The Long Tail of Web Video

Luca Rossetto and Heiko Schuldt

Databases and Information Systems Research Group
Department of Mathematics and Computer Science
University of Basel, Switzerland
{firstname.lastname}@unibas.ch

Abstract. Web Video continues to gain importance not only in many areas of computer science but in society in general. With the growth in numbers, both of videos, viewers, and views, there arise several technical challenges. In order to address them effectively, the properties of Web Video in general need to be known. There is however comparatively little analysis of these properties. In this paper, we present insights gained from the analysis of a data set containing the meta data of over 100 million videos from YouTube. We were able to confirm common wisdom about the relationship between video duration and user engagement and show the extreme long tail of the distribution of video views overall. Such data can be beneficial in making informed decisions regarding strategies for large scale video storage, delivery, processing and retrieval.

1 Introduction

Web video continues to grow, not only in quantity but also in importance. With the continued improvements in video recording technology, recording devices not only become cheaper and therefore more common, the videos created by these devices also increase in frame rate and resolution, further amplifying the data growth. This growth produces challenges across many fields, not only for video storage but also for processing, analysis, delivery, and retrieval. While a lot of research is conducted to overcome these challenges, little is known about the root cause, the videos as a whole. In this paper, we present an analysis based on a large set on web video meta data from YouTube[1]. We show that many properties of such web video as found in the wild show a long-tail distribution which has relevant consequences for many applications involving such videos.

The remainder of this paper is structured as follows: Section 2 presents other analyses of YouTube video done in the past and Section 3 outlines the methods employed in our analysis. Section 4 presents the results which are discussed in greater detail in Section 5. Section 6 concludes.

[1] https://youtube.com

2 Related Work

There are a few datasets which are built from video material on YouTube, and these are mostly built with a very specific purpose in mind. The newest and largest one as of September 2017 is the YouTube-8M [1] dataset published by Google Research. It was compiled with the intention to further the research in the field of video understanding and it contains various features extracted from labeled videos. It does not, however, include the videos themselves nor their meta data which renders the dataset unsuitable for our analysis.

There have been efforts in the past to analyze overall properties of the videos on YouTube [3–5, 9]. For all of those, a custom crawler was used to gather video meta data from the site itself which was subsequently analyzed. Most of these efforts used the data from one or several million videos and are already a few years old. The most recent of these analysis is based on meta data from roughly 100 million videos [9].

Additionally, the network implications of streaming video from YouTube have been analyzed several times over the years [2, 6–8, 10]. Some of those works also considered the video meta data in a similar way to the ones mentioned above. Since these measurements are however made from a point close to the edge of the Internet, the data produced does not have a strong claim for being representative of the distribution found on the entire platform since it is biased by the video watching habits of the users of the particular network in question.

3 Methods

The data used in this paper was first published in [9] and consists of the meta data collected from 20 million videos from vimeo[2] as well as from 100 million videos from YouTube. This data was collected in 2016 by a purpose-built distributed crawling setup and made available[3] together with the paper [9] as PostgreSQL[4] database export. For our analysis, we limit ourselves to the YouTube part of the data, not only because it is the larger part but also because it contains richer meta information including view- and like-count which the vimeo part does not. Table 1 shows the schema of the used data. The fields we focus on in this analysis are the numerical values *duration*, *views*, *likes*, *dislikes* as well as the age[5] of the video. In what follows, we show analysis results and aggregations on the basis of this data. Some of the relations between these properties were already analyzed in [9] which we will not repeat here.

[2] https://vimeo.com
[3] http://download-dbis.dmi.unibas.ch/WWIN/
[4] https://www.postgresql.org/
[5] With 'age', we denote the difference in days between the date the video was uploaded and the date the metadata of the video was harvested.

Table 1: The schema of the dataset used for the analysis

Property Name	Description
id	11 character string which uniquely identifies a video on YouTube
name	Name of the video as shown on the page
description	Description of the video as shown on the page
license	Binary value: true for a creative commons licensed video
duration	The video duration in seconds as an integer
upload_date	The day on which the video was uploaded to YouTube
views	Number of views for the video at the time of crawling
likes	Number of likes for the video or -1 if disabled
dislikes	Number of dislikes for the video or -1 if disabled
subtitles	Number of subtitles available for the video
unlisted	Binary value: true if the site reports the video as not publicly listed
family_friendly	false if the video contains content deemed offensive
crawl_date	The day on which the crawler generated this particular entry
channel	24 character string uniquely identifying the channel of the video
genre	The selected genre (out of the 18 available)

4 Results

In this section, we present the performed analysis as well as the results obtained by them. First, we will analyze how certain metrics behave with respect to the age of a video. Next, we will look at the relationship between a video's popularity and its duration and finally, we will see how long-tailed the distributions in video popularity actually are.

4.1 Changes over Time

We start by observing the relationship between the age of a video and the number of times it has been viewed within this time period. If one assumed a similar popularity for all videos and therefore a uniform random distribution of views across all videos, one would expect a roughly linear relationship between the number of days for which a video was available for viewing and the number of views it collected. As can clearly be seen in Figure 1, this is not the case.

The figure shows a density heat-map with respect to video age and the number of views accumulated during that time. The color is proportional to the logarithm of the density, changing from blue for small values to red for large values. With the exception of some artifacts on the left side of the figure, the distribution of views per video appears to change very little over time. For the most recent 4 years, it appears to be mostly independent of the actual age of the video. The changes which can be seen on the right side of the figure are less due to the video age as such and more due to the comparatively fewer videos in existence (or at least present within the used dataset) which are of such high age. This interpretation is further supported by Figure 2 which shows the daily

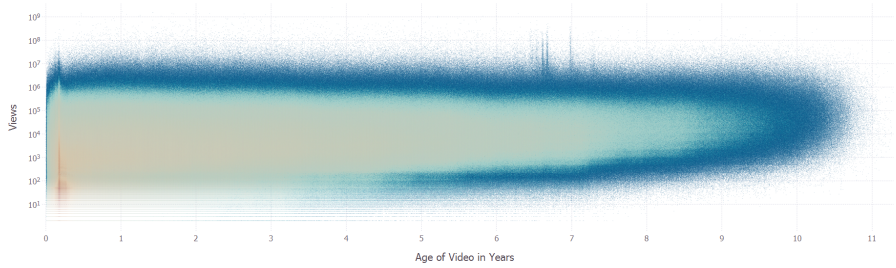

Fig. 1: Distribution of views per video with respect to video age

mean (blue) and median (green) aggregation of views per video, again with respect to the videos' age. It can be seen that a vast majority of views must be produced relatively shortly after a video is uploaded. After this initial period, the aggregated view count stays relatively stable with only a slight increase over time. The median aggregation rendered in green shows a more prominent trend towards view accumulation over time.

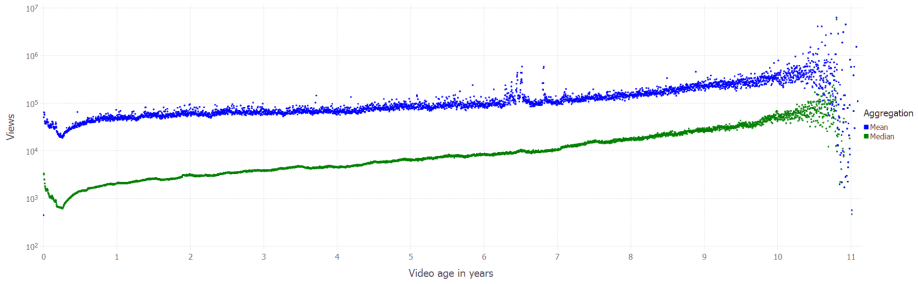

Fig. 2: Daily mean and median aggregation of views per video with respect to video age for the entire time covered by the dataset

The dip in both traces in Figure 2 which can be seen on the left of the plot can be explained by the higher number of relatively recent videos in the dataset which also caused the artifact in Figure 1.

Since videos appear to already start out with most of the views they will receive, we can focus our attention to Figure 3 which shows the same data as Figure 2 but zoomed in on the first 100 days after a video's initial upload. It can be seen that the vast majority of views is accumulated in the first few days after its upload. The data is rather sparse for the first few days, which is probably due to the way the crawler used to collect it found videos to visit. It is therefore not possible to make meaningful statements about the first few days of a video's presence on the platform.

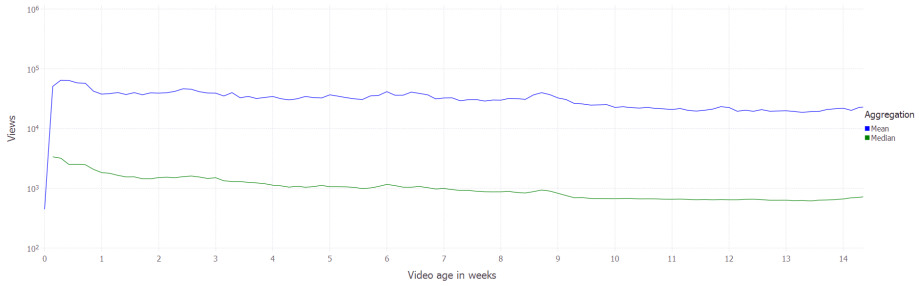

Fig. 3: Daily mean and median aggregation of views per video with respect to video age, limited to videos at most 100 days old

The data used there is not perfectly suited for the analysis presented up until now as it only contains the view count for a video at exactly one point in time. In order to have a more reliable analysis, a dataset would ideally contain several view counts of a video from different points in time. In the absence of such a dataset, we will content ourselves with this analysis.

When looking at the accumulation of likes over the time, we see a similar independence of video age as with the views. Figure 4 shows the distribution of video likes with respect to the age of the video. As with the views depicted above, the distribution of likes appears to be mostly independent of the age of the video, at least when only considering videos with an age sufficiently large to put them out of the initial accumulation period.

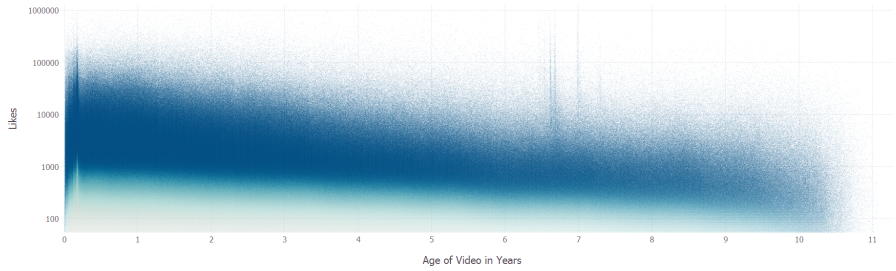

Fig. 4: Distribution of likes per video with respect to video age

This initial accumulation period is indeed the same as for the views, which can be seen in Figure 5 which shows the mean views, likes and dislikes of a video with respect to its age. Note that the blue points are the same as the ones in Figure 2. The cyan and purple points show the aggregated likes and dislikes per video, respectively.

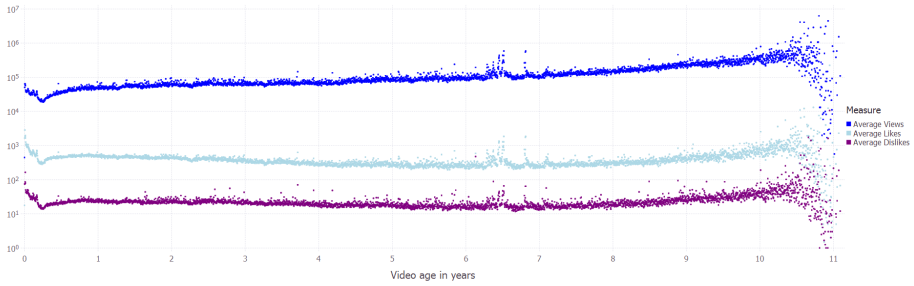

Fig. 5: Aggregated mean of views, likes and dislikes per video with respect to video age

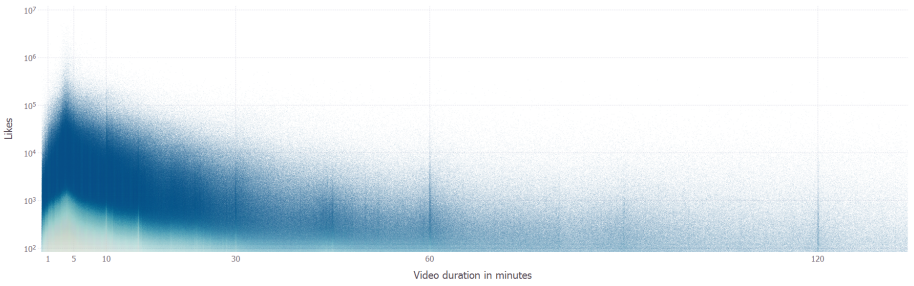

Fig. 6: Distribution of likes per video with respect to video duration

We can see that the shapes of the three traces are very similar and none of them shows a substantial dependency on the video age. There also appears to be a close to constant factor between the three measures. Independently of the age of the video, there is approximately one like for every 331 views[6] and one dislike for every 17 likes[7].

4.2 Duration

In this section, we will look into the influence of a video's duration over how often it is watched or liked. Figure 6 shows the distribution of video likes with respect to video duration.

It can be seen that, in accordance with commonly repeated recommendations often found in the context of YouTube video creation, the most-liked videos are somewhere between 3 and 4 minutes in length. Note that this distribution is not normalized with respect to views per video or numbers of videos per duration. Figure 7 shows the aggregated average of views, likes, and dislikes with respect to the video duration, similar to the way shown in Figure 5 with respect to video age. In contrast to Figure 5 which shows the entire available data range, the data

[6] average(views / likes): 332.9592, median(views / likes): 331.0722
[7] average(likes / dislikes): 17.2896, median(likes / dislikes): 16.8257

shown in Figure 7 was limited to videos with a total duration of at most 150 minutes, after which the data becomes too sparse for a meaningful aggregation.

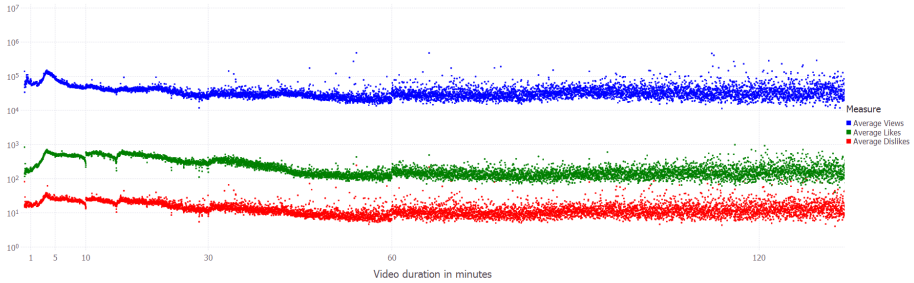

Fig. 7: Aggregated mean of views, likes and dislikes per video with respect to video duration

Like before, we can see three traces with similar shapes and a seemingly constant factor between them. If we again estimate this factor, we get roughly 200 views per like[8] and 15 likes per dislike[9]. The fact that these values not only differ from the ones above but also from the ratios of the entire dataset which has on average 162.25 views per like and 18.54 likes per dislike indicates that in both cases, the data was not entirely uncorrelated with respect to the aggregation. Figure 7 also confirms that videos with a duration of between 3 and 4 minutes appear to receive more views than both longer and shorter videos. The difference in expected likes and dislikes is even more substantial for shorter videos.

4.3 Genres

Tables 2 and 3 show the breakdown of views, likes, and dislikes grouped by *genre* and aggregated by average and median, respectively. It can be seen that there are substantial differences between genres in all three measures. The most 'liked' videos come from the *Gaming* genre, according to both average and median aggregations. Most 'disliked' are *Shows* and *Movies* which form the top two in number of dislikes per video in both aggregation, albeit not in a consistent order. Interestingly, those genres, together with the *Trailers* form the top three in terms of views per video in both aggregations, followed by *Music* on place four.

While not towards the top in terms of views or likes per video, members of the *Howto & Style* genre show a high view to like ratio in both aggregations, indicating that these videos while not excessively likely to be viewed are more likely to be 'liked' by their viewers. This discrepancy could also be due to the fact that any user with an account on YouTube can watch a video multiple times and

[8] average(views / likes): 201.6157, median(views / likes): 195.7886
[9] average(likes / dislikes): 15.4702, median(likes / dislikes): 14.6817

Table 2: Average Views, Likes and Dislikes per video with respect to Genre

Genre	Videos	Views	Likes	Dislikes	$\frac{Views}{Likes}$	$\frac{Views}{Dislikes}$	$\frac{Likes}{Dislikes}$
People & Blogs	21'507'972	33'496.2	211.49	14.13	158.38	2'370.46	14.97
Entertainment	13'789'270	84'829.4	5'11.52	29.86	165.84	2'840.63	17.13
Music	11'256'684	184'613.8	738.22	34.26	250.08	5'388.29	21.55
News & Politics	107'30'269	15'341.5	57.57	7.22	266.47	2'123.66	7.97
Education	6'619'603	33'029.5	114.38	10.85	288.77	3'044.13	10.54
Gaming	6'598'415	78'590.74	1'351.64	46.99	58.14	1672.46	28.76
Sports	6'040'181	35'518.5	141.19	8.43	251.57	4'212.42	16.74
Howto & Style	4'285'544	54'007.3	542.14	24.42	99.62	2'211.75	22.20
Film & Animation	4'202'044	105'460.2	345.33	29.11	305.39	3'622.93	11.86
Autos & Vehicles	3'926'800	36'142.3	104.10	9.18	347.18	3'935.76	11.33
Travel & Events	3'083'467	18'039.8	72.23	4.66	249.76	3'874.88	15.51
Science & Technology	2'900'186	36'040.6	220.03	15.69	163.80	2'296.72	14.02
Comedy	2'630'365	135'632.6	1'184.19	59.40	114.54	2'283.21	19.93
Nonprofits & Activism	2'536'371	13'154.1	74.28	5.79	177.10	2'273.59	12.84
Pets & Animals	1'235'834	59'092.5	169.90	13.46	347.80	4'390.05	12.62
Shows	323'115	288'280.4	1'181.55	99.45	243.99	2'898.63	11.88
Movies	18'681	194'249.4	307.42	69.01	631.86	2'814.99	4.46
Trailers	7'407	196'615.2	226.94	24.60	866.39	7'993.99	9.23

therefore generate an arbitrary amount of views but only 'like' or 'dislike' a video once. A low view to like ratio could therefore either mean that videos are more likely to be liked when watched or that they are less likely to be watched several times by the same user. The data at hand does not enable us do differentiate between these cases.

Table 4 shows the average and median video duration by genre. The longest videos are unsurprisingly labeled as *Movies* while *Trailers* are the shortest. Compared to these two extremes, the difference in duration between the other genres is relatively minor. There is however a difference between the two aggregations. The last column in Table 4 shows the time difference between the average and the median duration. For all genres except *Movies*, the average video is longer than the median one, often by a significant fraction.

Such discrepancies between the two aggregations can also be observed in Tables 2 and 3 where the values generated by the average are usually substantially larger than the median values. This indicated a skewed distribution with a small number of large values and a large number of small values, commonly referred to as a *long-tail* distribution.

4.4 Long tails

In this last part of our analysis, we look at how views, likes, and dislikes are distributed across all videos. We have already seen above that it would be incorrect to assume a completely uniform distribution of views, etc. since this would

Table 3: Median Views, Likes and Dislikes per video with respect to Genre

Genre	Videos	Views	Likes	Dislikes	$\frac{Views}{Likes}$	$\frac{Views}{Dislikes}$	$\frac{Likes}{Dislikes}$
People & Blogs	21'507'972	1'962	10	1	196.20	1'962.00	10.00
Entertainment	13'789'270	4'894	18	1	271.89	4'894.00	18.00
Music	11'256'684	10'309	40	1	257.73	10'309.00	40.00
News & Politics	10'730'269	1'206	4	0	301.50	-	-
Education	6'619'603	1'888	9	0	209.78	-	-
Gaming	6'598'415	7'203	74	4	97.34	1'800.75	18.50
Sports	6'040'181	3'013	10	0	301.30	-	-
Howto & Style	4'285'544	5'615	37	2	151.76	2'807.50	18.50
Film & Animation	4'202'044	5'620	19	1	295.79	5'620.00	19.00
Autos & Vehicles	3'926'800	4'156	11	1	377.82	4'156.00	11.00
Travel & Events	3'083'467	1'766	6	0	294.33	-	-
Science & Technology	2'900'186	3'448	11	1	313.45	3'448.00	11.00
Comedy	2'630'365	7'961	30	2	265.37	3'980.50	15.00
Nonprofits & Activism	2'536'371	976	1	0	976.00	-	-
Pets & Animals	1'235'834	3'621	11	1	329.18	3'621.00	11.00
Shows	323'115	19'416	52	5	373.38	3'883.20	10.40
Movies	18'681	22'458	39	9	575.85	2'495.33	4.33
Trailers	7'407	25'888	1	0	25'888.00	-	-

lead to a strong correlation between video age and view count which we did not observe.

Figure 8 shows what fraction of videos, sorted by most viewed in descending order, is needed to account for that part of all views within the dataset. It is rather surprising to see that already the most viewed 0.07‰ videos account for 10% of all views. Increasing this fraction to 1‰ encompasses 26.68% of the view total, meaning that the remaining 99.9% of videos only produced 73.32% of all views. With 0.73%, we can even account for half of all views and rounding up to 1% of videos gives us an additional 4.4% of views, meaning that the bottom 99% of videos are accountable for 45.56% and therefore not even half the views.

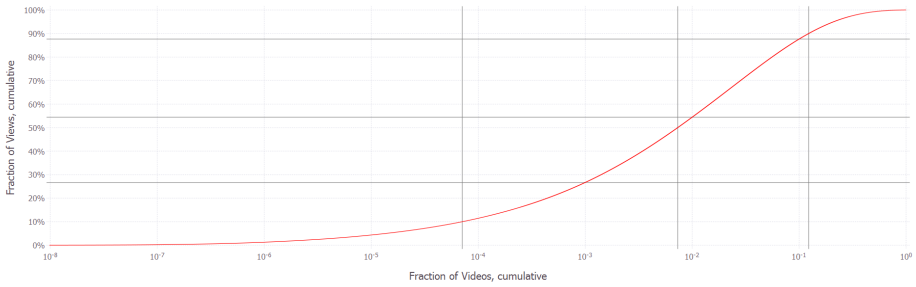

Fig. 8: Cumulative fraction of views per fraction of videos

Table 4: Average and Median video duration by Genre

Genre	Average	Median	Difference
Trailers	00:02:34	00:02:08	00:00:26
Autos & Vehicles	00:08:12	00:04:35	00:03:37
Pets & Animals	00:08:50	00:03:32	00:05:18
Comedy	00:09:31	00:04:35	00:04:56
Howto & Style	00:10:24	00:06:50	00:03:34
Music	00:10:29	00:04:34	00:05:55
Travel & Events	00:12:04	00:05:54	00:06:10
Sports	00:12:30	00:05:12	00:07:18
News & Politics	00:14:04	00:04:31	00:09:33
Entertainment	00:14:12	00:05:29	00:08:43
Science & Technology	00:14:44	00:06:14	00:08:30
People & Blogs	00:15:24	00:06:24	00:09:00
Film & Animation	00:17:28	00:05:52	00:11:36
Shows	00:18:22	00:11:08	00:07:14
Gaming	00:23:24	00:13:47	00:09:37
Education	00:24:06	00:10:00	00:14:06
Nonprofits & Activism	00:25:43	00:10:00	00:15:43
Movies	01:54:06	02:03:35	-00:09:29

Increasing this ratio further to 10% of videos gives us 87.64% of views and in order to get nine in ten views, we would need the top 12.29% of videos. The picture for likes and dislikes is rather similar, as depicted in Figures 9 and 10 respectively.

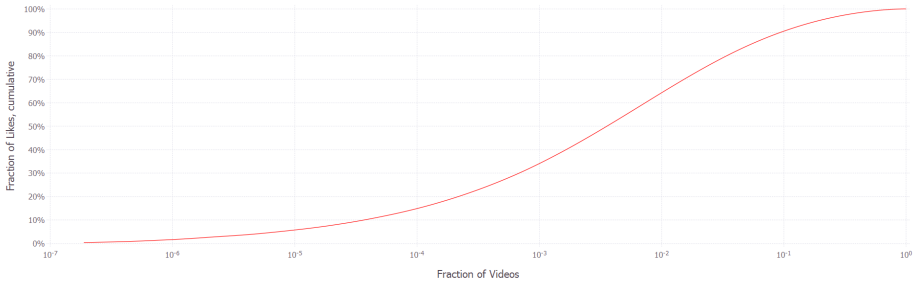

Fig. 9: Cumulative fraction of likes per fraction of videos

5 Discussion

Multiple conclusions can be drawn from the analysis presented above. The data appears to confirm that new content is more relevant than old content, which, by

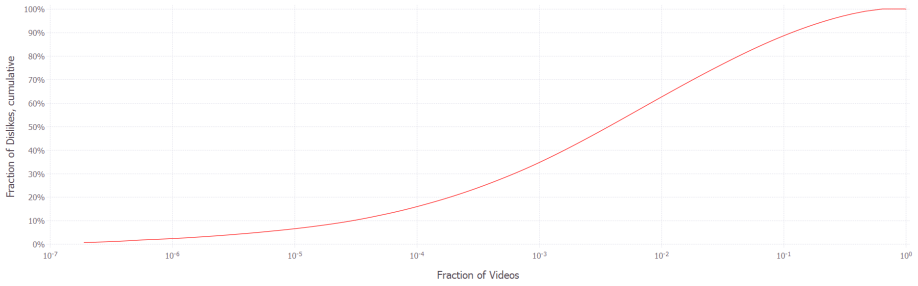

Fig. 10: Cumulative fraction of dislikes per fraction of videos

itself, is not very surprising. Having this intuition empirically confirmed however enables informed decisions about data storage and caching policies. It can be argued that it is not only more important to treat recently accessed data as hot data but also to prime the caches with newly added content in anticipation of its consumption. Since the used dataset holds no information about global access patterns for individual videos over time, let alone regional ones, we cannot draw any conclusions about useful distribution strategies from this data.

We were able to confirm a long standing common recommendation about video duration since, for this dataset, videos with a duration of about three minutes show on average more views and 'engagement' as measured by the number of likes and dislikes. This trend, while not very substantial, can again be used to inform decisions about video caching and processing. Since video duration is strongly correlated with its file size, choosing a caching policy which favors videos in this popular duration range could lead to more efficient use of storage. Similarly, since video processing effort is commonly dependent on video duration, employing a similar strategy could also prove beneficial.

A further interesting insight gained from this data is the extremely long-tailed distribution of views, likes, and dislikes per video. If YouTube were for some reason unable to deliver 9 out of every 10 videos, they could still retain over 87% of views. Since shorter videos produce commonly more views than longer ones and longer videos require more storage space, deleting these 90% would presumably free up substantially more than 90% of the current storage requirements. Since we have no way to predict a video's popularity in advance, we would however advise against this data deletion strategy.

6 Conclusion

In this paper, we have presented insights gained from the analysis of a large set of web video meta data. We have shown that, at least in the case of YouTube, video popularity as expressed by views and likes (as well as dislikes) exhibits a substantially long-tailed distribution in which a small percentage of the most popular videos accounts for a vastly over-proportional fraction of views. The

data also suggests that video age might be a usable indicator for interest in a particular video as most views appear to be generated shortly after the video publication. Further analysis of data which contains multiple samples per video from different points in time would however be required to make more conclusive statements regarding this topic. This analysis aims at helping to make informed decisions when working with large scale multimedia in general and video in particular, not only in the areas of video storage and distribution but also for video processing, analysis, and retrieval.

Acknowledgements

This work was partly supported by the Chist-Era project IMOTION with contributions from the Swiss National Science Foundation (SNSF, contract no. 20CH21_151571).

References

1. Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. *CoRR*, abs/1609.08675, 2016.
2. Pablo Ameigeiras, Juan J Ramos-Munoz, Jorge Navarro-Ortiz, and Juan M Lopez-Soler. Analysis and Modelling of YouTube Traffic. *Transactions on Emerging Telecommunications Technologies*, 23(4):360–377, 2012.
3. Xianhui Che, Barry Ip, and Ling Lin. A Survey of current YouTube Video Characteristics. *IEEE MultiMedia*, 22(2):56–63, 2015.
4. Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and Social Network of YouTube Videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238. IEEE, 2008.
5. Xu Cheng, Jiangchuan Liu, and Cameron Dale. Understanding the Characteristics of Internet Short Video Sharing: A YouTube-based Measurement Study. *IEEE Transactions on Multimedia*, 15(5):1184–1194, 2013.
6. Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. YouTube Traffic Characterization: a View from the Edge. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 15–28. ACM, 2007.
7. Géza Horváth and Péter Fazekas. Characterization and Modelling of YouTube Traffic in Mobile Networks. *ICN 2015*, page 127, 2015.
8. Juan J Ramos-Muñoz, Jonathan Prados-Garzon, Pablo Ameigeiras, Jorge Navarro-Ortiz, and Juan M López-Soler. Characteristics of Mobile YouTube Traffic. *IEEE Wireless Communications*, 21(1):18–25, 2014.
9. Luca Rossetto and Heiko Schuldt. Web Video in Numbers – An Analysis of Web-Video Metadata. *CoRR*, abs/1707.01340, 2017.
10. Michael Zink, Kyoungwon Suh, Yu Gu, and Jim Kurose. Characteristics of YouTube Network Traffic at a Campus Network – Measurements, Models, and Implications. *Computer networks*, 53(4):501–514, 2009.