Article

CONVERGENCE

YouTube channels, uploads and views: A statistical analysis of the past 10 years

Convergence: The International
Journal of Research into
New Media Technologies
2018, Vol. 24(1) 16–32
© The Author(s) 2017
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1354856517736979
journals.sagepub.com/home/con
SAGE

Mathias Bärtl
Hochschule für Technik, Wirtschaft und Medien Offenburg, Germany

Abstract

To this date, it is difficult to find high-level statistics on YouTube that paint a fair picture of the platform in its entirety. This study attempts to provide an overall characterization of YouTube, based on a random sample of channel and video data, by showing how video provision and consumption evolved over the course of the past 10 years. It demonstrates stark contrasts between video genres in terms of channels, uploads and views, and that a vast majority of on average 85% of all views goes to a small minority of 3% of all channels. The analytical results give evidence that older channels have a significantly higher probability to garner a large viewership, but also show that there has always been a small chance for young channels to become successful quickly, depending on whether they choose their genre wisely.

Keywords

API, success probability, statistics, time series, UGC, video usage data, video sharing, video popularity, YouTube, YouTube categories, YouTube channels

Introduction

In December 2016, the Forbes Magazine published a list of the 12 most influential YouTube stars, estimating their annual income at a total of US$70.5 million (Berg, 2016). YouTube on their own web page stated in January 2017 that ‘the number of channels earning six figures per year on YouTube is up 50% y/y’ (YouTube, n.d Statistics). Given that the prospect of revenue may be one of the motivations to share videos online (Burgess and Green, 2009), such press releases must make YouTube seem a viable career opportunity for many. The data collected as part of this study in fact suggest that the number of content-providing channels grew, on average, by 20% each year since 2006.

Corresponding author:
Mathias Bärtl, Hochschule für Technik, Wirtschaft und Medien Offenburg, Badstraße 24, Offenburg 77652, Germany.
Email: Mathias.Baertl@hs-offenburg.de

After Google purchased YouTube, the number of professional productions on YouTube significantly increased (Kim, 2012; Lobato, 2016; Vonderau, 2016). Some scholars support the notion that it is generally harder for user-generated content (referred to as UGC for brevity in the remainder of this article) to attract views (Kruitbosch and Nack, 2008). Other studies identify a strong preference for UGC (Welbourne and Grant, 2016) or at least a higher viewer interaction with UGC through commenting and voting (Liikkanen and Salovaara, 2015). However, the sheer volume of already existing competition must make it difficult for both professionals and amateurs to deliver their content to any larger audience (Cunningham et al., 2016).

A basic prerequisite for providing a quantitative assessment of a new channel's chances to attract a relevant number of views is the availability of adequate high-level statistics, but to this date, such statistics are hard to find. YouTube in its own communications is most inexplicit with numbers, and even plain questions such as 'How many channels and videos exist on YouTube?' can only be answered vaguely. For example, as of 2015, Vonderau finds estimates that range from 80 million to more than 3 billion videos (Vonderau, 2016). To contribute to a discussion that delivers a more global view, this article presents a quantitative analysis of the composition of the platform in terms of channels, uploads and views by video category (a genre-like classification used by YouTube) over time. It also addresses the impact of a channel's age and category on its success and provides estimates for young channels to become successful. The analysis is based on data from 19,025 channels, covering 5,591,400 videos uploaded between 2006 and 2016. The data were obtained through the YouTube application programming interface (API), by applying statistical sampling techniques. These, with some caveats, enable conclusions that can be seen as representative for all of YouTube. It is hoped that the findings are not only useful to individuals or companies considering whether to establish a YouTube presence but also to researchers with an interest in online video-sharing practices.

The article is organized as follows: the next section introduces related research that has also analysed YouTube data to draw conclusions about video sharing on the website. The following section specifies the aim of this article, drawing on findings from existing work. Subsequently, I discuss the technicalities of sampling representative data from YouTube and explain the approach taken to obtain the data in support of this study. Then, I present the key findings of the data analysis and conclude by offering a discussion on the analytical results and their implications.

Related work

YouTube, according to alexa.com, is the second most visited website worldwide, and its rise to one of the most relevant mass communication media of the past decade calls for significant attention from academia. YouTube also keeps detailed records of all user interaction and shares some of the so-produced data publicly, which vastly enhances quantitative research opportunities. A range of studies has picked up on this and exploited YouTube data to produce insights into social video sharing. The majority of such work has focused on the description and analysis of videos and viewer interaction, the prediction of video popularity, and external factors affecting how videos are being used and shared. This section gives a brief overview of such work and its key findings.

Various studies have explored the length, upload and access patterns, lifespan, ratings and comments, resolution and file size of YouTube videos to assess the extent to which they differ from traditional streaming content and help service providers optimize their networks (Cha et al., 2009; Che et al., 2015; Cheng et al., 2008). One study introduces a random prefix sampling approach to address the issue of unknown number of videos hosted on YouTube and estimates roughly 500 million videos by May 2011 (Zhou et al., 2011).

Analyses of viewer interaction showed that many YouTube videos experience a peak of attention, normally within a few days after their publication, although the vast majority of videos does not receive any relevant attention at all (Cha et al., 2009; Crane and Sornette, 2008). Based on a more detailed analysis of view patterns before and after the peak, Crane and Sornette suggested a classification into 'viral', 'quality', 'junk' and 'memoryless' videos. Figueiredo et al. (2011) developed a simplified approach to label videos in accordance with that classification but found that videos can have several peaks of popularity.

To further understand a video's viewing evolution more analytically, some studies developed models to predict longer term popularity by extrapolating from early view patterns. This idea seems plausible when based on the assumption that most popular videos experience their peak of attention during the first days after their release; it was found to produce reliable results by Szabo and Huberman (2010) and Pinto et al. (2013). However, Borghol et al. (2011) demonstrated that over a longer period of time views can fluctuate significantly from week to week and concluded that early lifetime views are not necessarily a good indicator of a video's future popularity. With reference to the high variance of viewer interaction, they propose to base prediction models on collections of uploads rather than individual videos.

Following up on their initial study, Borghol et al. published a second article that uses a framework of external content-agnostic factors (e.g. size of the uploader's network, number of keywords associated with the video, video age) as a starting point to analyse video popularity by comparing video clones. They find the total number of previous views to have the largest impact on future video popularity except for very young videos (where the size of the uploader's network is most important), followed by video age. Their study provides evidence for the common notion that such factors, alongside a first-mover advantage, strongly contribute to the 'rich-get-richer' phenomenon (Borghol et al., 2012). Other studies focused on the networking character of YouTube and investigated how videos are discovered and shared: Searching or clicking on YouTube suggestions was identified as the most common way to access a video (Figueiredo et al., 2011), while physical proximity of users, locality of interest (e.g. sports, news, politics) and language create barriers to geographic diffusion but, to some extent, can be overcome by social sharing (Brodersen et al., 2012). Further aspects, such as active user interaction through commenting (Liikkanen and Salovaara, 2015) or content factors (such as UGC or professional, style of delivery, gender of presenter; Welbourne and Grant, 2016) and their relation to video popularity have also been subject to quantitative research.

Most data-centric studies on YouTube aim to draw general conclusions related to individual videos, although some produce high-level statistics as a by-product (e.g. Brodersen et al., 2012; Che et al., 2015; Cheng, 2013; Cheng et al., 2008; Figueiredo et al., 2011). What is lacking is a more global quantitative description of YouTube, focused on larger aggregates, such as channels or video categories, as suggested by Borghol et al. (2011).

Aim of the study

For the purpose of understanding YouTube from a high-level perspective, this study aims to provide an overall characterization by:

- describing how channels, uploads and views on YouTube evolved over time;
- analysing the distribution of views among content-providing YouTube channels; and
- identifying factors that affect a channel's potential to garner a significant viewership.

Approach

Existing quantitative work on YouTube varies in a range of technical aspects, such as sample size and collection frequency (from less than 100 to several millions of videos, captured through just one or several snapshots in time), data collection approach (crawling, API querying, manual, or a combination of these) and in whether it looks across all of YouTube's content or focuses on specific types of videos. However, depending on how and when data are retrieved, results can vary widely and even conflict with each other. It is therefore crucial to carefully consider the specifics of YouTube data collection mechanisms in light of the desired analytical outcomes. This section describes the technicalities associated with obtaining data from YouTube and explains the approach taken for the purposes of this study. It starts with a general description in subsection 'Data collection overview' and provides a more in-depth explanation of sampling details in subsection 'Sampling details and biases'.

Data collection overview

Given that retrieving and analysing data on all videos hosted on YouTube is currently impossible (Wu et al., 2014), the key challenge is to compile a sample of data that are representative for all of YouTube but technically manageable. In principle, there are two basic methods for collecting YouTube samples: crawling or querying the YouTube API. Crawling is retrieval of data by having a data collection software visit a YouTube video page and capture predefined data items (e.g. video title, number of views, likes, dislikes, etc.). The tool then follows 'related videos' links listed on that page and carries on until some predefined completion criteria are met. Crawling allows the collection of any data that can also be viewed by a person looking at the page, but, as Zhou et al. have shown, produces samples that are highly biased towards popular videos, hence significantly underestimates the number of videos with little attention (2011). The second alternative, querying the YouTube API, allows the retrieval of data for specified videos or channels. The API also responds to keyword searches. Returns to keyword searches are ordered in some form of YouTube-defined popularity metric and, therefore, also tend to be skewed towards popular content (Borghol et al., 2011).

The way to obtain an unbiased data set is randomized sampling that is not directly supported by the YouTube API (Pinto et al., 2013). However, in principle, it seems possible to create a near-random collection of YouTube channels by using keyword searches which themselves can be seen as random samples. For this study, a tool was implemented to generate random string searches and retrieve channels with a creation date between 1 January 2006 and 31 December 2016. In a first step, the tool keeps sending search requests to the YouTube API until, after removal of duplicates, a predefined number of channels has been collected. An additional filter was applied to collect only the channels with at least five uploads (a disadvantage for very young channels but necessary to focus the analysis on active content providers). In a second step, all channel and video data associated with the list from step one are retrieved via the API as specific data requests.

The complete data collection process was repeated four times, creating four batches of approximately 5000 channels each with between 2,121,332 and 2,263,076 associated videos. The

Table 1. Number of results reported by YouTube in relation to search string length.

Search string length	Cases	Minimum	Lower quartile	Median	Upper quartile	Maximum
0	4	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000
1	10	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000
2	33	22,952	167,289	643,045	1,000,000	1,000,000
3	116	52	1253	4064	16,666	1,000,000
4	194	0	7	20	46	1,000,000
5	244	0	0	0	1	145,603

analytical results presented in the Findings section are based on all four batches combined, with duplicate channels and videos removed. The removal of duplicates left 19,025 unique channels and 5,591,400 videos in the combined sample.

Sampling details and biases

Several methods exist to create random searches, such as randomly selecting words from a dictionary or using randomly chosen time frames. To avoid bias towards a certain language or time span, the data collection tool developed for the purposes of this study creates a search string by combining a number of randomly selected letters. Each search string is forwarded to the YouTube API as a keyword search; the API responds by reporting a number of ‘results’ (a sole value, e.g. ‘About 105,000 results’) and a list of retrievable ‘returns’, limited to a maximum of 500 (YouTube, n.d., Developers); limiting the number of returns is common for any Internet search engine. If more returnable results exist, YouTube applies some form of popularity sorting and, in theory, makes a subset of 500 returns available. Consequently, a key question of the random search string approach is the ideal length of that string: Short strings tend to yield far more than 500 results and returns will be biased by YouTube’s popularity sorting algorithm. Short strings also create much overlap in search results (e.g. the channel ‘La Voz del Gol’ will be found by both the searches ‘voz’ and ‘gol’), negatively affecting data collection performance. Long strings mostly yield too few or no results at all, thereby making the data collection process time-consuming. Long strings also might give preference to channels with awkward abbreviations and discriminate against channels with common names.

Given that the ideal string length is unknown, in this study that parameter is also selected by chance (chances were arbitrarily set to: 2% for zero letters, 4% for one letter, 6% for two letters, 20% for three letters, 29% for four letters and 39% for five letters). The extent to which the search string length affected the results and returns is illustrated by the statistics provided in Tables 1 and 2.

As shown in Table 1, search strings of length 0 (four search requests) or 1 (10 search requests) always generate at least one million results (while the YouTube search page can present any value for number of results, this value is set to a maximum of 1,000,000 for searches via the API). The numbers in Table 2 also indicate that YouTube does not rigorously enforce the maximum 500 returns restriction; some searches retrieved more returns. As expected, both the number of results and returns decrease significantly with an increasing string length, although it is still possible for longer strings to generate very many results and returns (if the string happens to be a common word, e.g. ‘why’, ‘next’, ‘pause’).

The statistics in Tables 1 and 2 confirm that there is not one optimal search string length to generate a random sample of YouTube channels. However, for many searches, the actual number

Table 2. Number of actual returns in relation to search string length.

Search string length	Cases	Minimum	Lower quartile	Median	Upper quartile	Maximum
0	4	257	301	490.5	574.25	583
1	10	11	104.75	145.5	334.75	534
2	33	10	182.5	475	524	588
3	116	3	58.25	133.5	274.5	512
4	194	0	0	1	5.25	505
5	244	0	0	0	0	440

Table 3. View statistics by return classes.

Return classes	Cases	Minimum	Lower quartile	Median	Upper quartile	Maximum
[0; 100)	1489	0	230	1375	14,034	1,220,140,166
[100; 200)	1966	0	232	1645	46,693	765,693,072
[200; 300)	2206	0	176	1081	8096	972,181,419
[300; 400)	2562	0	355	2599	65,175	1,252,420,368
[400; 500)	4351	0	261	1791	14,808	2,099,923,377
[500; ∞)	6471	0	1548	20,918	675,109	2,111,879,188

of returns can be well below 500 even when the number of results is much greater. A possible explanation is that channels with 0 uploads (many channels only exist to subscribe to other channels; such channels are irrelevant to this study) appear to never be returned by the API but might still be counted as results when their name matches the search string. If this assumption is correct, it would follow that searches with less than 500 returns should be largely unaffected by the API's popularity sorting, and their popularity characteristics should be significantly different from searches with 500 returns or more. To test this assumption in more detail, returns were separated into six return classes, indicating whether they were retrieved as part of a search with just a few or many returns (e.g. class [0; 100]: all channels returned as part of search requests with between 1 and 99 returns, class [100; 200]: all channels returned as part of search requests with between 100 and 199 returns, etc.); descriptive statistics are provided in Table 3.

As can be seen from Table 3, all classes include channels with no views and very many views, but a comparison of the median (and also the lower and upper quartile) demonstrates that the class with 500 or more returns is significantly biased towards popular channels. In conclusion, a popularity bias might be reduced by excluding channels from searches with 500 or more returns. However, it depends on the analytical question of whether such an exclusion is helpful: if, for example, a minority of popular channels generates the vast majority of views on YouTube, removing them would result in a substantial underestimation of viewing traffic, hence be inappropriate for all related analyses. Therefore, in this study, the described segmentation is not used to exclude potentially biasing returns but to assess their impact on analytical findings. This is done by comparing the results for channels from searches with less than 500 returns and the results for all channels and evaluating the degree of uncertainty introduced by the popularity sorting. In addition, a cross-check for consistency is conducted across all four individual sample batches. Only results that prove to be robust against the sensitivity check and are consistent across all batches are reported in this study; caveats are stated if required.

While the data collection and cross-checking approach described above should help understand the degree of popularity bias, there is at least one systematic bias that could not be avoided by the sampling process: all search strings are generated from Latin letters and therefore discriminate to some large extent against channels with titles that are solely in Arabic, Cyrillic, Greek, and so on, script. While such channels are not completely excluded from the data set (because zero-, one- and two-letter searches did collect a sizeable number of non-Latin script channels), the extent of that bias is currently unclear.

For those 19,025 channels that are included in the search, the maximum margin of error for all proportion estimates is smaller than 1 percentage point at a 99% confidence level.

Findings

Evolution of channels, uploads and views over time

Currently, 18 different genre-like categories exist on YouTube (Autos & Vehicles, Comedy, Education, Entertainment, Film & Animation, Gaming, How To & Style, Movies, Music, News & Politics, Nonprofits & Activism, People & Blogs, Pets & Animals, Science & Technology, Shows, Sports, Trailers, Travel & Events), from which one needs to be chosen to label both channels and individual videos. In theory, channels and videos can be labelled independently of each other, for example, an Entertainment channel might upload a video and label it as Comedy. However, 75% of all channels have the same category assigned to at least 80% of their videos. Therefore, it seems fair to represent each channel by its mainly used category. This approach is chosen whenever statements about channels are made and is referred to as ‘channel category’ in the remainder of this article; the term ‘video category’ will be used to refer to category assignments made to individual videos.

From 2006 to 2009, most channels fall into the Music category, reaching a high in 2009 when more than 20% of all created channels would upload mainly Music videos. Entertainment, People & Blogs and Education were also continuously popular among creators during these years. However, from 2010 onwards, the majority of newly created channels were People & Blogs, coming close to 75% in 2016. Gaming started to be the second most channel category in 2012 and has remained in that place ever since. Across all years, Pets & Animals, Autos & Vehicles, Nonprofits & Activism and Travel & Events channels occur the least. Overall, results on channel creation remain stable when potentially biased API responses are removed (see subsection ‘Sampling details and biases’), with the exception of People & Blogs channels being around 5 percentage points higher and even more dominant over the rest. Table 4 details estimates of the proportions of channels by the year of their creation; the number of channels in the sample is included for information. For manageability, only values for the six most occurring channel categories are listed in order of their overall frequency from left to right; the remaining 12 categories are summarized into ‘Other’. Figure 1 is a visual representation of the evolution of channel categories over time, using 2011 as a baseline of 100%.

Less than 3% of all channels are News & Politics, but this category has always been extremely active and accounts for 45% of all uploads overall. From 2008 to 2015, the second most uploads were Entertainment, accounting for 12% of all uploads. People & Blogs did not grow nearly as fast in terms of uploads as it did in numbers of channels, but taking all years together, it is the third most active category, accounting for 9% of all uploads. When potentially biased API responses are removed (see subsection ‘Sampling details and biases’), the higher proportion of People & Blogs channels in the reduced data set translates into a slightly higher proportion of uploads from that category (almost exactly at the cost of Entertainment and News & Politics), making People &

Table 4. Channels: Proportions of categories by year of channel creation.

		Channel category						
Year	Channels (sample)	People & Blogs	Gaming	Music	Entertainment	Education	Sports	Other
2006	576	0.130	0.063	0.177	0.134	0.066	0.066	0.365
2007	616	0.135	0.063	0.190	0.146	0.083	0.078	0.305
2008	664	0.107	0.057	0.190	0.120	0.128	0.075	0.322
2009	965	0.152	0.082	0.205	0.110	0.089	0.062	0.299
2010	1089	0.178	0.078	0.164	0.113	0.103	0.059	0.305
2011	1853	0.276	0.098	0.145	0.096	0.073	0.043	0.270
2012	1917	0.318	0.138	0.117	0.100	0.064	0.046	0.216
2013	2380	0.411	0.147	0.106	0.080	0.048	0.043	0.166
2014	2438	0.560	0.117	0.057	0.067	0.030	0.034	0.134
2015	3147	0.639	0.111	0.045	0.050	0.023	0.023	0.109
2016	3437	0.739	0.117	0.022	0.038	0.013	0.013	0.058

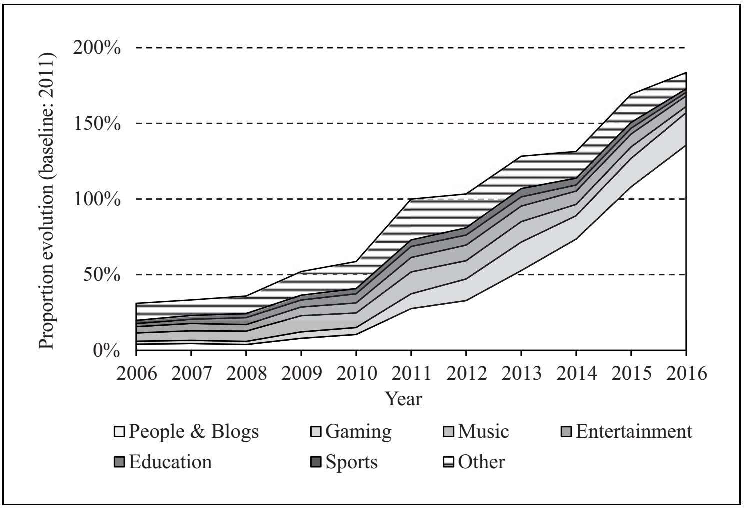

Figure 1. Channels: Proportions of categories by year of channel creation.

Blogs the second most active video provider. However, overall results on upload proportions remain stable and grew on average by 50% each year since 2009. This study's sample properties, taken together with the estimate made by Zhou et al. of 500 million videos by 2011 (Zhou et al., 2011), suggest a total of almost four billion videos on YouTube by the end of 2016.

Table 5 gives estimates of the proportions of video categories by the year of upload for the six most popular categories in order of their frequency; the remaining 12 categories are summarized into 'Other'. The number of uploads in the sample is included for information. Figure 2 is a visual representation of the development of uploads, using 2011 as a baseline of 100%.

In contrast with the supply of uploads, most popular with viewers are Entertainment videos, overall accounting for 24% of all views, followed by Music (17%) and Gaming (13%). However,

Table 5. Uploads: Proportions of categories by year of upload.

		Video category						
Year	Uploads (sample)	News & Politics	Entertainment	People & Blogs	Gaming	Education	Sports	Other
2006	1704	0.131	0.192	0.078	0.001	0.000	0.178	0.420
2007	18,813	0.434	0.106	0.108	0.000	0.020	0.133	0.200
2008	58,848	0.496	0.117	0.040	0.013	0.101	0.066	0.167
2009	116,881	0.494	0.085	0.052	0.019	0.084	0.063	0.204
2010	185,296	0.445	0.097	0.056	0.042	0.097	0.051	0.212
2011	318,063	0.476	0.116	0.038	0.058	0.074	0.05	0.189
2012	492,935	0.425	0.124	0.053	0.056	0.071	0.058	0.212
2013	611,402	0.433	0.113	0.058	0.069	0.068	0.044	0.216
2014	1,006,496	0.455	0.182	0.075	0.060	0.047	0.034	0.146
2015	1,223,308	0.506	0.110	0.104	0.068	0.038	0.037	0.138
2016	1,557,654	0.412	0.108	0.140	0.079	0.044	0.066	0.150

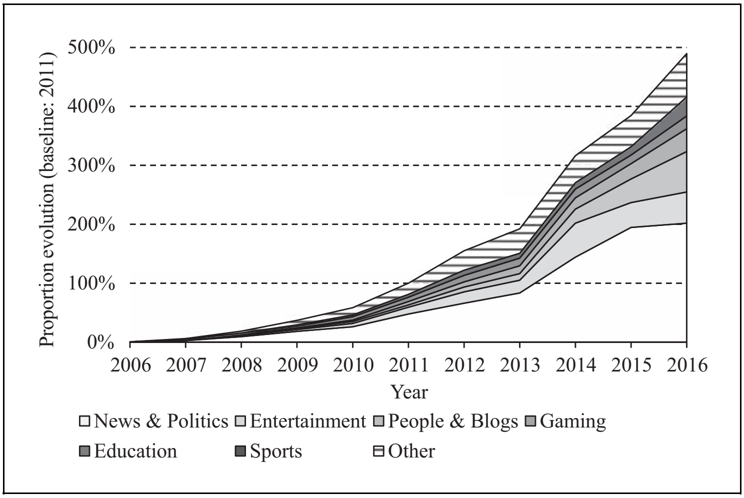

Figure 2. Uploads: Proportions of categories by year of upload.

looking at years individually, the most viewed category kept changing over time: in 2007 and 2008, News & Politics videos were watched most; in the years 2009, 2011 and 2012, it was Music; and from 2013 onwards, Entertainment became the most viewed category. Gaming started gaining popularity in 2010 and gradually grew year by year to be the second most viewed category in 2015 and 2016. The most popular category in 2010, and in fifth place across all years, is Shows, although this result is based on only 10 highly viewed channels, hence may not be representative for all of YouTube. A removal of potentially biased API responses (see subsection 'Sampling details and biases') changes view proportion estimates only insignificantly. However, for categories with similar view proportions, the order of relevance might change (for example: when potentially

Table 6. Views: Proportions of categories by year of upload

		Video category						
Year	Views (sample)	Entertainment	Music	Gaming	Comedy	Shows	People & Blogs	Other
2006	2.87E8	0.300	0.072	0.000	0.172	0.049	0.030	0.376
2007	1.27E9	0.150	0.116	0.001	0.169	0.076	0.056	0.433
2008	4.15E9	0.125	0.119	0.022	0.172	0.125	0.033	0.403
2009	1.24E10	0.069	0.494	0.033	0.084	0.101	0.013	0.205
2010	1.50E10	0.114	0.126	0.059	0.085	0.412	0.016	0.188
2011	1.73E10	0.195	0.227	0.089	0.098	0.134	0.025	0.234
2012	2.95E10	0.213	0.240	0.093	0.106	0.087	0.036	0.225
2013	4.93E10	0.241	0.178	0.134	0.089	0.106	0.028	0.225
2014	7.50E10	0.266	0.196	0.138	0.101	0.056	0.039	0.203
2015	8.56E10	0.255	0.149	0.160	0.103	0.042	0.060	0.232
2016	8.68E10	0.288	0.109	0.152	0.074	0.040	0.105	0.232

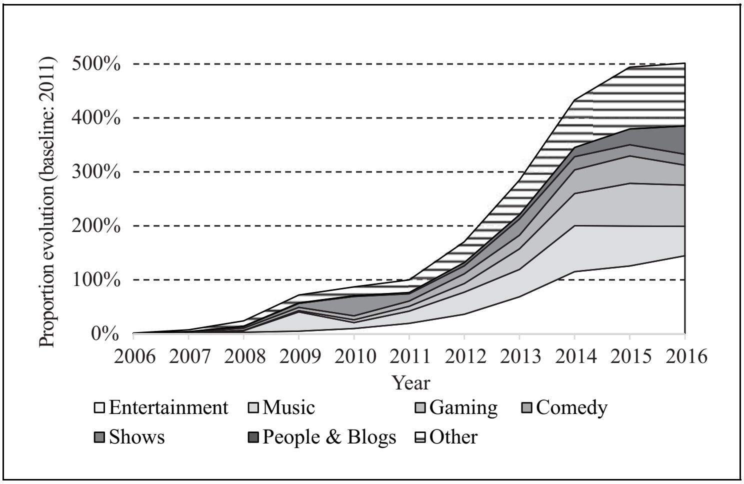

Figure 3. Views: Proportions of categories by year of upload.

biased returns are removed, Entertainment remains the most viewed category, but Music and Gaming swap places to two and three). Table 6 gives estimates of the proportions of views by the video category and year, again displaying values for the six most viewed categories and summarizing the rest into ‘Other’; the number of views in the sample is included for information. Figure 3 is a visual representation of the evolution of views, using 2011 as a baseline of 100%.

Taking all categories together, not much difference can be seen between the total number of views for uploads from 2015 and from 2016, although the number of uploads increased by approximately 27% between the 2 years. An explanation for this is that older videos generally tend to have more views, an observation already documented by Borghol et al. (2011). At first glance, this seems to be in conflict with the phenomenon of many videos experiencing their peak of attention a few days after their release (Cha et al., 2009; Crane and Sornette, 2008). However, it is

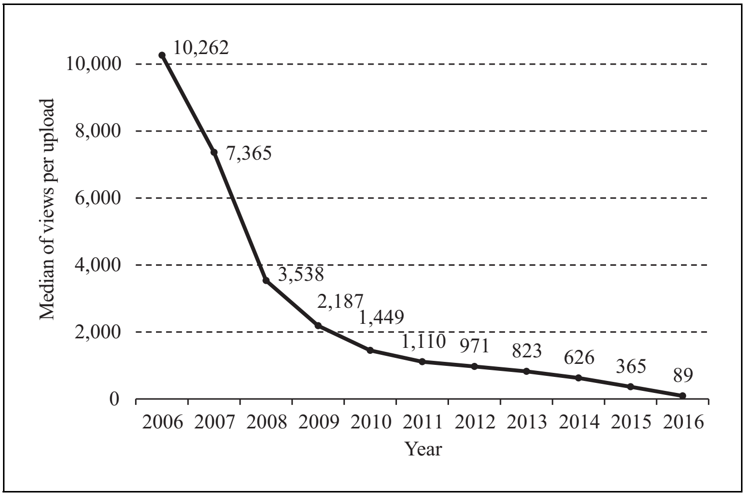

Figure 4. Median of views per upload by year of upload.

not: A long period of low viewing rates can easily outweigh the number of views received during the peak. To further illustrate the relation between a video's age and its total number of views, Figure 4 shows the median number of views per upload by year, starting with 10,262 views for videos from 2006, plummeting quickly and going as low as 89 views for videos from 2016 (meaning that 50% of all videos uploaded in 2016 had a total of 89 views or less at the time when the data were collected).

Distribution of views among content-providing channels

Figure 4 indicates how little attention most videos receive and can therefore serve as an introduction to an analysis of how views are distributed between channels. To explore this question further, for each year separately, all channels were separated into the 'top 3%' most viewed channels and the rest referred to as 'bottom 97%.' The proportion of uploads were calculated for both the top 3% and bottom 97% segments by year; the same was done for views. It shows that, very consistently, the top 3% most viewed channels upload at least 20% of all videos and receive the vast majority of all views, up to nearly 90% in 2016. Taking all years together, the top 3% most viewed channels account for 28% of all uploads and 85% of all views. Table 7 lists detailed proportions by year; Figure 5 illustrates the relationship between channels, uploads and views using the top 3% to bottom 97% division from 2006 to 2016, again using 2011 as a baseline. However, the annual estimates of the top 3% most viewed channels' dominance over the rest are sensitive to the removal of potentially biased API returns (see subsection 'Sampling details and biases'): Queries with 500 or more returns overestimate the dominance of popular channels in earlier years and underestimate it from 2010 onwards for uploads; the bias is similar but somewhat smaller for views. Consequently, Table 7 and Figure 5 should mainly be seen as a rough indication of the irrelevance of the vast majority of YouTube channels; actual conditions are likely to be more extreme (i.e. the top 3% most viewed channels might account for up to almost 50% of all uploads and 95% of all views).

Table 7. Upload and view share between the 3% most and the 97% least viewed channels.

	Uploads		Views	
Year	Top 3%	Bottom 97%	Top 3%	Bottom 97%
2006	0.272	0.728	0.638	0.362
2007	0.455	0.545	0.649	0.351
2008	0.545	0.455	0.785	0.215
2009	0.384	0.616	0.873	0.127
2010	0.332	0.668	0.845	0.155
2011	0.293	0.707	0.799	0.201
2012	0.270	0.730	0.816	0.184
2013	0.281	0.719	0.835	0.165
2014	0.286	0.714	0.843	0.157
2015	0.212	0.788	0.856	0.144
2016	0.286	0.714	0.898	0.102

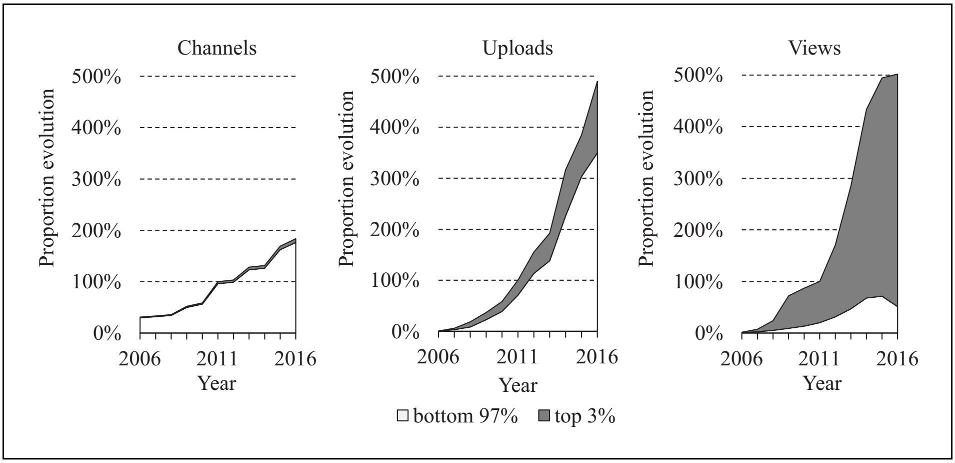

Figure 5. 3% of all channels uploaded 28% of all content and received 85% of all views.

Figures 4 and 5 can be seen as a macro-manifestation of the 'rich-get-richer' phenomenon, which has been extensively researched already at the individual video level, pointing at 'previous views' being highly relevant to predict 'future views' (Borghol et al., 2012; Cha et al., 2009; Crane and Sornette, 2008; Szabo and Huberman, 2010). Taken to the aggregate level of channel popularity, this raises questions as to whether it is possible at all for younger channels to attract any relevant viewership.

Factors impacting a channel's potential to garner significant viewership

Recapping the initial observations of (a) stark contrasts between categories in terms of existing channels, uploads and views; (b) younger videos having a lower number of views; and (c) the vast majority of all views going to a small minority of all existing channels, the final part of this section will inspect the impact of age and category on a channel's probability to move to the top 3% most

Table 8. Success probabilities by category and year of activity.

	Year					
	2011	2012	2013	2014	2015	2016
Cases	3131	4239	5557	6708	8730	12,751
df	11	11	11	11	11	11
χ^2	50.27	59.7	97.17	128.51	215.71	445.45
p	5.60E-07	1.06E-08	6.49E-16	3.48E-22	4.03E-40	1.34E-88
Channel category	Success probability					
Comedy	0.095	0.066	0.065	0.066	0.077	0.085
Entertainment	0.054	0.054	0.059	0.069	0.084	0.093
Howto & Style	0.059	0.070	0.068	0.050	0.045	0.078
News & Politics	0.032	0.039	0.038	0.050	0.054	0.109
Gaming	0.037	0.048	0.055	0.058	0.055	0.056
Film & Animation	0.030	0.028	0.021	0.039	0.035	0.045
Music	0.034	0.034	0.036	0.028	0.033	0.037
Science & Technology	0.019	0.026	0.032	0.030	0.031	0.045
Sports	0.013	0.010	0.011	0.016	0.019	0.036
Education	0.003	0.004	0.003	0.007	0.008	0.018
Nonprofits & Activism	0.007	0.006	0.006	0.006	0.010	0.010
People & Blogs	0.012	0.009	0.007	0.007	0.006	0.004

viewed segment. Using 3% as a cut-off point is somewhat arbitrary and was done to highlight the heavy concentration of views on very few channels. However, for clarity, this threshold will be maintained for the following analysis. The flip side of this is that there are an insufficient number of cases for most categories in the years from 2006 to 2010, and for Autos & Vehicles, Pets & Animals, Shows and Travel & Events for most years, to allow for results to be satisfactorily reliable from a statistical viewpoint. Data from 2006 to 2010 and the four aforementioned categories are therefore excluded from all further analysis.

To assess the impact of a channel's category on its success, for each category and year, the number of channels within the top 3% most viewed segment is divided by the number of all channels of that same category (e.g. number of Comedy channels of 2011 in top 3% / number of all Comedy channels of 2011). This proportion is referred to as 'success probability' in the remainder of this article (if chances were even, all categories would always have a probability of 0.03 to become part of the top 3%, but some categories will have better chances than others). The same procedure is repeated for channel age.

The obvious approach to test and indicate the importance of channel age and category would be a logistic regression model. However, a parallel split into several years and categories does not leave a sufficient number of cases in the top 3% segment for the regression to deliver statistically significant results. Instead, a χ^2 test for independence is employed individually for channel category and age.

Starting with categories, Table 8 lists the test statistics and the probabilities by category and year. 'Cases' refer to the number of active channels in the sample by year; the categories are sorted in order of their likelihood across all considered years.

The odds of Comedy, Entertainment, How To & Style and Gaming to make it to the top 3% have always been better than average. In 2016, however, News & Politics stands out with a 10.9%

Table 9. Success probabilities by year of channel creation and year of activity.

	Year					
	2011	2012	2013	2014	2015	2016
Cases	3131	4239	5557	6708	8730	12,751
df	5	6	7	8	9	10
χ^2	72.58	82.47	109.83	113.82	160.19	314.9
p	2.97E-14	1.10E-15	9.96E-21	6.23E-21	6.79E-30	1.10E-61
Year of channel creation	Success probability					
2006	0.097	0.103	0.102	0.089	0.093	0.120
2007	0.027	0.035	0.034	0.056	0.059	0.071
2008	0.042	0.044	0.047	0.051	0.052	0.068
2009	0.030	0.026	0.030	0.050	0.058	0.077
2010	0.017	0.031	0.054	0.055	0.060	0.080
2011	0.008	0.018	0.028	0.028	0.035	0.046
2012		0.011	0.017	0.034	0.044	0.049
2013			0.006	0.014	0.026	0.031
2014				0.004	0.014	0.020
2015					0.003	0.015
2016						0.005

chance. The chances of Sports, Education, Nonprofits & Activism and People & Blogs are consistently worse than average. Success probabilities appear to be generally insensitive to potentially biased API responses (see subsection ‘Sampling details and biases’); by and large, Table 8 can be regarded as giving a fair answer to the question of whether a certain category improves or deteriorates the chances to become successful. Three exceptions to this are Gaming, which goes down to nearly 3% in the reduced data set, and How To & Style and News & Politics, which seem to have even better chances to become successful.

The descriptive statistics presented on video age and views point to a strong advantage for older videos. To explore the relationship between channel age and views, the approach used to test the impact of category on success probability is repeated for channels, using the year of channel creation as independent variable; the results are presented in Table 9.

As expected, older channels, with few minor exceptions in some years, do have a higher probability to belong to the top 3% most viewed channels; the results are statistically significant in all years. The data should not be misread, though, as most successful channels are old. Because there are many more younger channels than older ones, many channels within the top 3% are from relatively recent years. For example, in 2016, 33% of the successful channels were from 2013 or younger. However, given the competition from an exponentially increasing number of channels, it becomes more and more difficult to be successful against the peer age, hence the success probability wanes quickly. Also of note, there has always been a small chance for young channels to become successful quickly: Throughout all years, there were some channels making it to the top 3% within their first year of existence.

A removal of potentially biased API responses (see subsection ‘Sampling details and biases’) results in channels from 2007 appearing to have the greatest advantage; otherwise, results remain relatively stable.

In summary, the considerable concentration of a majority of views on a minority of videos identified several years ago by Cha et al. (2009) and, to some extent, also Crane and Sornette (2008), still holds true and can also be observed at the level of YouTube channels. Both channel category and channel age are statistically significant predictors of a channel's annual size of viewership, with an advantage for older channels and for channels that mainly upload Comedy, Entertainment and How To & Style videos.

Conclusion and discussion

One of the key observations presented in this article is the overwhelming dominance of very few channels over the rest of content on YouTube. The findings of this study provide strong indications that this macro-manifestation of the rich-get-richer phenomenon can be linked to two of probably three factors, namely (a) general processes of growth and sharing of information (Crane and Sornette, 2008) and (b) a mismatch between supply and demand of content. Firstly, it is normal for videos or channels that have already been viewed by many to get more new views, simply because they have a greater sharing base. As channels and videos collect more views in the course of their lifetime, there is always a possibility to gather a critical mass of attention, but for most, it will take long and a lot of patience. Secondly, an overwhelming growth of, for example, People & Blogs channels against a moderate consumer interest creates an increasing mismatch between demand and supply and will naturally leave many channels with very few views. It is therefore not surprising that the channel category is a highly significant predictor of channel success. The data show that there has always been some space for newcomers to become successful quickly in categories with an expedient demand–supply ratio. A preliminary review of the channels from 2016 that made it to the top 3% within their first year of existence indicates a mix of UGC and professionally produced videos, but this needs more analysis. However, it does give hope that YouTube's 'broadcast yourself' rhetoric is not a complete fiction. A third factor (c), which probably accelerates the convergence of views towards a few content providers, is YouTube's search and recommendation algorithms and video promotion sales (Kim, 2012; Zhou et al., 2011). That third factor is in the hands of YouTube and could be adjusted to help steer the platform at least somewhat towards a social networking ideal, improve the chances of young content providers to build a worthwhile viewership and thereby help keep the production of UGC interesting.

A second takeaway from this study and its comparison with the existing work is that analytical findings from social media data can differ dramatically, depending on the data collection method, the time frame covered and the analytical approach. High variety and high velocity are well-known properties of big data, and small variations in analytical parameters can lead to very different results. However, the data are in large part a reflection of human behaviour, therefore high variety and high velocity may be properties of society, too, and should therefore be studied with care. In that context, finding the right level of aggregation instead of fine-tuning, as suggested by Borghol et al. (2012), might be a good approach to exploit social media data to help draw large-scale maps of where society is headed. However, this needs subject matter expertise, corroborating observations and narrative, as looking at numbers of uploads, views, likes or view durations can hardly tell the full story. Nevertheless, it is hoped that media researchers find the high-level view of the platform provided by this study a useful input in the effort to understand the fascinating medium that is YouTube.

Supplementary material

Supplementary material for this article is available online at https://elearning.hs-offenburg.de/moodle/course/view.php?id=3776 Password: YouTube-Conv2018.

References

Berg M (2016) The highest-paid YouTube stars 2016: pewdiepie remains no. 1 with $15 million. In: Forbes, 5 December 2016. Available at: http://www.forbes.com/sites/maddieberg/2016/12/05/the-highest-paid-youtube-stars-2016-pewdiepie-remains-no-1-with-15-million/#6b45b38f6b0f (accessed 23 August 2017).

Borghol Y, Ardon S, Carlsson N, et al. (2012) The Untold Story of the Clones: Content-Agnostic Factors that Impact Youtube Video Popularity. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, 12–16 August 2012, pp 1186–1194. New York: ACM.

Borghol Y, Mitra S, Ardon S, et al. (2011) Characterizing and modelling popularity of user-generated videos. *Performance Evaluation* 68(11): 1037–1055.

Brodersen A, Scellato S, and Wattenhofer M (2012) YouTube Around the World: Geographic Popularity of Videos. In: *Proceedings of the 21st International Conference on World Wide Web*, Lyon, France, 16–20 April 2012, pp. 241–250. New York: ACM.

Burgess J and Green J (2009) The Entrepreneurial Vlogger: Participatory Culture Beyond the Professional-Amateur Divide. In: Snickars P and Vonderau P (eds) *The YouTube Reader*. Stockholm: National Library of Sweden, pp. 89–107.

Cha M, Kwak H, Rodriguez P, et al. (2009) Analyzing the video popularity characteristics of large-scale user generated content systems. *Transaction on Networking* 17(5): 1357–1370. New York: ACM

Che X, Ip B, and Lin L (2015) A Survey of Current YouTube Video Characteristics. *IEEE Multimedia* 22(2): 56–63.

Cheng X (2013) Understanding the characteristics of internet short video sharing: a youtube-based measurement study. *Transactions on Multimedia* 15(3): 1184–1194.

Cheng X, Dale C, and Liu J (2008) Statistics and Social Network of YouTube Videos. In: *Proceedings of the 16th International Workshop on Quality of Service*, Enschede, Netherlands, 2–4 June 2008, pp. 229–238. New York: IEEE.

Crane R and Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105(41): 15649–15653.

Cunningham S, Craig D, and Silver J (2016) YouTube, multichannel networks and the accelerated evolution of the new screen ecology. *Convergence* 22(4): 376–391.

Figueiredo F, Benevenuto F, and Almeida JM (2011) The Tube over Time: Characterizing Popularity Growth of YouTube Videos. In: *Proceedings of the fourth ACM international conference on Web search and data mining*, Hong Kong, China, 9–12 February 2011, pp.745–754. New York: ACM.

Kim J (2012) The institutionalization of YouTube: From user-generated content to professionally generated content. *Media, Culture & Society* 34(1): 53–67.

Kruitbosch G and Nack F (2008) Broadcast yourself on YouTube: Really? In: *Proceedings of the 3 rd ACM International Workshop on Human-Centered Computing*, Vancouver, British Columbia, Canada, 31–31 October 2008, pp. 7–10. New York: ACM.

Liikkanen LA and Salovaara A (2015) Music on YouTube: User engagement with traditional, user-appropriated and derivative videos. *Computers in Human Behavior* 50: 108–124.

Lobato R (2016) The cultural logic of digital intermediaries: YouTube multichannel networks. *Convergence* 22(4): 348–360.

Pinto H, Almeida JM, and Gonçalves MA (2013) Using Early View Patterns to Predict the Popularity of YouTube Videos. In: *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, Rome, Italy, 4–8 February 2013, pp. 365–374. New York: ACM.

Szabo G and Huberman BA (2010) Predicting the Popularity of Online Content. *Communications of the ACM* 53(8): 80–88.

Vonderau P (2016) The video bubble: Multichannel networks and the transformation of YouTube. *Convergence* 22(4): 361–375.

Welbourne DJ and Grant WJ (2016) Science communication on YouTube: Factors that affect channel and video popularity. *Public Understanding of Science* 25(6): 706–718.

Wu X, Zhu X, Wu GQ, et al. (2014) Data Mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering* 26(1): 97–107.

YouTube (n.d) Statistics. Available at: http://youtube.com/yt/press/statistics.html (accessed 3 September 2017).

YouTube (n.d.) Developers. Available at: https://developers.google.com/youtube/v3/docs/search/list (accessed 3 September 2017).

Zhou J, Li Y, Adhikari VK, et al. (2011) Counting YouTube Videos via Random Prefix Sampling. In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, Berlin, Germany, 2–4 November 2011, pp. 371–380. New York: ACM.

Author biography

Mathias Bärtl is a professor of mathematics and statistics at Offenburg University of Applied Sciences (Offenburg/Germany). His research examines the potential of social media data in general and the role of educational videos on YouTube. He is the author of a bestselling textbook that integrates the benefits of video-based teaching from his Statistics Tutorials Channel on YouTube.