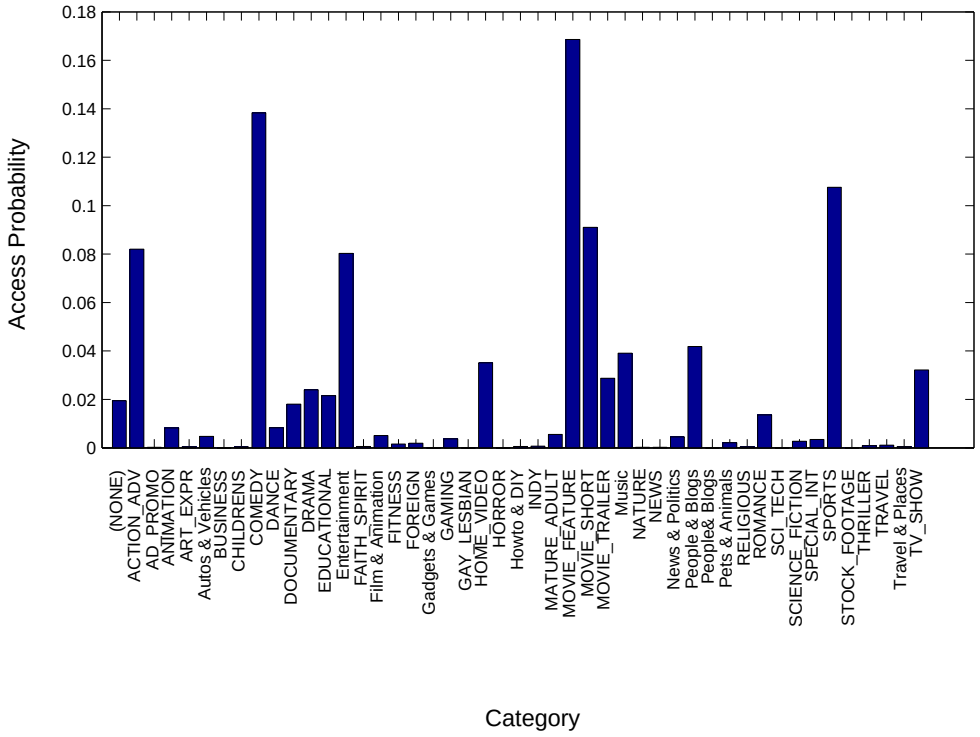
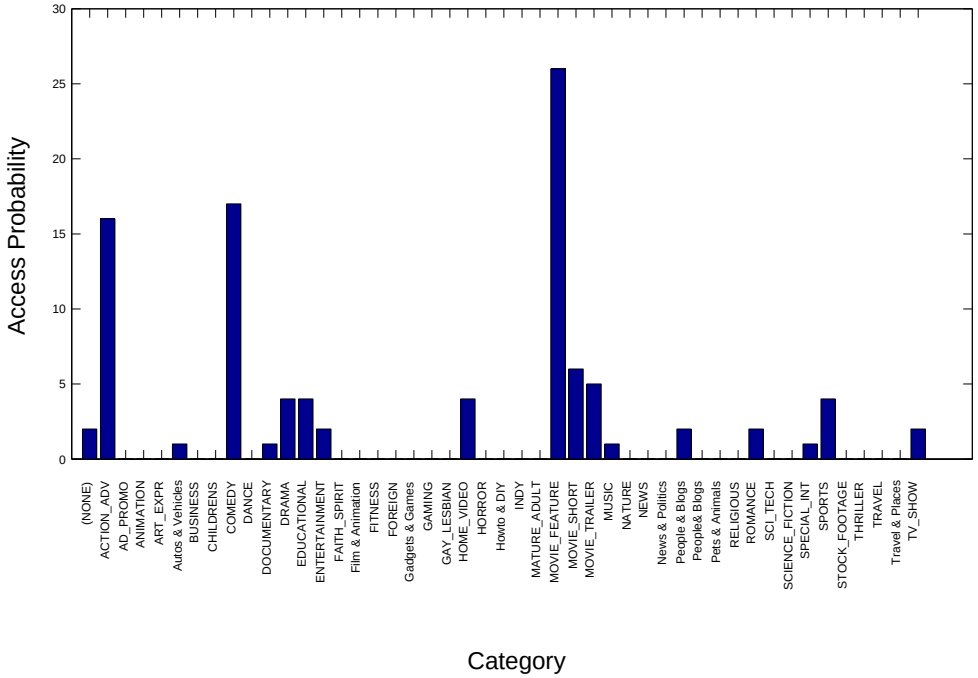


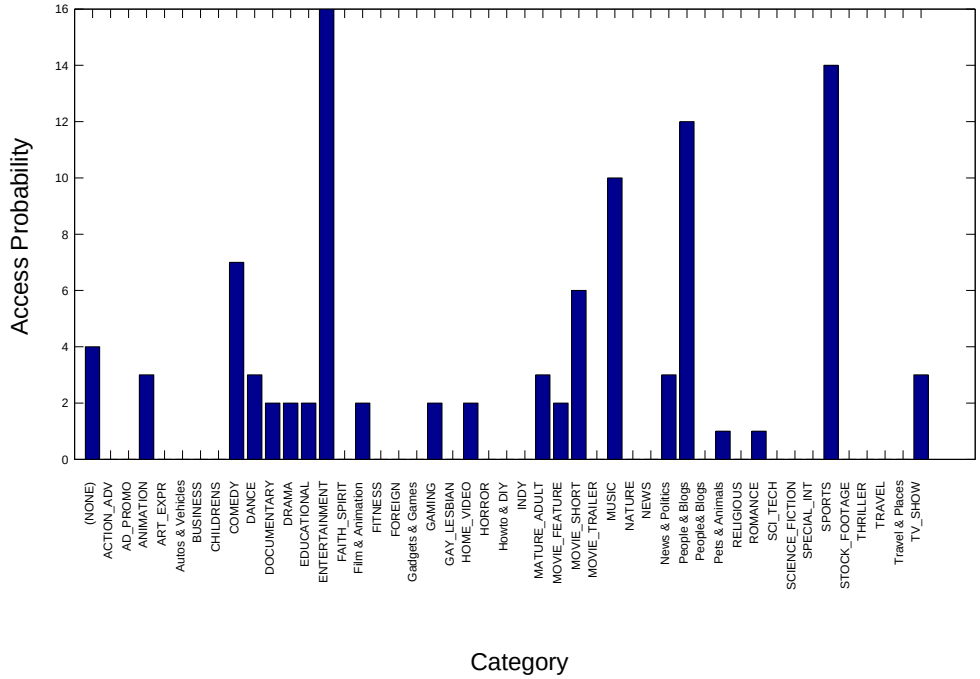
Figure 26: Near East:Top 100 Google Videos

5.6.5 Near East

The Near East does not follow the world trend and actually shows much more distributed access, as seen in Figure 26. The feature movie category is the most popular at 17% - the only region to show such strong access for this social video category. Comedic videos fall in second place at 14%, while sports, short movies, entertainment, and action adventure social videos each fall within the 8 - 12% access range. The Near East has six categories above 8% access probability, more than any other world region.



(a) Turkey



(b) Saudi Arabia

Figure 27: Near East: Country-Level Category Comparison

Turkey and Saudi Arabia are the two most popular countries in the Near East region, and we see varied interest in social video categories in each region. In Turkey, feature movies account for more than 25% of the access probability distribution, higher than any other country reported by Google for this category. As we see in Figure 27(a), while short movies and trailers together account for more than 10% of additional access probability, comedy at 17% and action adventure at 16% account for large portions of the watched videos. It is, again, interesting to note that the viewing of action adventure videos are the highest in Turkey compared to any other country. Saudi Arabia shows more distributed results in Figure 27(b), with entertainment videos at 16% access probability, sports at 14% access probability, and people and blogs at 12% access probability. It is interesting to note that people and blogs social videos have the highest access probability in Saudi Arabia compared to the rest of the countries reported, while its access probability of entertainment videos is above both the media and average (10% and 9.96%, respectively) when compared to all other countries. A very good starting point for social study is the following questions: 1) Why are social videos containing feature movie content so popular in Turkey in comparison to other countries? 2) Why are people and blog social videos in Saudi Arabia higher than in the rest of the world reports?

5.6.6 Northern America

In Figure 28, Northern America, similar to Latin America & the Carribean, tends to follow the world trend as well. Music, comedy, and entertainment show access probabilities of 32%, 15%, and 15% respectively. The only other category above 5% access probability are People & Blogs at 8% and TV shows at 5%. It is interesting to note that People & Blogs has approximately the same access probability in Northern America as it does in the Commonwealth of Independent States, and both of the these regions have more access probability than any other region for this social video category.

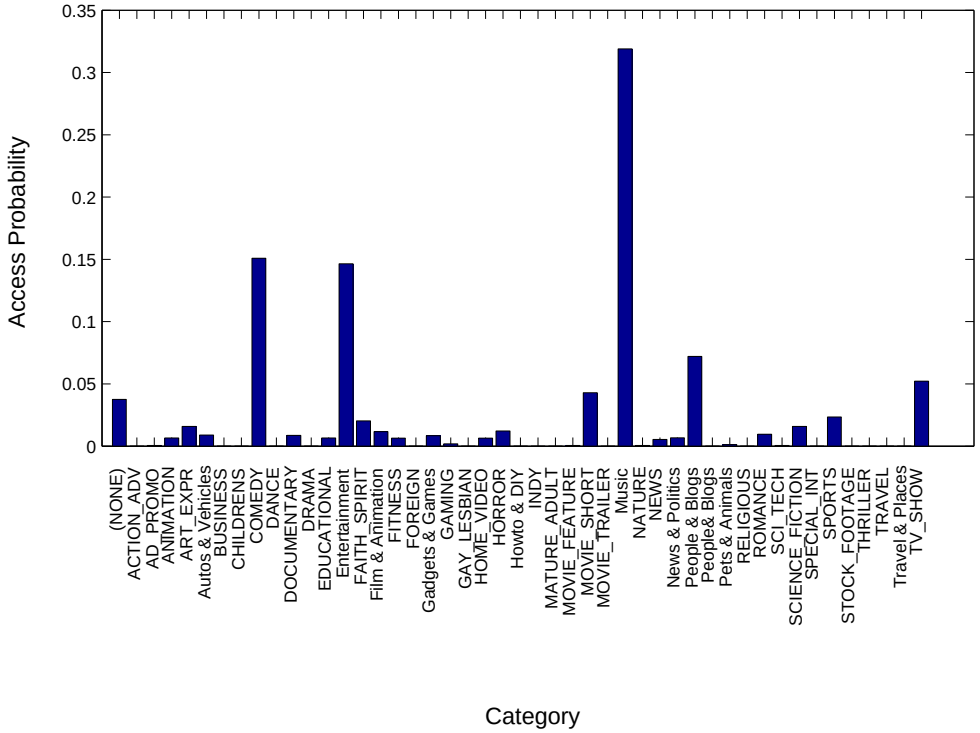
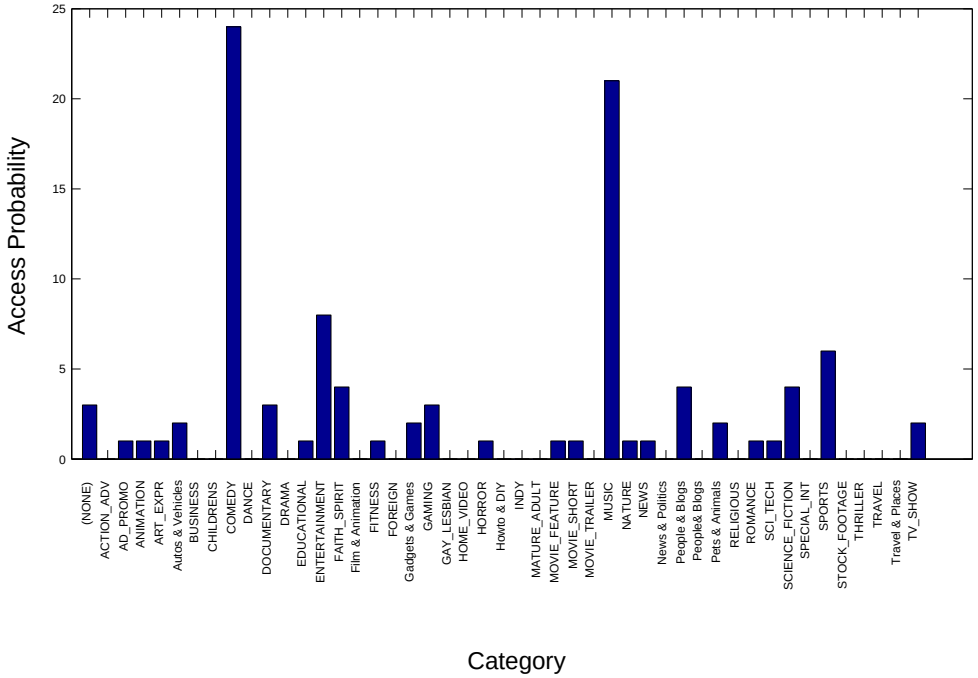
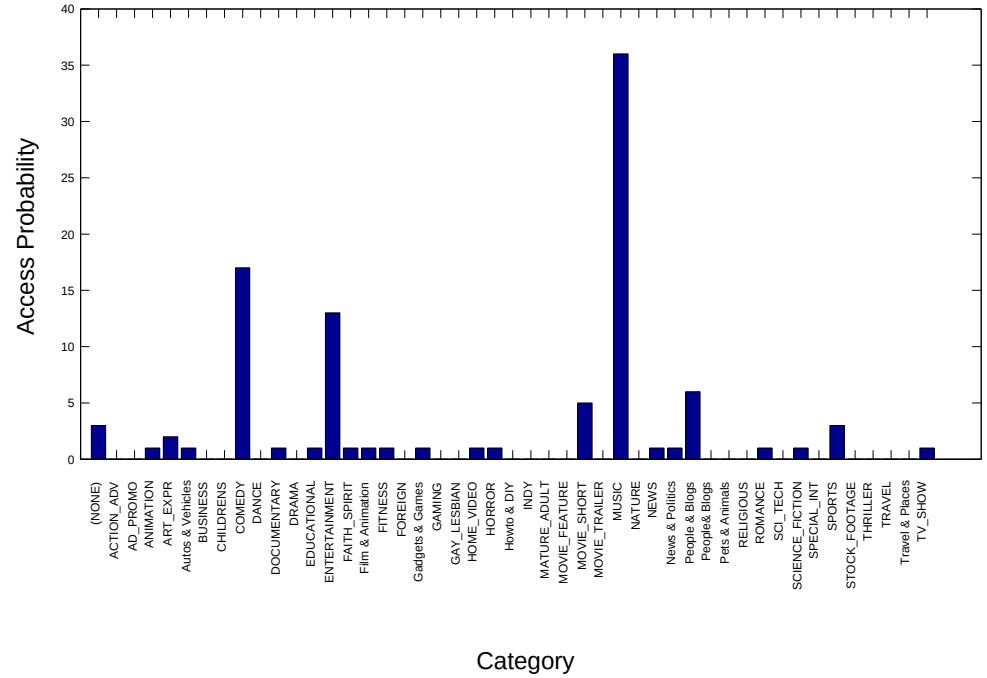


Figure 28: Northern America: Top 100 Google Videos

The Google data collected for Northern America is based on two countries: Canada and the United States. With the populations of the United States (approx. 301.1 million) and Canada (approx. 33.3 million) at a 9:1 margin, we can see in Figure 29 that the category distribution for the region does mostly follow the distribution for the United States. What is interesting to note, however, is that Canadians do watch a larger percentage of comedic videos in comparison to music videos, favoring comedic videos at a larger access probability than the United States. It is also interesting to note that while the United States has a higher affinity for music videos than does the Canadian population, the access probability for Canada is much more distributed amongst categories. Although the differences are small, categories such as faith & spirit, gaming, science fiction, sports, and TV shows do exhibit higher access probabilities in Canada compared to the United States.



(a) Canada



(b) United States

Figure 29: Northern America: Country-Level Category Comparison

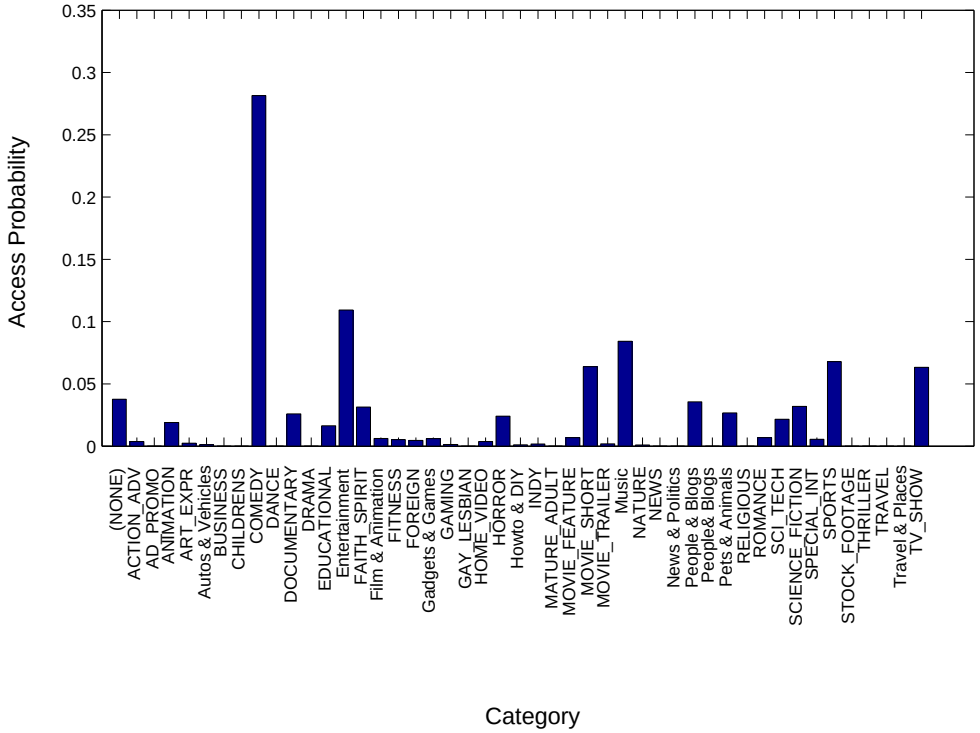
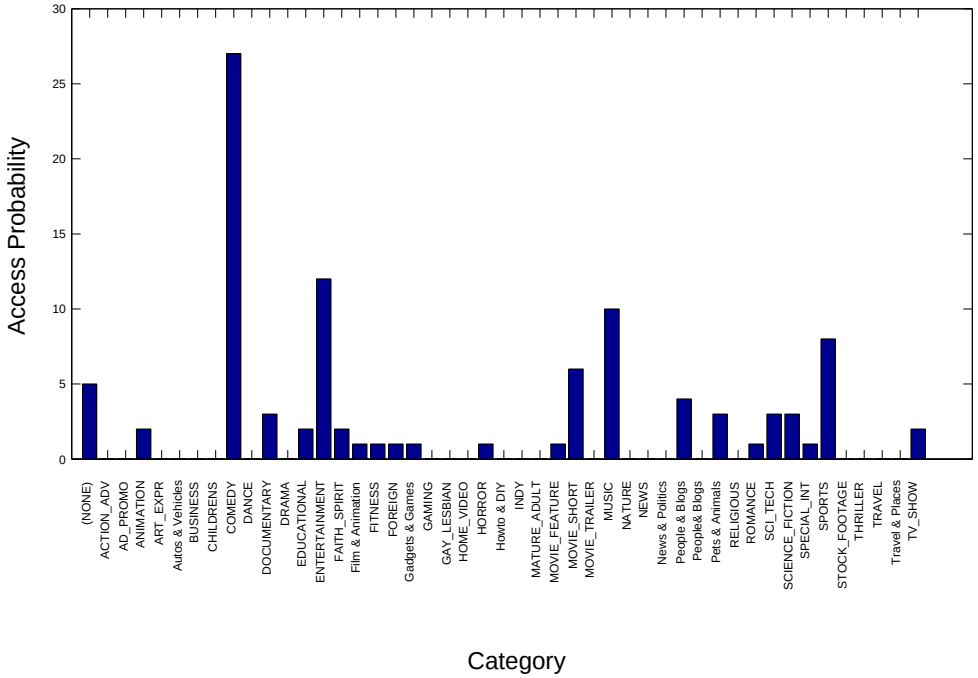


Figure 30: Oceania: Top 100 Google Videos

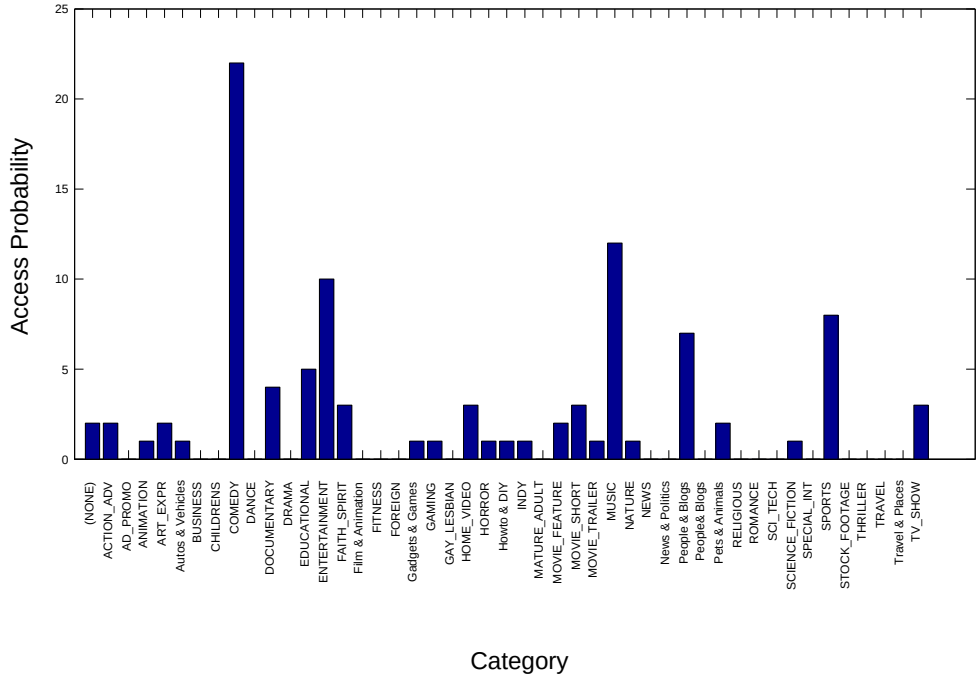
5.6.7 Oceania

Comedy is the most dominant social video category accessed in Oceania at 29%, followed by entertainment at 11% and music at 9% as seen in Figure 30. Other categories above the 5% access probability threshold include short movies, sports, and TV shows each at approximately 7%, respectively. While not exact, we do note approximate similarities between Eastern Europe and Oceania here, with each having comedy, music, short movies, sports, and TV shows near the top of their access rankings.

Oceania data presented in Figure 30 is composed of data from two countries: Australia and New Zealand. With a ratio of 5 to 1 between the populations of Australia and New Zealand, we note from Figure 31 that the top four categories in terms of access probability are common for each country - comedy (27% and 22%, respectively), entertainment (12% and 10%, respectively), music (10% and 13%, respectively), and sports (each 8%, respectively). In New Zealand, short movies are the only other category that shows more



(a) Australia



(b) United States

Figure 31: Oceania: Country-Level Category Comparison

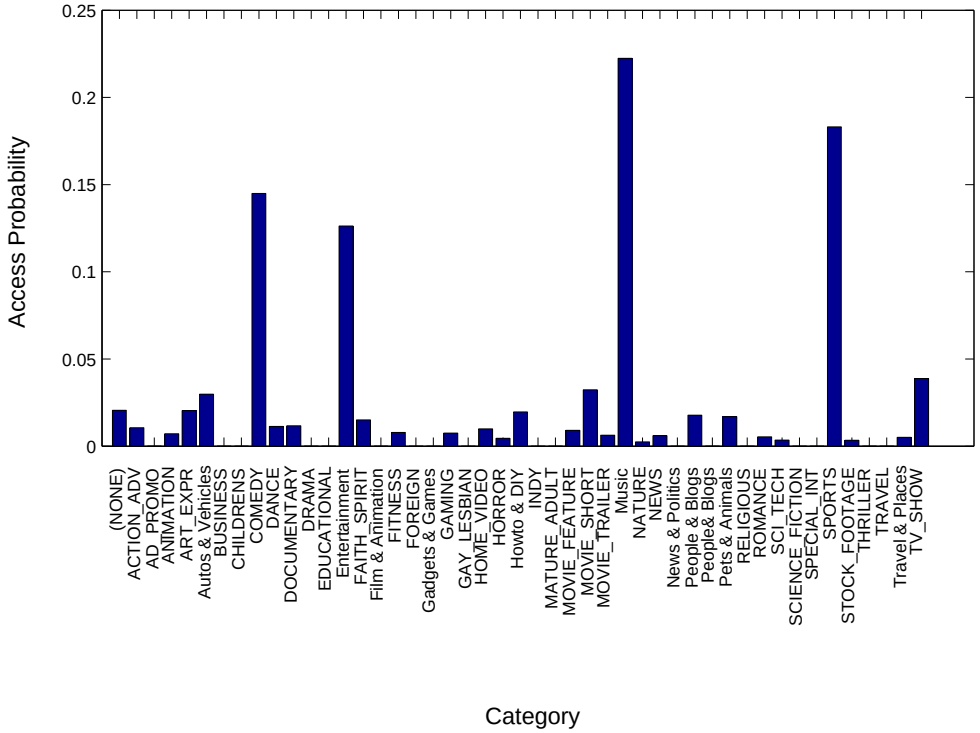


Figure 32: Sub-Saharan Africa: Top 100 Google Videos

than 5% access probability in Australia. Educational, as well as people and blogs, social videos are the only other categories which show more than 5% access probability. Due to their close proximity, it is not surprising that both Australia and New Zealand show such common social video interests, contributing to a consistent view of social video popularity in the region of Oceania.

5.6.8 Sub-Saharan Africa

It is clear from Figure 32 that while Sub-Saharan Africa views music videos most often, the access of sports videos as the second most viewed category at 18% is the most unique here among the other world regions. Comedy and entertainment video access at nearly 15% and 13% respectively follow the world trend. No other social video category is watched more than 5% in comparison, so it is clear that music, sports, comedy, and

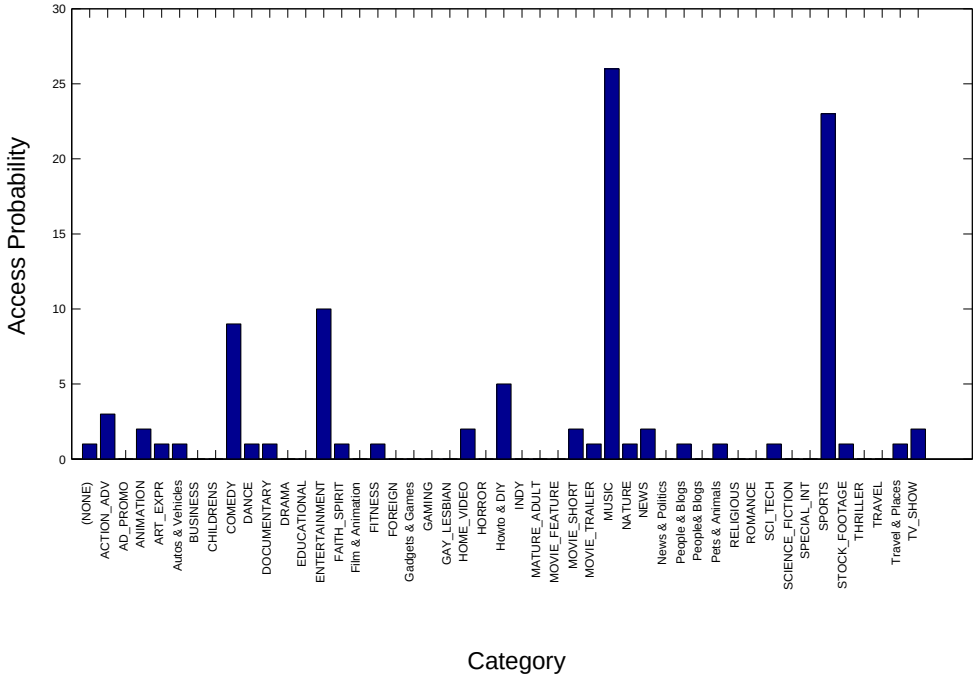
entertainment social videos are the most watched categories in this region.

Kenya and South Africa each contribute to the Sub-Saharan Africa data. When looking closer at each country in Figure 33, we are not suprised that music, sports, comedy and entertainment videos are significant in the access probabilities of each country. What we do see is that while the sports category is significant in each country, music and sports is more popular in Kenya compared to South Africa, with access probabilities in Kenya of 26% for music and 23% for sports. On the other hand, comedic social videos are much more popular in South Africa with 20% access probability compared to 9% in Kenya. The only other categories with more than 5% access probability in these countries include autos and vehicles in South Africa at 5% and how-to and do-it-yourself in Kenya at 6%.

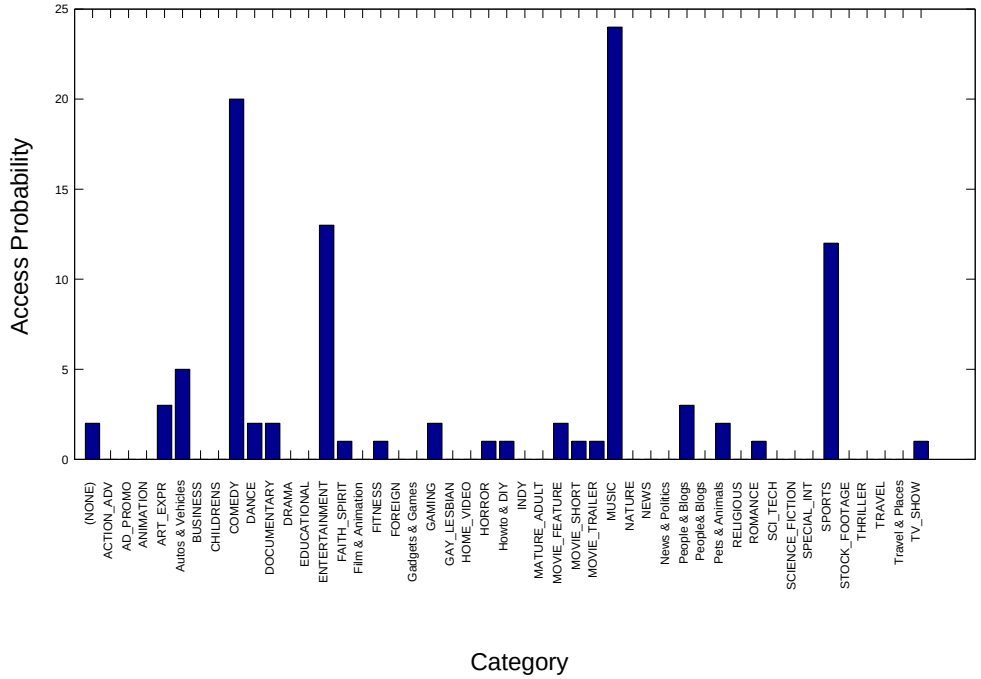
5.6.9 Western Europe

Western Europe has a similar distribution to that of Sub-Saharan Africa as seen in Figure 34; however, its access of music videos is approximately 32% in comparison to 23%. Comedy is accessed approximately 15%, sports approximately 10%, and entertainment videos approximately 8%. Again, like Sub-Saharan Africa, no other social video category is watched more than 5% in this region. One outstanding question remains: why are the social videos in Sub-Saharan Africa and Western Europe so similar in category, and, further, why is music accessed more in Western Europe than Sub-Saharan Africa in comparison?

Examining the two most populous countries in Western Europe, we see similar results in terms of the most popular category. Music accounts for 41% of social video access in Germany and 43% in France. Comedy is the next most popular social video category in Germany with 15% of the access probability, while France's next most popular category is sports at 14%. Feature movies, short movies, and movie trailers hold a collective



(a) Kenya



(b) South Africa

Figure 33: Sub-Saharan Africa: Country-Level Category Comparison

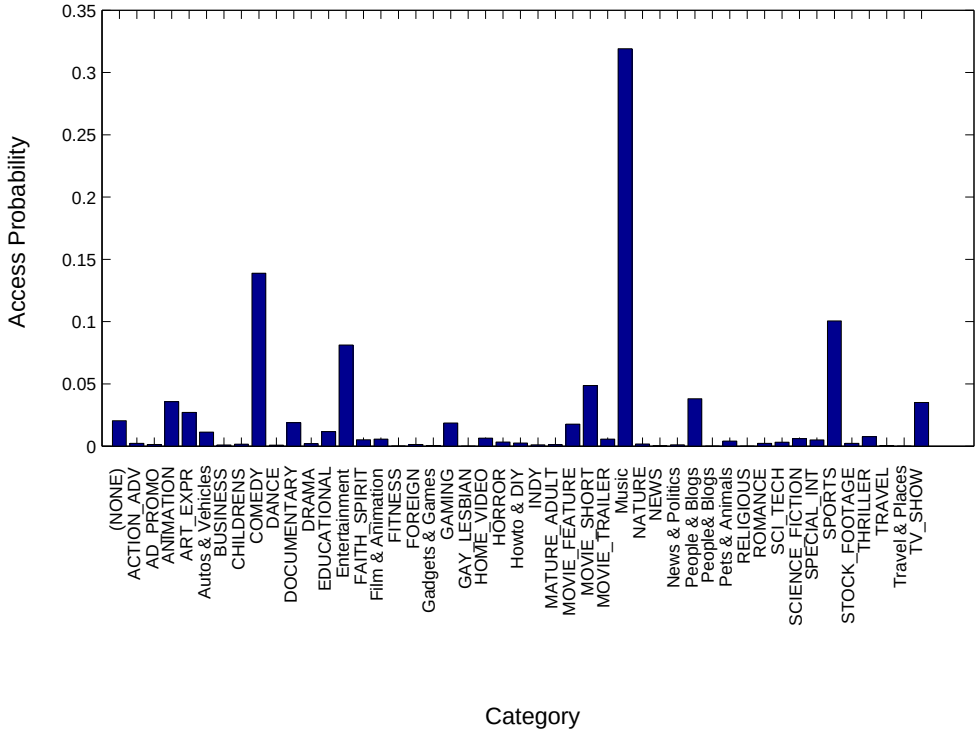
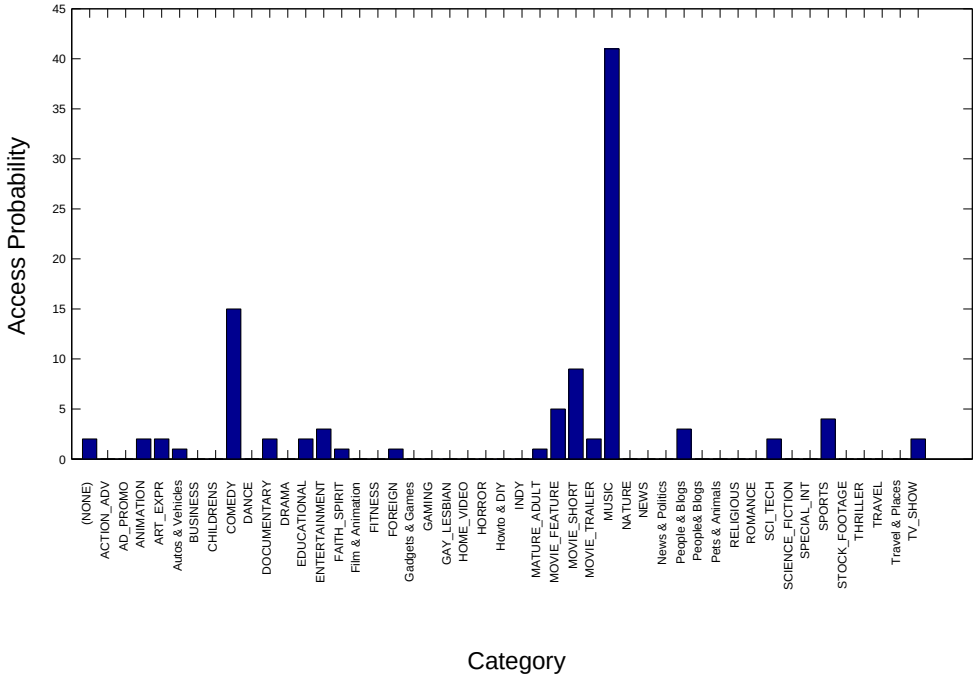


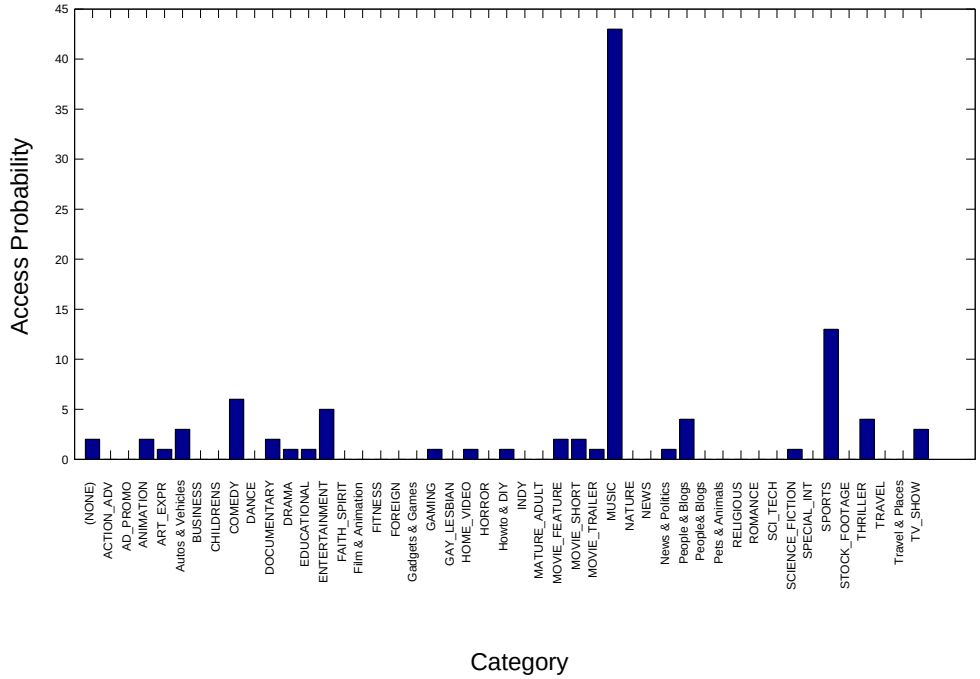
Figure 34: Western Europe: Top 100 Google Videos

17% of the category popularity in Germany, while France shows less interest in movies yet approximately 5% of its access probability in entertainment videos and 4% in thriller videos.

Further study around the reasons for these video choices need to be determined; however, we present these findings as a starting point for further analysis and study. Cultural study and sociological reasoning may indeed be necessary to analyze the importance of popular social video genres and their content. What we do present is the characterization of these video categories on popular videos around the world and emphasize the distribution of video served to those regions.



(a) Germany



(b) France

Figure 35: Western Europe: Country-Level Category Comparison

CHAPTER 6
CONCLUSIONS

This thesis has investigated social media and, more specifically, the characterization of online social video from a global perspective. We introduced the concept of social media, the purpose of our study, and an overview of how our study relates to past studies. We then provided a definition of social media, explained the types of social media that currenty exist, and reviewed popular online websites which fit within these social media categories. Next, we provided a detailed comparison of how our work relates to past studies. We mentioned studies which compared social video access probabilities to Zipf-like and stretched exponential distributions. Further, we reviewed studies which examined online video category, video access evolution, video length, and geographical aspects, mentioning the limitations of these past studies. We then presented our evaluation methodology and showed how we collected data from YouTube, Google Video, and Truveo Video Search using Ruby programming language scripts. Downloading data from YouTube twice a day for a period of forty-two weeks from the daily, weekly, monthly, and yearly Top 100 lists, we were able to process this data and extract key elements such as video duration, view count, category, and URL information. We used similar methods to collect data from Truveo Video Search and Google Video to produce data relevant to our study. Lastly, we presented our results from over ten months of data collection and demonstrated that video popularity on YouTube follows Zipf-like distribution with $t = 0.30$ for daily access and $t = 0.50$ for weekly, monthly, and yearly access, while video popularity from the Truveo Video Search data follows a Zipf-like distribution with $t = 0.20$. In the category analysis, we have demonstrated that music, comedic, and entertainment videos are the most popular. When correlating both category and video length, we have shown that the average video length of these popular videos have a significant effect on the total average video length of the YouTube and Truveo videos. Access patterns demonstrate an oscillating curve as video popularity varies per day of the week, while the most fre-

quent views happen on Fridays and Saturdays during 9PM and 1PM UTC. The rate of change in the access of the most popular videos exhibits medium negative correlation with the rank. Finally, we have provided extensive popularity distributions based on upon the nine regions of the world defined by the U.S. Census Bureau. Using Google Video data, we have provided comparative data on the video interests of those nine regions. We found that music dominates the social video popularity in Latin America and the Caribbean, Northern America, and Western Europe, while Russia and the Ukraine have a unique interest for social video documentaries.

We contribute the findings in this paper as a starting point for future study. Through these findings, we hope to encourage future work and study within the topic of social video characterization. There are many opportunities to explore. First, more data collection from YouTube, Google Video, Truveo Video Search, and other social video websites over a period of the next several years will only strengthen and improve the findings claimed in this paper. We have noticed the need for social video research over long periods of time to strengthen hypotheses and social video characterization claims. While social video is a rather new form of social media, it is clear that its popularity requires intense study. Second, there should be attention paid to the interaction of individual users on social video websites. While not covered in this paper, understanding how individual users interact with social video content may be important in understanding why certain videos become more popular than others. Furthermore, understanding how search engines increase or decrease this popularity would also be an interesting area of study. The method to discovering social video may be a key element when characterizing workload on social video websites. Lastly, we strongly feel that collecting more geographical data around social video and looking for answers for these results in the realm of social studies would be a worthwhile exercise. Using social video data from each geographical region, there is much analysis that can be performed in conjunction with the culture, sociology, economics, and other social scientific methods in a particular region to explain reasons

for social video trends. Why is music so popular in Northern America, for example, yet not as popular in Oceania compared to comedy? Do how-to and do-it-yourself videos show a willingness of the Kenyan people to learn from earch other through social video? There are a large amount of questions such as these that still need to be answered as a result of our findings. We welcome the ability to participate in such an important study in the future.

REFERENCES

[1] Alexa, the web information company. http://www.alexa.com.

[2] Blogger. http://www.blogger.com.

[3] Breakthrough internet device. http://www.apple.com/iphone/features/index.html#internet.

[4] Delicious. http://www.delicious.com.

[5] Digg. http://www.digg.com.

[6] Flickr growth and bumped images. http://www.flickr.com/photos/gustavog/2324101087/.

[7] Google-video project. http://rubyforge.org/projects/google-video/.

[8] Linkedin. http://www.linkedin.com.

[9] Online gaming revenues to triple by 2009. http://www.parksassociates.com/press/press_releases/2005/gaming-1.html.

[10] Takingitglobal. http://www.takingitglobal.org.

[11] Truveo video search. http://developer.truveo.com/index.php.

[12] U.s. census bureau international data base. http://www.census.gov/ipc/www/idb/tables.html.

[13] Wikipedia. http://www.wikipedia.org.

[14] Youtube project. http://rubyforge.org/projects/youtube.

[15] S. Acharya, B. Smith, and P. Parns. Characterizing user access to videos on the world wide web. In *Multimedia Computing and Networking Conf. (MMCN)*, January 2000.

[16] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[17] J. M. Almeida, J. Krueger, D. Eager, and M. Vernon. Analysis of educational media server workloads. In *11th international workshop on network and operating systems support for digital audio and video (NOSSDAV)*, pages 21–30, June 2001.

[18] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the worlds largest user generated content video system. In *ACM Internet measurement Conference (IMC)*, San Diego, CA, October 2007.

[19] L. Cherkasova and M. Gupta. Characterizing locality, evolution, and life span of accesses in enterprise media server workloads. In *12th Int'l. Workshop on Network and Operating System Support for Digital Audio and Video (ACM NOSSDAV 2002)*, 2002.

[20] M. Chesire, A. Wolman, G. Voelker, and H. Levy. Measurement and analysis of a streaming media workload. In *USENIX Symposium on Internet Technologies and Systems (USITS)*, pages 1–12, March 2001.

[21] F. Duarte, F. Benevenuto, V. Almeida, and J. Almeida. Geographical characterization of youtube: a latin american view. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*, pages 13–21, Washington, DC, USA, 2007. IEEE Computer Society.

[22] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic characteristics and communication patterns in blogosphere. In *International Conference on Weblogs and Social Media*, 2007.

[23] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Characterizing user sessions on youtube. In *Multimedia Computing and Networking (MMCN)*, San Jose, CA, USA, January 2008.

[24] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang. Does internet media traffic really follow zipf-like distribution? *SIGMETRICS Perform. Eval. Rev.*, 35(1):359–360, 2007.

[25] M. J. Halvey and M. T. Keane. Exploring social dynamics in online media sharing. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1273–1274, New York, NY, USA, 2007. ACM.

[26] C.-H. Lin, L.-Y. Li, W.-C. Hu, G.-D. Chen, and B.-J. Liu. Constructing an authentic learning community through wiki for advanced group collaboration and knowledge sharing. *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, pages 342–344, July 2007.

[27] P. Svoboda, W. Karner, and M. Rupp. Traffic analysis and modeling for world of warcraft. *Communications, 2007. ICC '07. IEEE International Conference on*, pages 1612–1617, June 2007.

[28] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: Youtube network traffic at a campus network measurements and implications. In *Multimedia Computing and Networking (MMCN)*, San Jose, CA, USA, January 2008.

ABSTRACT

CHARACTERIZATION OF SOCIAL VIDEO

by

JEFFREY R. OSTROWSKI

December 2008

Advisor: Nabil J. Sarhan

Major: Computer Engineering

Degree: Master of Science

The popularity of social media has grown dramatically over the World Wide Web. In this thesis, we provide an overview of the types of social media. We then analyze the video popularity distribution of well-known social video websites (YouTube, Google Video, and the AOL Truveo Video Search engine) and characterize their workload up to an eight month period. We identify trends in the categories, lengths, and formats of those videos, as well as characterize the evolution of those videos over time. We further provide an extensive analysis and comparison of video content categories amongst the main regions of the world, uncovering region and country-specific trends. Music dominates social video popularity when viewing world data, while music, comedy and entertainment are categories which are consistently popular at the region level.

AUTOBIOGRAPHICAL STATEMENT

JEFFREY R. OSTROWSKI

EDUCATION

- Bachelor of Science in Electrical Engineering, May 2003

  Wayne State University, Detroit, MI, USA

PUBLICATIONS

1. Characterization of Social Video, Accepted by Multimedia Computing and Networking (MMCN) 2009.