Journal of Quantitative Description: Digital Media 3(2023), 1–85 10.51685/jqd.2023.022

Dialing for Videos: A Random Sample of YouTube

RYAN MCGRADY
KEVIN ZHENG
REBECCA CURRAN
University of Massachusetts Amherst
JASON BAUMGARTNER
PushShift.io
ETHAN ZUCKERMAN
University of Massachusetts Amherst

YouTube is one of the largest, most important communication platforms in the world, but while there is a great deal of research about the site, many of its fundamental characteristics remain unknown. To better understand YouTube as a whole, we created a random sample of videos using a new method. Through a description of the sample's metadata, we provide answers to many essential questions about, for example, the distribution of views, comments, likes, subscribers, and categories. Our method also allows us to estimate the total number of publicly visible videos on YouTube and its growth over time. To learn more about video content, we hand-coded a subsample to answer questions like how many are primarily music, video games, or still images. Finally, we processed the videos' audio using language detection software to determine the distribution of spoken languages. In providing basic information about YouTube as a whole, we not only learn more about an influential platform, but also provide baseline context against which samples in more focused studies can be compared.

Keywords: YouTube, random sampling, methodologies, social media, digital infrastructure

Introduction

YouTube is the second most popular website in the world as of November 2022, receiving 33 billion visits that month. The users who visited spent significantly more time there than visitors to any of the other sites in the top 20 (SimilarWeb, 2022). Founded in 2005 and owned by Google since 2006, the video sharing website has grown to become the default video hosting site for the global internet. It is a platform for self-expression, a mechanism to transmit recordings of events to participants, an alternative news outlet, film archive, music streaming service, and ubiquitous driver of popular culture. It has given rise to new forms of entertainment (Burgess & Green, 2018), new modes of communication (Tolson, 2010), new economies (Ørmen & Gregersen, 2022), new forms of marketing (Mowlabocus, 2018), new educational tools (Duffy, 2008), and new avenues for influence and propaganda (Barrett & Hendrix, 2022). It has transformed older media industries in profound ways, from music (Cayari, 2011) to news and television (Nashmi, et al., 2017), and through its algorithms plays an active role in shaping knowledge, attitudes, and opinions (Rieder, et al., 2018; Bryant, 2020; Chen, et al., 2021).

The importance of YouTube as an object of study has been well established (Arthurs, et al., 2018; Burgess & Green, 2018; Snickars & Vonderau, 2010), but while there is a body of research about the site, there is still much we do not know, for two primary reasons. The first reason is the difficulty of studying audiovisual content. For years, projects like Brandwatch,[1] CrowdTangle,[2] and Media Cloud[3] have provided researchers with tools to analyze and process large amounts of text scraped from social networks and news media, but there is no clear equivalent for YouTube. Unlike sites like Facebook and Twitter, YouTube's primary medium is video, which is harder for computers

[1] https://brandwatch.com/

[2] https://crowdtangle.com/

[3] https://mediacloud.org, https://ojs.aaai.org/index.php/ICWSM/article/view/18127/17930

to make sense of than text. Humans can analyze small groups of videos to evaluate their content, but machines can process text in much larger quantities. Research that examines the content of online speech - research on hate speech and cyberbullying, for example - often relies on computational methods designed to analyze text content like Twitter rather than video on YouTube. The YouTube API enables researchers to extract video metadata, but metadata by its nature is information about a video, like a video's length or the number of views it has received, rather than the contents of that video.

The second major reason YouTube is difficult to study is not unique to YouTube but typical of all the large platforms run by for-profit companies: the opacity of their code and practices. While it is true that nearly anyone can upload a video to YouTube and then share the link with friends or family, traffic on YouTube is largely driven by internal recommendation algorithms (Salsman, 2018) which organize and prioritize videos according to a wide range of factors, some of which are personalized based on the information it has about individual viewers. This may benefit media consumers pleased to discover content that matches their tastes, but it presents a challenge to researchers.

As described by Kate Starbird, YouTube is "almost inscrutable" to researchers, compared to platforms like Twitter (Martin, 2021). YouTube has not allowed outside researchers to audit its algorithms, so we cannot be certain about why some videos are promoted and others exist in obscurity. The combined opacity and importance of these algorithms in the larger information ecosystem have led to a large body of work attempting to study them. For example, in the late 2010s there was a great deal of interest among journalists and academics in the way YouTube seemed to be guiding people to increasingly radical or harmful content (Tufekci, 2018; Fisher & Bernnhold, 2018; Lewis, 2018a; Lewis, 2018b; Rieder, at al., 2018; Munn, 2019). These researchers had to employ a wide range of methodologies involving identifying particular groups of political actors, identifying harmful videos, finding links between videos, collecting large numbers of recommendations, analyzing search results, or comparing comment data.

Then, in 2019, YouTube announced it changed its algorithms to avoid moving people to more extreme content. It was a rare case of the company disclosing such a change, perhaps in response to the negative publicity. In other cases, researchers would not know why subsequent studies had different findings. Indeed, although YouTube did not disclose exactly what kinds of changes it made, research of its algorithms after they were implemented did show that it was in fact usually moving people away from the more extreme content and towards videos from more mainstream or general interest channels (Ledwich & Zaitsev, 2020; Chen, at al., 2021; Ribeiro, et al., 2021).

Further complicating matters, while YouTube has a useful API, it does not provide any way to obtain a random sample of its videos, so most studies of YouTube either begin with a list of known channels or videos or create opportunistic samples by collecting videos recommended by YouTube's algorithms. We believe - and will demonstrate in forthcoming work - that these methods create samples that are significantly different from a pure random sample of videos. Specifically, we see that videos collected via recommendation are more likely to be popular than purely random videos, both because YouTube promotes them and because a human or algorithm determined they were worth recommending. While there is vitally important work which operates within these limitations on sampling, we are left with fundamental questions about the entirety of the YouTube corpus.

Among these fundamental unknowns are the number of videos YouTube hosts, how many views those videos typically receive, how much the site as a whole has grown over time, and how frequently various languages are represented in videos. Missing data like this creates "denominator problems" and "distribution problems" which limit the kinds of claims we can make (Zuckerman, 2021). A study may identify a set of 500 YouTube channels which contained misinformation about the 2020 United States presidential election, but it would be limited in the claims it can make about the prevalence of this misinformation. In other words, 500 out of how many videos? To understand how likely someone is to find this harmful information, how popular it is relative to other videos, how much other content exists, or how many videos are in a given language, we need to know the overall size of collections (the denominator) and the distribution within those collections. Studies about harmful content on YouTube can shed light on how much of that content there is and, sometimes, how popular it is, but we need to know the denominators and distributions in order to better understand the role that content plays on the broader YouTube environment.

This paper is, in large part, an attempt to solve denominator and distribution problems. YouTube is an important, immensely popular, and critically understudied platform in part due to the difficulties it poses to researchers. By describing our methods and findings, we hope to contribute to a better understanding of YouTube as a whole and provide a basis for comparison when researchers want to contextualize samples generated for focused projects.

The first part of this paper reviews some of the ways researchers have tried to produce random samples of YouTube videos in the past using a defunct recent uploads feed, random keyword searches, and a bug in the search engine. We then present our method, Dialing for Videos, a modified brute force technique which takes advantage of the search engine's case insensitivity. The methodology section also explains how we extracted metadata for our sample of about 10,000 videos from YouTube's API, downloaded the audio tracks, ran them through a language detection model, and hand-coded a subset of 1,000 videos to better understand the content which would otherwise be missing when only looking at metadata.

Most of the paper is a presentation of our results. We start with the overall size of YouTube, which we are able to estimate with high confidence because of the sampling method we used. Upload dates from metadata further allow us to estimate YouTube's growth over time, showing for example a marked increase in yearly uploads starting in 2020. The next part of the paper describes the characteristics of the sample, combining both metadata analyses and results of hand-coding. We present four dimensions of video popularity first: views, likes, comments, and subscriptions. Subscriptions are a metric associated with channels, not videos, but we were able to extract channel data based on our video sample, so we included it as another commonly used measure for which there is not good data applicable to YouTube as a whole. Given the widely assumed relationships between these four metrics, we calculate linear correlations. The next characteristics are video duration and live-streaming, which were affected by changing site rules about maximum length and the availability of live-streaming. Categories and tags, ways that YouTube uploaders organize their content and make it discoverable, are frequently studied as they provide a quick sketch of the prominence of different kinds of content on the platform. Unfortunately, most uploaders just use the default category and most videos do not include tags.

Through hand-coding, we are able to provide some basic descriptions of audio and video quality and content, such as how many videos may just be still images, how many are just video games or have only music. We were especially interested in spoken language, which is so important to transcription and language detection. We coded for advertisements and other forms of monetization, too, but found advertisements too inconsistent to measure properly and scant evidence of monetization in our sample. Finally, we give the results of our language analysis, including a breakdown of most spoken languages in our sample. Though there are limitations in the technology we used, we believe this is a crucial finding in the way it reveals YouTube as a more diverse place than any one user will realistically be exposed to. The last part of this paper compares our sample, for which we have high faith in its randomness, to what we think is the most promising other method to create a random sample.

Methods for creating and analyzing a random sample of YouTube videos

There are many ways to sample YouTube. Researchers have used surveys and browser tracking to identify sets of videos or channels (Chen, et al., 2021), queried YouTube's search function to find videos matching certain keywords or tags (Rieder, et al., 2018; Ribeiro, et al., 2021; De Saa & Ranathunga, 2020), followed YouTube's algorithmic suggestions to collect videos recommended to viewers of one or more pre-selected videos (Cheng, et al., 2008; Ledwich & Zaitsev, 2020; Hussein, et al., 2020), and crowdsourced examples of certain types of content (McCrosky & Geurkink, 2021).

The sampling strategy a study uses constrains the kinds of claims that can be made about a broader population. A study of videos about COVID-19 from channels known for promoting fringe theories may yield interesting findings about what kinds of misinformation exist within a particular part of YouTube and how those channels relate to each other, but cannot tell us about the prevalence of vaccine misinformation on YouTube as a whole. Short of gaining access to YouTube's full database, random sampling is the only way to confidently make claims about the site as a whole. Random sampling helps to ensure the subset being studied is representative of the larger population, reducing the potential for bias. Unfortunately, YouTube does not provide a mechanism to create such a sample.

A common way to obtain a random sample in early YouTube research was to use a feature that is no longer available: the recent uploads feed. As of 2009, the list was limited to the most recent 1,000 videos, which was a small fraction of the total uploaded each day (Hráček, 2009). Wesch (2008) relied on student labor to systematically retrieve videos from the list multiple times per day while Hráček (2009) automated the process. The automated process ensured that no videos were missed, querying the API every two hours throughout the month of April 2009 and producing a list of 161,643 videos. There were two problems with reliance on the recent videos feed, however. The first is that the feed did not appear to be unfiltered. Hráček found “periodically recurring blind spots” and times when the feed seemed to stop being populated, and also noted that the size of his set did not align with much larger figures published about the number of uploads YouTube receives (2009). The second problem with the approach is that, by its nature, it is bound to a certain time frame. The earliest dated video would be the date upon which the study began. To conduct the kind of study in this paper using the recent uploads feed, even if it were still available, we would have had to be collecting videos consistently since YouTube's launch.

A creative alternative strategy was used by Bartl (2018), who moved the site of randomness from the individual videos to keywords. He used a program to generate random combinations of characters, varying in length, and input the string into YouTube's search engine, extracting the channels. His aim was to consider trends across the first ten years of the site's existence rather than to characterize the site as a whole, so he restricted the sample to channels created during that time frame, removed channels with fewer than five total uploads, and was only minimally concerned with the role of the site's algorithms in ordering results when they exceeded the number of "returns" (videos actually provided to the user). The set used in his study included 20,000 channels with about 9 million videos.

The most promising random sampling technique up to now is the "Random Prefix Sampling" approach by Zhou, et al. (2011). In a paper which sets out to estimate the size of YouTube, the authors take advantage of an unusual feature (or, more likely, a bug) in the YouTube search software. Every YouTube video is assigned a unique 11-character identifier which is visible in the url after "youtube.com/watch?v=". Zhou, et al. found that when you search for the string "watch?v=xy…z" where "xy…z" is the first part of a valid YouTube ID (prefix) that does not contain a dash ("-"), the search engine will return results beginning with that prefix followed by a dash. This may have to do with the way search indexing algorithms typically treat a dash as a space. Since YouTube appears to index videos by URL, the part of the url before the dash and the part of the url after the dash are both valid search terms. The method Zhou, et al. used involved using a range of prefix lengths, observing some noise with shorter prefixes. Twelve years later, it is still possible to produce a sample of YouTube videos with dashes in their IDs using this method. There is some challenge in determining the length of the prefix, and some trust required that the selection of prefixes, the existence of a dash, and inner workings of YouTube's search algorithms are not creating non-random results, but we suspect this continues to be the most practical way for researchers to produce random samples.

A potential weakness in this method is the overstudying of the part of YouTube that has a dash in the ID because of this method's accessibility. In particular, if this method is widely used, we would expect researcher views to skew view counts for videos with a dash in their descriptor. (As detailed later, random videos often have very low view counts, to the point where researcher skew could significantly alter results.) Part of our aim is to make small improvements to the degree of randomness of this technique by eschewing the use of prefixes, expanding the range of possible videos to include more than those with dashes in their IDs, and avoiding the reliance on a bug in an already opaque system. Our method also more easily lends to estimating YouTube's overall size and growth. After describing our sample, we show how its fundamental characteristics compare with the Random Prefix Sampling method.

Dialing for Videos

Every video on YouTube is assigned a case-sensitive 11-character identifier which is visible in the URL as youtube.com/watch?v=[identifier]. Unlike video titles, descriptions, or tags, these identifiers are guaranteed to be unique for every video. Each of the first ten characters in an identifier can be an uppercase (A-Z) or lowercase (a-z) letter of the English language, a number (0-9), dash ("-"), or underscore ("_"). The eleventh character can only be one of these sixteen characters: A, E, I, M, Q, U, Y, c, g, k, o, s, w, 0, 4, or 8.

The most straightforward, if hypothetical, method to randomly sample YouTube would be brute force: keep guessing random YouTube IDs until you get enough hits for a sample. However, the identifier configuration allows for 18.4 quintillion possibilities (2^{64}, or 64^{10} * 16), making it computationally infeasible to search for each potential video identifier. Among other concerns, in order to produce a random sample within a reasonable time frame using brute force, we would have to query YouTube's API at a rate which may draw unwanted attention from the company or negatively affect its normal operations. By constructing a search query that joins together 32 randomly generated identifiers using the OR operator, the efficiency of each search increases by a factor of 32.

To further increase search efficiency, randomly generated identifiers can take advantage of case insensitivity in YouTube's search engine. A search for either "DQW4W9WGXCQ" or "dqw4w9wgxcq" will return an extant video with the ID "dQw4w9WgXcQ". In effect, YouTube will search for every upper- and lowercase permutation of the search query, returning all matches. Each alphabetical character in positions 1 to 10 increases search efficiency by a factor of 2. Video identifiers with only alphabetical characters in positions 1 to 10 (valid characters for position 11 do not benefit from case-insensitivity) will maximize search efficiency, increasing search efficiency by a factor of 1024. By constructing search queries with 32 randomly generated alphabetical identifiers, each search can effectively search 32,768 valid video identifiers.

We ran a Python script between October 5, 2022, and December 13, 2022, which recorded every randomly generated video identifier that was searched and every video that was randomly guessed. Searches were performed using the InnerTube API,[4] which is an internal API that enables searches without rate-limits. Upon a successful guess (or "hit"), the script would download the metadata, any available transcripts, and the video's audio track using the yt-dlp Python package.[5] We stopped the script when we collected 10,016 videos.

There is at least one limitation to this method: we have only collected public videos. YouTube videos can be assigned one of three privacy levels: public, unlisted, and private. Public videos are visible to anyone, appear on channel pages, show up in search results, can be recommended or linked as related, appear in subscribers' feeds, can be commented on, and have shareable links. Unlisted videos are similar, but do not show up in search results, are not linked through recommendations or related videos, and do not appear in

[4] https://github.com/tombulled/innertube
[5] https://github.com/yt-dlp/yt-dlp

subscribers’ feeds. Typically, an unlisted video is one that you must be given the url to view. Private videos are only visible to the uploader and individuals to whom the uploader grants access directly.

As our method uses YouTube search, our random set only includes public videos. While an alternative brute force method, involving entering video IDs directly without the case sensitivity shortcut that requires the search engine, would include unlisted videos, too, it still would not include private videos. If our method did include unlisted videos, we would have omitted them for ethical reasons anyway to respect users’ privacy through obscurity (Selinger & Hartzog, 2018). In addition to this limitation, there are considerations inherent in our use of the case insensitivity shortcut, which trusts the YouTube search engine to provide all matching results, and which oversamples IDs with letters, rather than numbers or symbols, in their first ten characters. We do not believe these factors meaningfully affect the quality of our data, and as noted above a more direct “brute force” method - even for the purpose of generating a purely random sample to compare to our sample - would not be computationally realistic.

Extracting metadata

The yt-dlp Python package, forked from the widely used but since-deprecated youtube-dl package, provides an extract_info function to extract metadata from YouTube’s internal InnerTube API for a given video. The resulting metadata file contains fields for both user-provided and YouTube-assigned values. Fields like “title,” “tags,” and “description” are provided by users when uploading a video, while fields like “id,” “channel_id,” and “view_count” have values that are set by YouTube. “id” and “channel_id” are examples of fields which are expected to be unchanging, while fields like “view_count” and “comment_count” may change over time.

Hand coding

YouTube's API makes it relatively easy for a technically savvy researcher to extract video metadata like the title, description, view count, tags, duration, number of comments, and default language. Those comprise an important component of our description below, and there are many valuable studies based just on the material accessible through the API. However, there are limitations to metadata inherent in its nature: it is information about the video rather than the content of the video itself. One of the great challenges of studying YouTube or any primarily visual medium is the extent to which available data does not capture much about what is in the video. To fill in some of these gaps, we randomly selected a subset of videos to hand-code. In preparation for this project, we conducted a preliminary hand-coding task in November-December 2021 using a Random Prefix sample, revising the coding scheme based on intercoder reliability, coder feedback, and research needs.

In November 2022, we randomly selected 1,000 videos from our random sample and hand coded them. Fourteen coders were assigned one or more blocks of 50 videos, and each video was coded by at least two people. The coders were a combination of faculty, staff, and students from the University of Massachusetts Amherst, Northeastern University, and the Media Ecosystems Analysis Group. All of the coders received onboarding training to discuss the coding process, and all had experience using YouTube.

Upon starting the task, coders were instructed to open a new Private Window (if using Firefox) or Incognito window (Chrome) to ensure their experience would not be affected by their own personal YouTube account settings or history. They were also asked to disable ad blocking software. For each video, coders were instructed to watch the first 60 seconds, then, for videos longer than 60 seconds, to jump halfway through the remaining video and watch another 30 seconds. This time frame served two purposes, in addition to completing the task within a reasonable time frame: to standardize the unit of analysis with a sample that includes videos between a few seconds and several hours in length, and because what happens in the beginning of the video may influence certain elements of how the video is processed by YouTube (YouTube, n.d.b). While our preliminary coding task indicated these 90 seconds would typically be sufficient in most cases, it is likely this strategy resulted in some phenomena being undercounted due to their presence outside of the watched portion. As noted below, this may especially apply to calls to action and other types of content which video creators often include at the ends of videos.

After verifying the existence of each video, coders were asked up to 32 questions divided into three sections dealing with audio, video, and content. Many of the questions about audio are intended to help us with future research, asking about features that may aid or impede transcription or comprehension. The questions about video are also about quality, but their primary purpose is to distinguish the diverse ways people use YouTube and the genres of videos they create. Finally, the questions about content ask about the presence of advertisements, monetization, and calls to action that are commonly associated with more prominent YouTube genres. The content questions also ask coders to assign videos to a category, using the same categorization system as YouTube in order to see how frequently user-selected categories align with the coders' judgment. A shared document tracked questions and answers to coders' initial questions, and all coders were asked to review the document before beginning.

Of the 1,000 videos coded, 11 were discarded due to errors during the coding process. The codes were broken into four groups based on the amount of data in each. In the first group, coders were asked whether a video existed (n=989; CI 99% ± 4.10%). Videos where at least one coder indicated it did not exist were removed from subsequent groups. For the second group, comprising all extant videos, coders were asked about video characteristics, visual content, and whether there was any audio at all (n=940; CI 99% ± 4.21%). The third group concerned extant videos with audio and included a question about whether there was any spoken language (n=910; CI 99% ± 4.28%). The fourth group was specific to extant videos with spoken language (n=466; CI 99% ± 5.97%).

We calculated intercoder reliability using Krippendorff's alpha. The primary limitation of the hand coding process is that many of the languages spoken in the videos are not understood by any of the coders. Coders were instructed to answer conservatively, erring on the side of coding “no” concerning the presence of certain characteristics unless they could affirmatively detect their presence. For this reason, questions about the presence of political content, religious content, calls to action, spoken advertising, and other linguistically dependent features are likely underestimations. The full list of questions and instructions are in Appendix A, alongside figures for agreement and reliability**.** Our analysis below includes figures about agreement, disagreement, and reliability, as well as some figures restricted to where there was affirmative agreement. For most of the descriptions of hand-coded data provided below, figures for agreement are provided alongside a breakdown of what there was agreement for. In other words, videos for which coders disagreed were not included except where expressly noted.

The size of YouTube

The most important “denominator” question about YouTube is its overall size. How many videos does it host? YouTube occasionally shares numbers of videos of a specific type, but more often focuses on statistics about the amount of time people spend watching the videos. An early figure shared by the company was that people watched 2.5 billion videos in June 2006. This statistic does not tell us whether 250 million videos received 10 views each or 10 videos received 250 million views each. The same announcement celebrated 65,000 new videos uploaded per day that month and 50,000 per day the month before (Reuters, 2006). The company also regularly updates the average amount of video uploaded in a given time period, measured in hours of video rather than number of videos. That number increased from 6 hours per minute in 2007 to 60 hours in January 2012 and 400 hours in July 2015, (YouTube, 2012; Brouwer, 2015)

Cheng, et al. relied on a wildcard-only query (“*”) in YouTube’s search to estimate how much of YouTube’s total videos they had collected through extensive, iterative collection of “related videos.” At that time, in 2008, the wildcard search returned about 77.1 million videos (Cheng, et al., 2008). Wesch also maintained a list of statistics including total videos using this approach. At the time of his last update, in March 2008, the total was 78.3 million. It is worth noting that this technique no longer works, and it is unclear to what extent it ever did provide an accurate total.

Using their "Random Prefix Sampling" method, Zhou, et al. (2011) estimated 500 million videos in May 2011. Vonderau searched for estimates, using multiple third party statistical services, and found they varied significantly, from 80 million on the low end to more than 3 billion at the high end (2016).

We base our estimate on the distance between successful video ID guesses. As described above, we generated our random sample by randomly guessing YouTube IDs, aided by the case insensitivity of YouTube's search engine. Each of our searches generates a string of ten random alphabetical characters followed by one random character from the set of sixteen characters allowed in the eleventh position. Each query is thus 2^{10} possible video guesses. Producing our set of 10,016 videos required generating 18,260,259,669 case-insensitive IDs, equivalent to 18,698,505,901,056 case-sensitive guesses. Dividing the number of video IDs we tested by the number of hits we found, there were 1,866,863,608 guesses in between hits. To put this figure into perspective, if it takes 3 seconds for a human to input a random case-sensitive ID into a URL and see if a video exists, it would take someone an average of 178 years of non-stop guessing to find a single video.

Based on a distance between successful guesses of 1,866,863,608 and the 2^{64} possible IDs in YouTube's system, we can estimate that YouTube hosts 9,881,141,822 publicly searchable videos as of the time of our sampling window in late 2022. We used two methods to determine our level of confidence in this figure. To calculate how many videos we believed we would need for an accurate sample, we created a simulation in which we generated YouTube corpora of random sizes and measured how many guesses were required to converge on an estimate within 1% of the size of the corpus. Based on our successful collection of 10,016 videos and roughly 18 trillion guesses, we calculated a 95%

confidence interval and 2% margin of error using standard error of the mean. Considering our method does not include unlisted or private videos, it is highly likely YouTube hosts more than 10 billion videos.

Estimated growth over time

The proportion of the randomly sampled YouTube videos that were uploaded in each year from 2005 to 2022, or from the year of the first uploaded video to the data collection period, is statistically equivalent to the proportion of videos that were uploaded per year on YouTube overall. We can then apply the proportion of sampled videos uploaded in each year to the estimated overall size of YouTube to estimate the number of videos uploaded in each year. Table 1 lists the percentage of videos in our sample uploaded each year, showing a generally upward trend with a more pronounced increase since 2020. Despite being undersampled due to our collection period taking place before the end of the year, 2022 accounts for more than a quarter of our sample.

Table 1. Percentage of videos uploaded each year.

Year	**Percentage of sample**
2005	0.00%
2006	0.05%
2007	0.22%
2008	0.43%
2009	0.74%
2010	1.13%
2011	1.67%
2012	1.86%
2013	1.97%
2014	2.34%
2015	3.02%
2016	4.25%

2017	5.39%
2018	6.73%
2019	8.81%
2020	15.22%
2021	20.29%
2022	25.91%

Figure 1 shows the estimated total size of YouTube each year, based on the proportion of videos uploaded per year in our sample, and Figure 2 charts annual uploads. YouTube reported significant growth starting in 2020, when a large part of the world was at home more during the COVID-19 pandemic. More people used YouTube, spending more time on the site and uploading more videos (Rodriguez, 2021; Staff, 2021). Videos about COVID-19 information and related subjects were a major draw, although the site was criticized for hosting a large amount of medical misinformation (Li, et al., 2020).

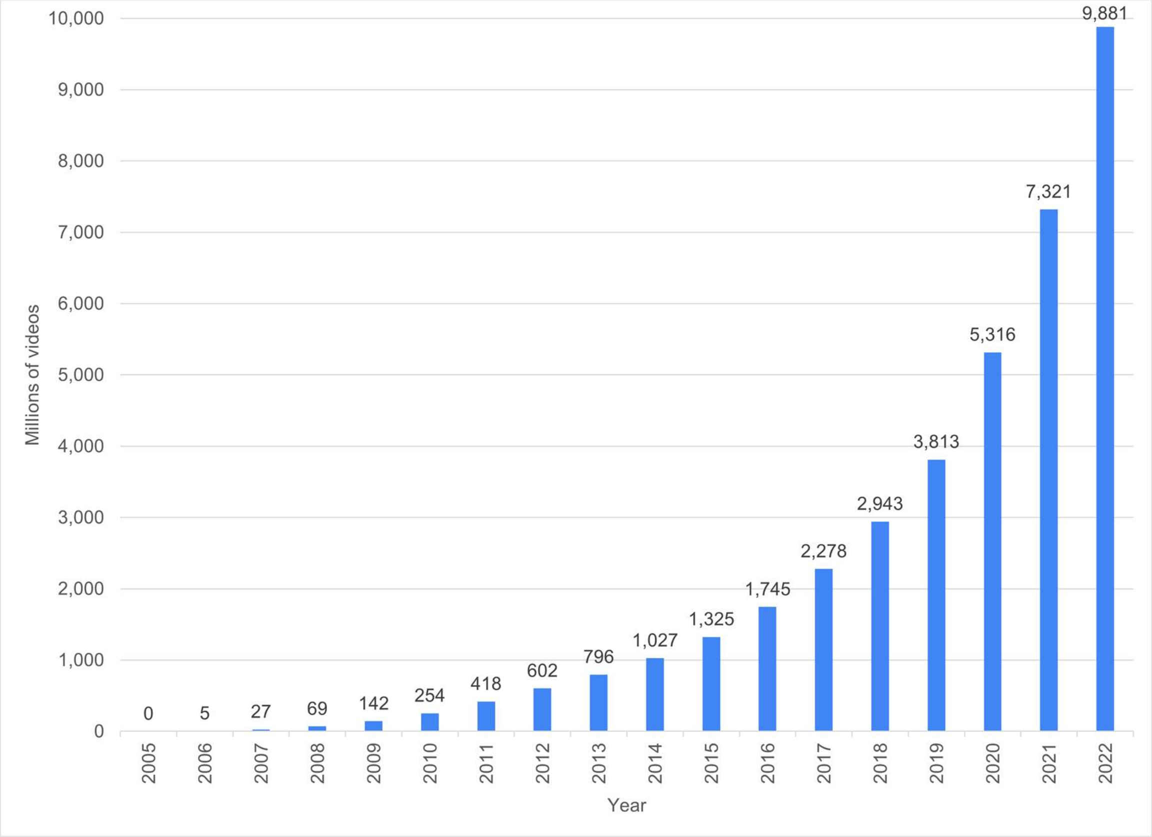


Figure 1. Estimated yearly total size of YouTube (millions).

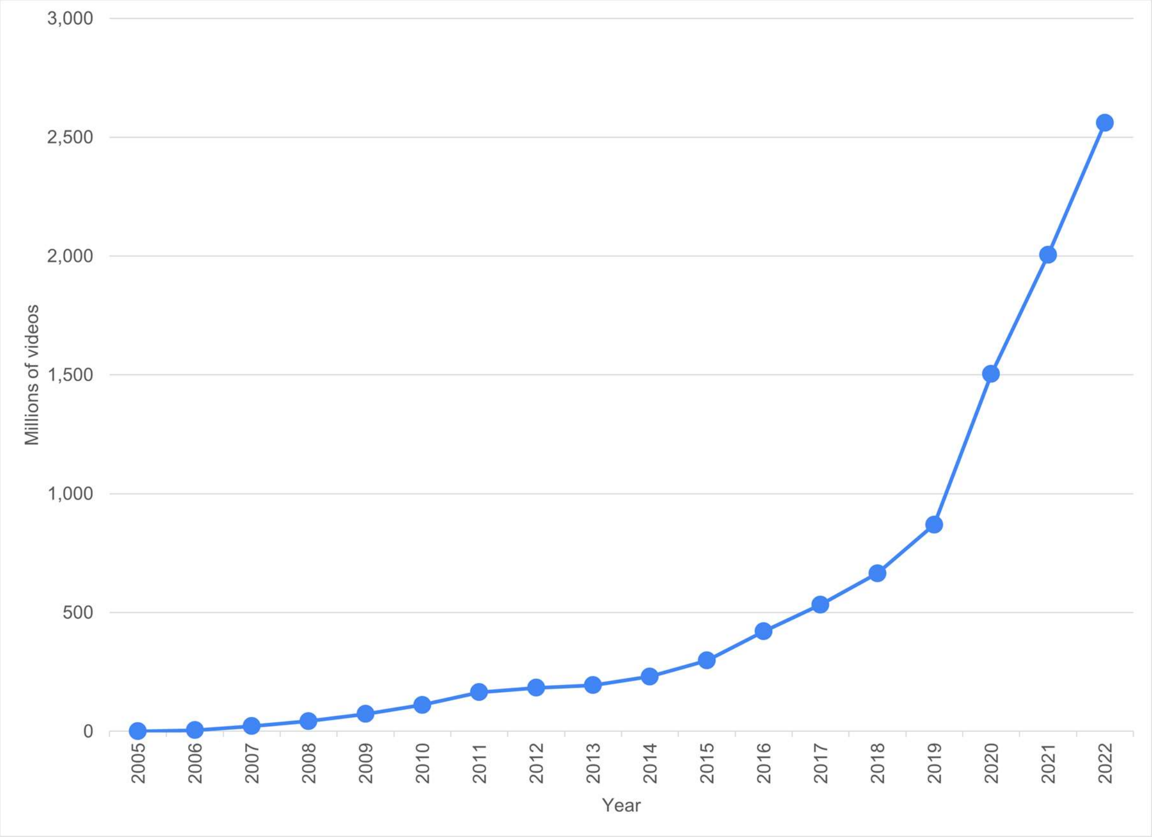


Figure 2. Estimated annual uploads to YouTube (millions).

Popularity of videos

YouTube has nearly 10 billion public videos, but how many of them are actually seen by anyone? While each video's view count is easily available, YouTube does not regularly share aggregate statistics. YouTube reported 600 million video views in March 2009 (Siegler, 2009), but more frequently shares figures about view time. In January 2012, it reported that it exceeded 4 billion daily views and in a February 2017 blog post, it claimed a billion hours of video watched every day (YouTube, 2012; Goodrow, 2017). Again, these broad characterizations do not tell us how those views are distributed - a few highly popular videos, or a larger number of less-popular videos?

There is some existing research on view counts. Cha, et al. reported in 2007 what they considered a surprisingly skewed distribution of views, with 10% of channels accounting for 80% of views, going on to suggest that refining recommendation algorithms could help to surface the more obscure content. Cheng, et al., 2008, found a views distribution (based on 100k videos sampled in April 2007) that found a heavy tail, though with more views to low popularity videos than a Pareto model would predict, which they speculate was because authors watch their own videos multiple times. Bartl (2018) found a generally upward trend of views being concentrated in a small number of channels. Highlighting even greater concentration than Cha, et al., 85% of views in his 10-year study went to 3% of channels, but that represented an average, while yearly figures increased after 2008, up to a high of 90% in 2016. The top 3% of channels also accounted for an average of 28% of uploads in each category.

From our sample, we can examine four dimensions of popularity: views, likes, comments, and subscribers. The mean number of views of videos in our sample was 5,868.02, with a median of 35. The lowest number of views was 0, and the highest was 4,586,494. Videos with no views whatsoever accounted for 4.88% of our sample, which is surprising in that one might expect an uploader to watch their own video at least once. It would be worth exploring the extent to which people might use scripts or software to automatically generate or back-up video content to YouTube in case it is needed for later retrieval. Somewhat fewer, 4.44%, had a single view, while 18.38% had fewer than five, 65.44% fewer than a hundred, and 86.93% fewer than a thousand.

Although videos with more than 10,000 views are likely the kinds of videos people think of first when they think of YouTube, only 3.67% of our sample had that many views or more, but they account for 93.61% of total views across the sample. In fact, just the 16 most viewed videos in the sample (0.16%) together account for more views (50.52%) than the rest of the sample combined. Figure 3 is a closer examination of the distribution of videos with 100 or fewer views. Due to the number of videos with just a handful of views and the long rightward tail of our distribution, we applied a logarithmic transformation to the dataset (natural log) to reduce skewness (Figure 4). A constant (1) was added to allow transformation of zero values. Figure 5 is a Q-Q plot comparing the log-transformed distribution to a normal distribution, illustrating that it is not quite log-normal due in part to the large number of videos with zero views.

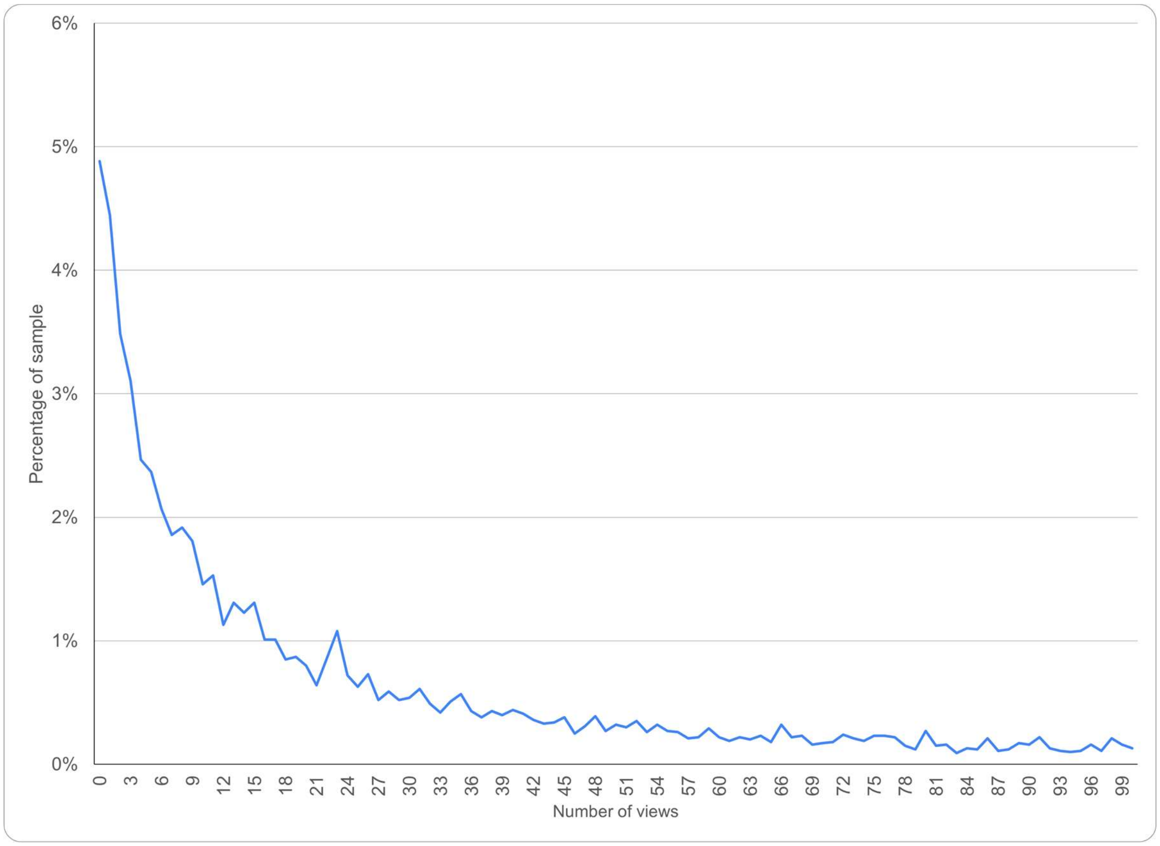


Figure 3. Distribution of views among videos with fewer than 100 views.

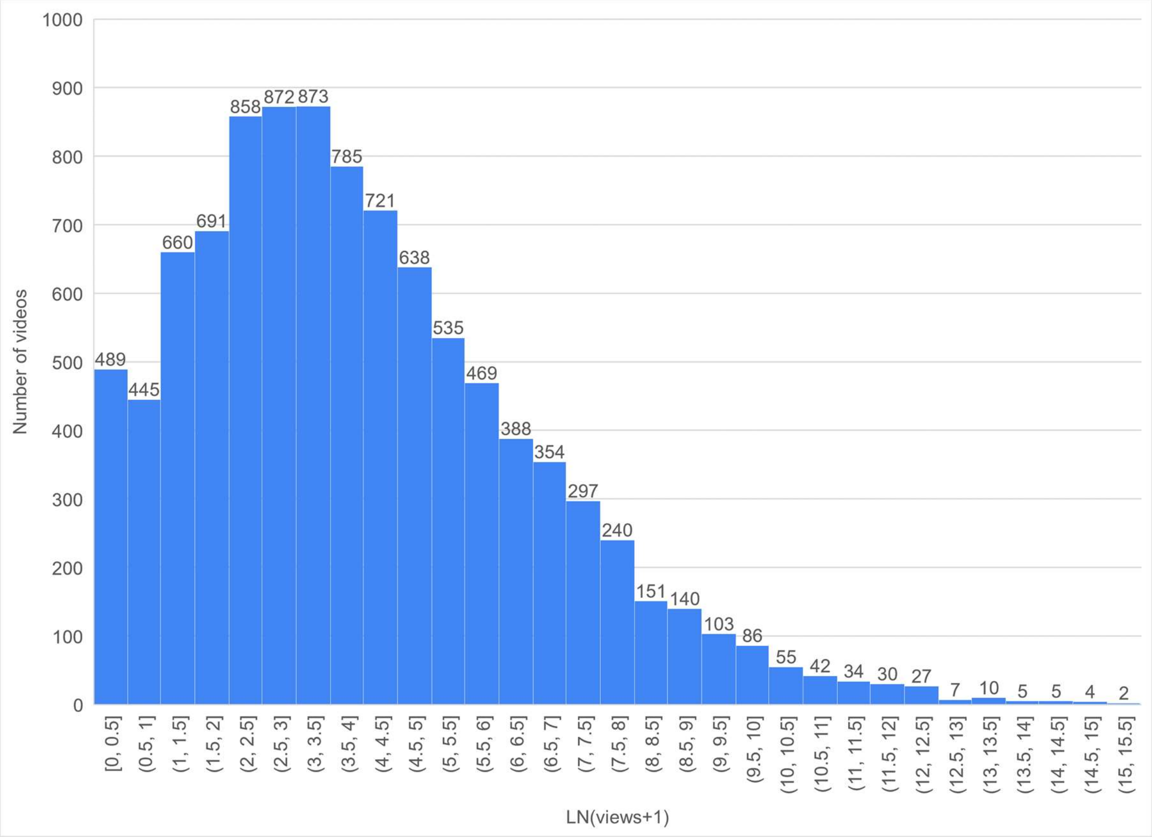


Figure 4. Log distribution of view counts.

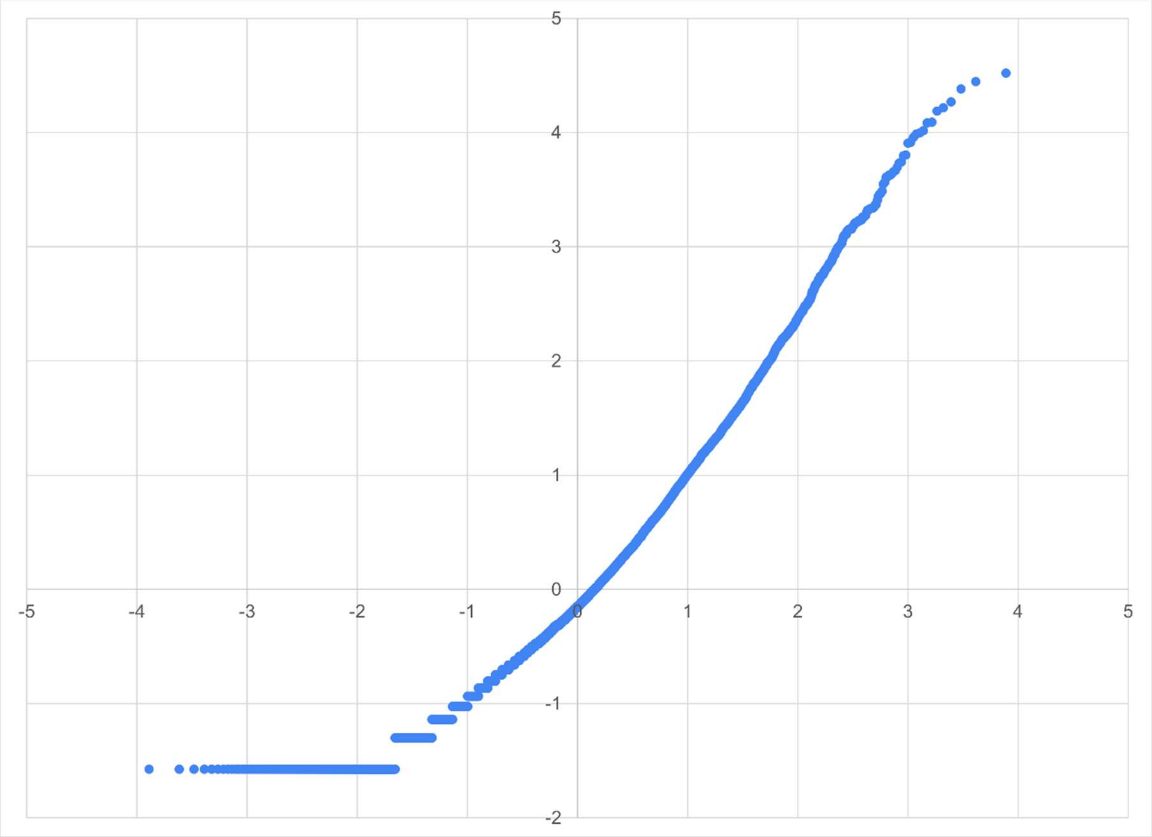


Figure 5. Q-Q plot of ln distribution of views.

The mean number of comments per video in the sample is 5.32, with a median of 0, maximum of 856, and minimum of 0. Most videos (72.64%) had no comments**.** The 1.04% of videos with more than 100 comments (104) account for more than half (54.60%) of all comments in the sample. The log-transformed distribution is Figure 6, again with a constant (1) added to allow transformation of zero values, with a Q-Q plot in Figure 7. A limitation to these figures is that metadata obtained from the YouTube API does not distinguish between videos with no comments and videos where comments have been disabled.

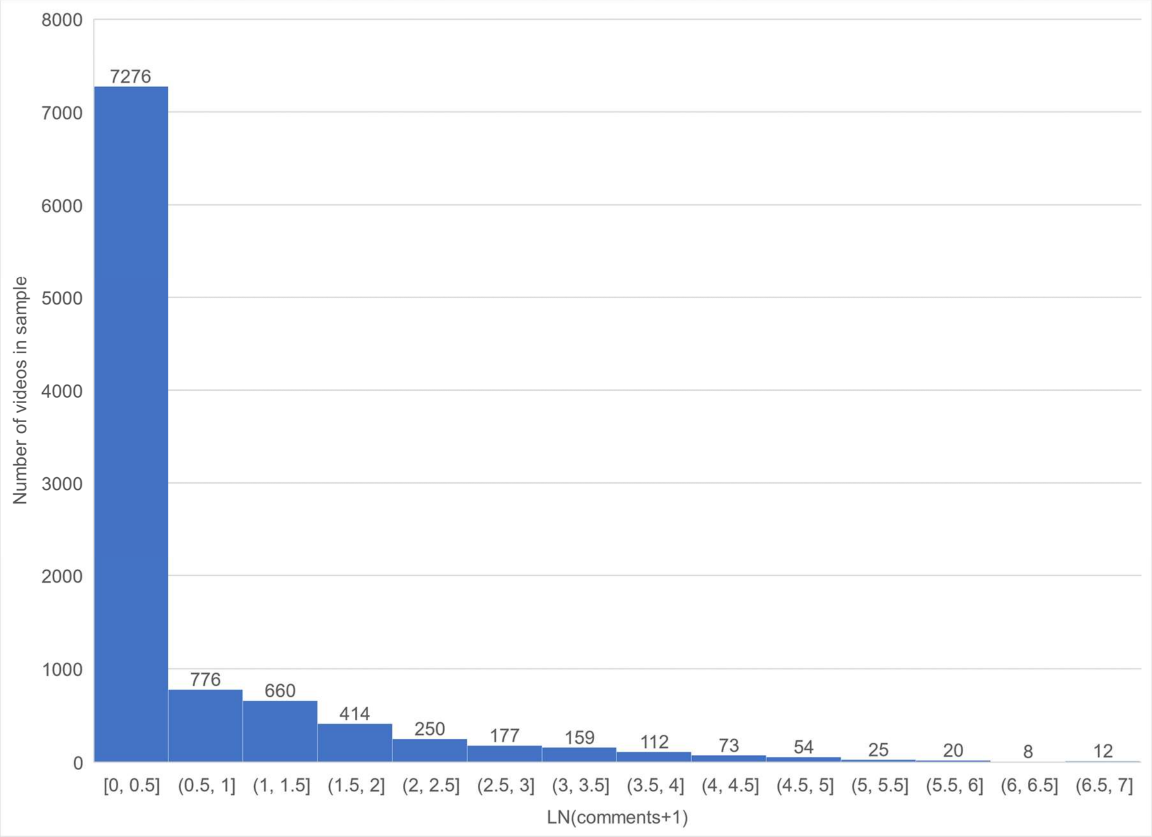


Figure 6. Log transformed distribution of comment counts.

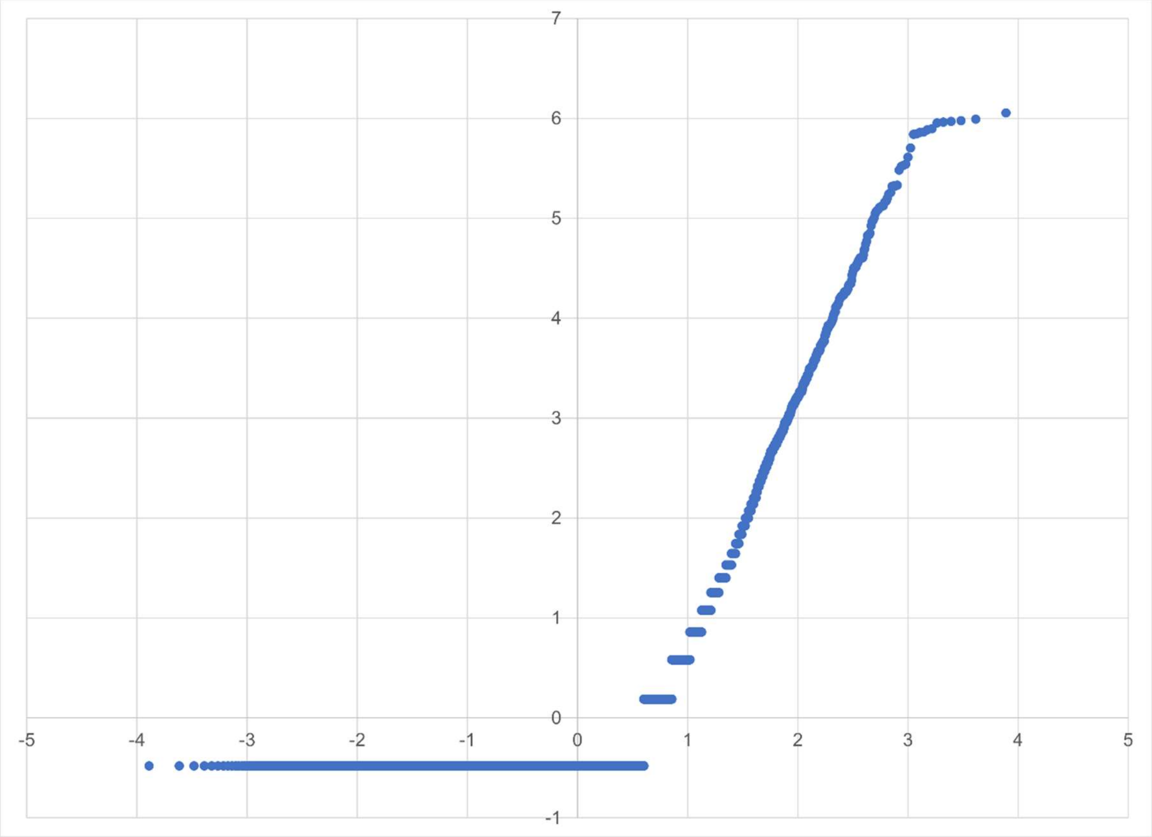


Figure 7. Q-Q plot of ln distribution of comment counts.

The mean number of likes in the sample was 16.48, with a median of 0, maximum of 17,517, and minimum of 0. Even more videos have no likes than have no comments (8,884 or 88.71%). The log-transformed distribution is in Figure 8, and Q-Q plot in Figure 9**.** Eight videos (.08%) account for 54.91% of all likes in the sample.

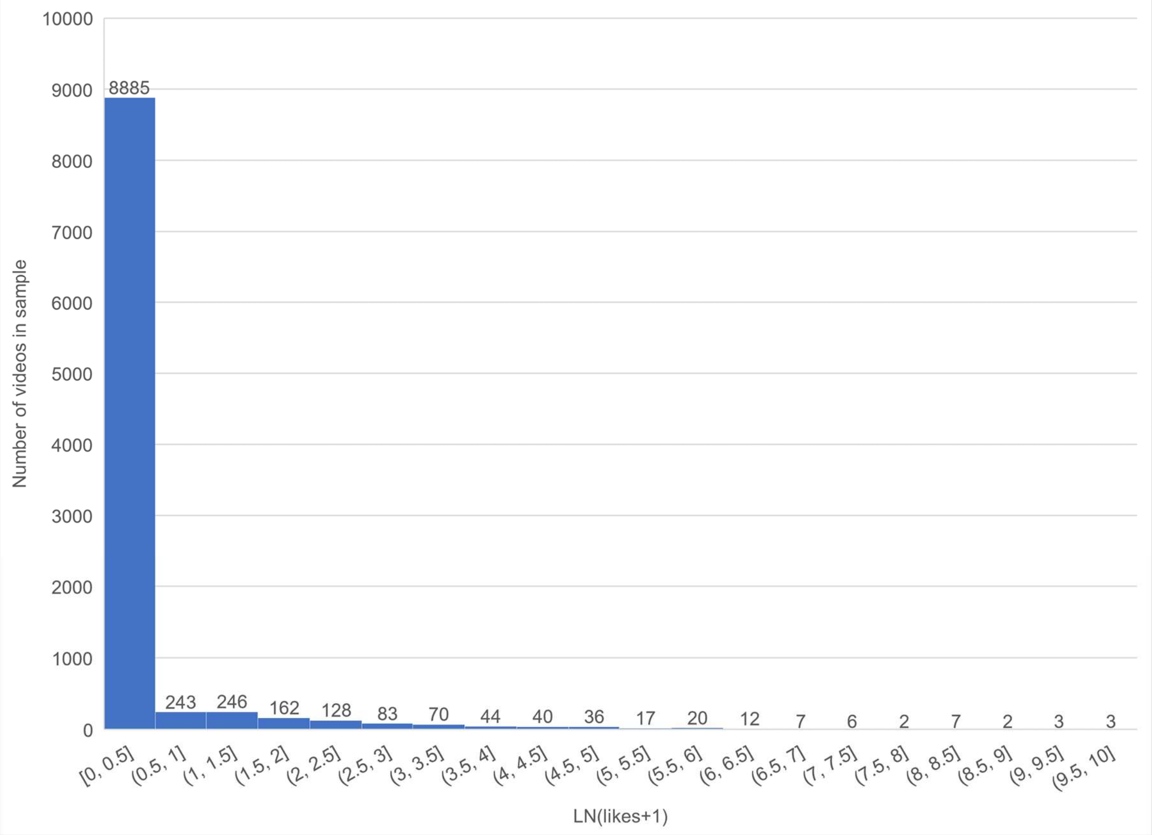


Figure 8. Log transformed distribution of like counts.

YouTube also has a "dislike" feature, but in December 2021 the company hid statistics about dislikes from public view. The decision came after two internal research findings: that dislikes frequently had more to do with viewers' opinions about the channels that hosted the videos rather than about the videos themselves, and that changing the visibility of dislikes was an effective way to curb certain types of attacks and harassment (Jung, et al., 2022). The number does still exist but is only visible to the uploader through either the default interface or the API. The move to disable the visibility of dislikes was controversial, and third parties have developed other mechanisms to attempt to quantify dislike figures, combining data from before December 2021 with numbers generated by users of the third-party tools. These should generally not be relied upon for research purposes, as the reliable, authentic data will continue to age and the reliability of data collected through one or more external tools is incomplete. Such data is also likely skewed towards the behavior of users who make use of the dislike button, while some research shows that the removal of the public numbers may have significantly decreased usage of that feature altogether (June, et al., 2022).

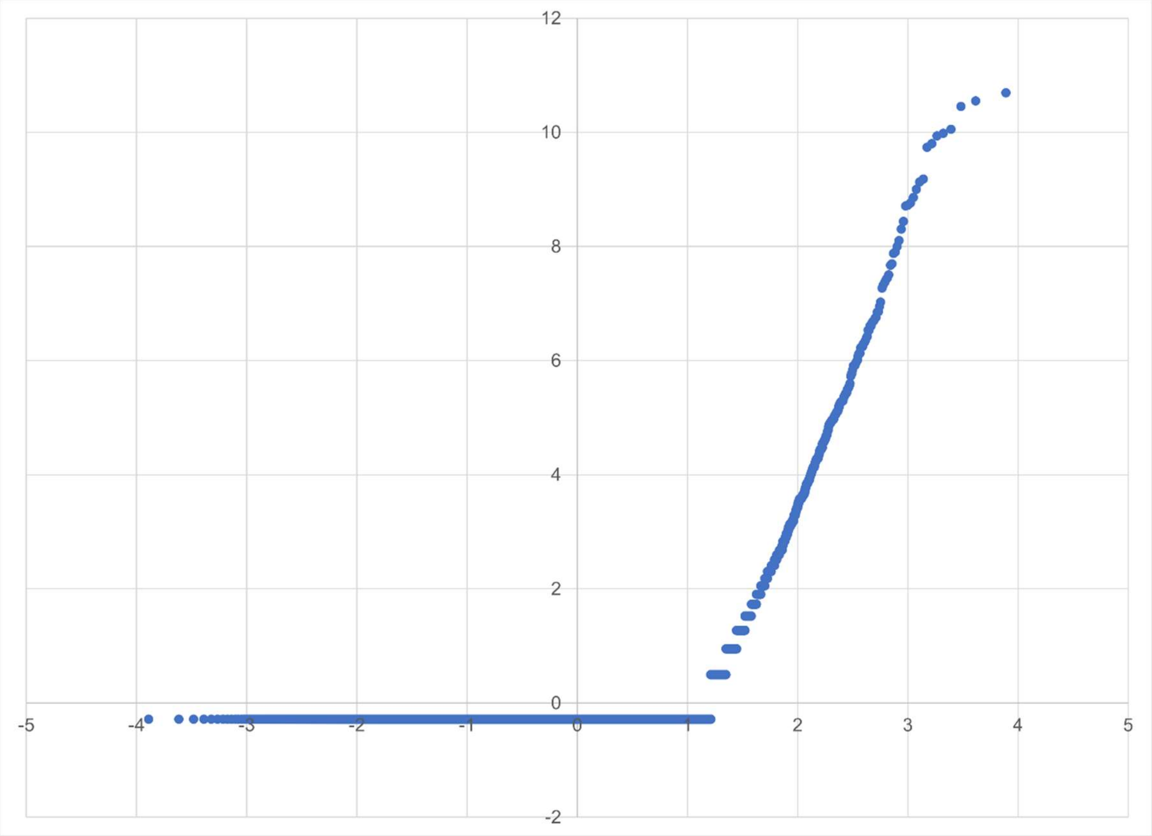


Figure 9. Q-Q plot of ln distribution of like counts.

Our sample is a collection of YouTube videos, not channels, but along with the metadata we retrieved from YouTube are the associated channel identifiers, allowing us to include their subscriber counts. Videos in the set come from 9,997 unique channels. Eight channels had two uploads in the set, and one had three uploads. None had more than three uploads in our set. The mean number of subscribers per channel was 55,100.04, with a median of 61, maximum of 33,100,000, and minimum of 0. Most channels (9,066) had 1 or more subscribers. While channel owners could previously opt to hide subscriber counts, that feature was disabled prior to the collection of this set (Hatmaker, 2022). Figure 10 is the log-transformed distribution of subscriber counts, and the Q-Q plot comparing it to a normal distribution is Figure 11. Although there were still many channels with zero subscribers, the distribution is markedly less skewed towards smaller values than the other popularity metrics.

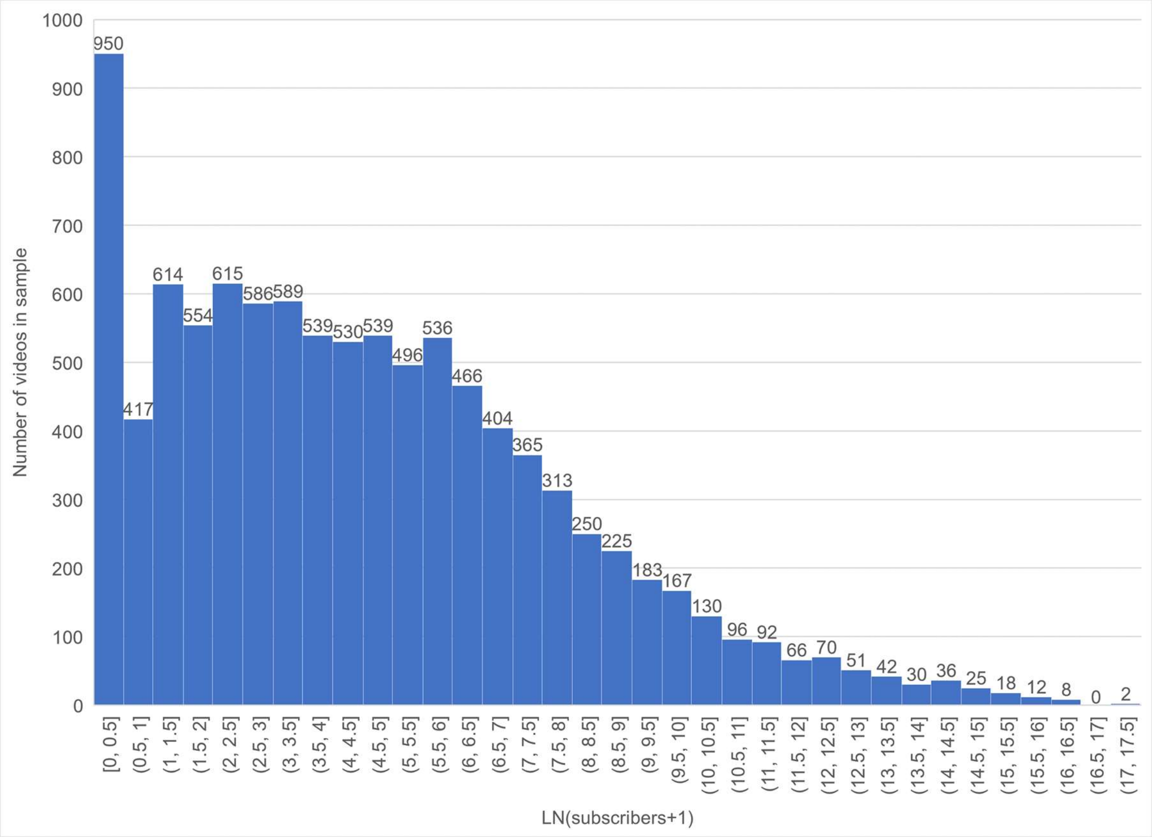


Figure 10. Log transformed distribution of subscriber counts.

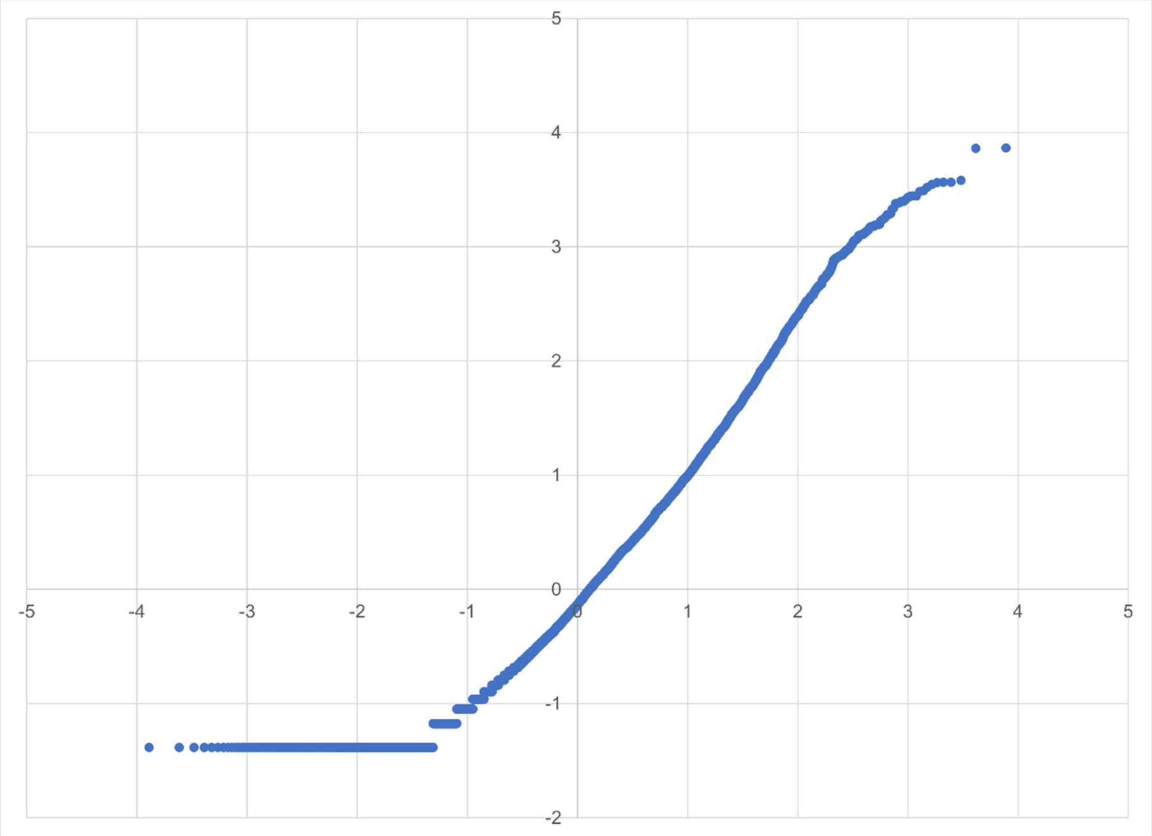


Figure 11. Q-Q plot of ln distribution of subscriber counts.

Distributions of views, comments, likes, and subscribers were all weakly to moderately correlated with each other. The strongest Pearson correlation was between views and comments (r=.46), followed by likes and comments (r=.35), and views and likes (r=.30). The weakest correlations were between likes and subscribers (r=.03) and comments and subscribers (r=.10). Finally, given the assumed function of subscribers to drive traffic to a channel's videos, it is initially surprising that there is only a very weak correlation between views and subscribers (r=.11). There are many instances of videos with a large number of views uploaded to channels with few subscribers, and many videos uploaded to channels with large subscriber counts that do not get many views. YouTube has acknowledged that more than 70% of traffic on the site is driven by its algorithms (Salsman, 2018), suggesting that the remaining 30% of views are directed by subscription, external linking, social sharing, and the other mechanisms through which someone comes across a video. The number of subscriptions may also be skewed towards creators who more effectively persuade viewers to subscribe rather than function as a simple measure of popularity. Some popular channels, like MrBeast, even run regular cash giveaways and contests to motivate subscribers, apart from view counts (Lloyd & Weiss, 2022). While views may be one motivator for promoting subscriptions, subscription counts also have their own internal functions: YouTube provides awards, access to features, and various forms of promotion to channels based on subscriber numbers, regardless of the number of views.

Duration and live-streaming

When YouTube launched, piracy on the site was rife. An early viral success on the site was the *Saturday Night Live* music video “Lazy Sunday,” uploaded without permission (Biggs, 2006). When employees analyzed the content it hosted, they found that among videos longer than 10 minutes, “the overwhelming majority of them were full length, copyrighted videos from tv shows and movies” (Fisher, 2006). This, rather than bandwidth issues, led YouTube to quickly institute a 10-minute limit on uploads. Videos uploaded before March 2006 and people in the premium Director Program were not subject to the limit, but everyone else was. As a workaround, users would upload longer videos in 10-minute pieces. This was reflected in Cheng, et al.’s set from 2008: 97.9% of videos were within 600 seconds and the distribution had peaks in the first minute, at 3-4 minutes (which they attribute to the popularity of music videos), and near 10 minutes, suggesting segments of larger works (Cheng, et al., 2008). YouTube raised the limit from 10 to 15 minutes in July 2010, only after it had implemented its Content ID system, which automatically detects certain copyrighted material (Siegel, 2010). The 15-minute limit remains for new users, but as soon as an account is verified the limit is now the lesser of 12 hours or 256 gigabytes.

The mean duration of videos in our set, measured in seconds, was 615 (just over 10 minutes), with a median of 126, maximum of 43,199 (reflecting the 12-hour limit), and minimum of 1. 6.24% of videos were 10 seconds or less, 37.94% were a minute or less, and 81.77% ten minutes or less. Only 3.9% of videos were an hour or more. Figure 12 shows the log-transformed distribution of video length in our sample.

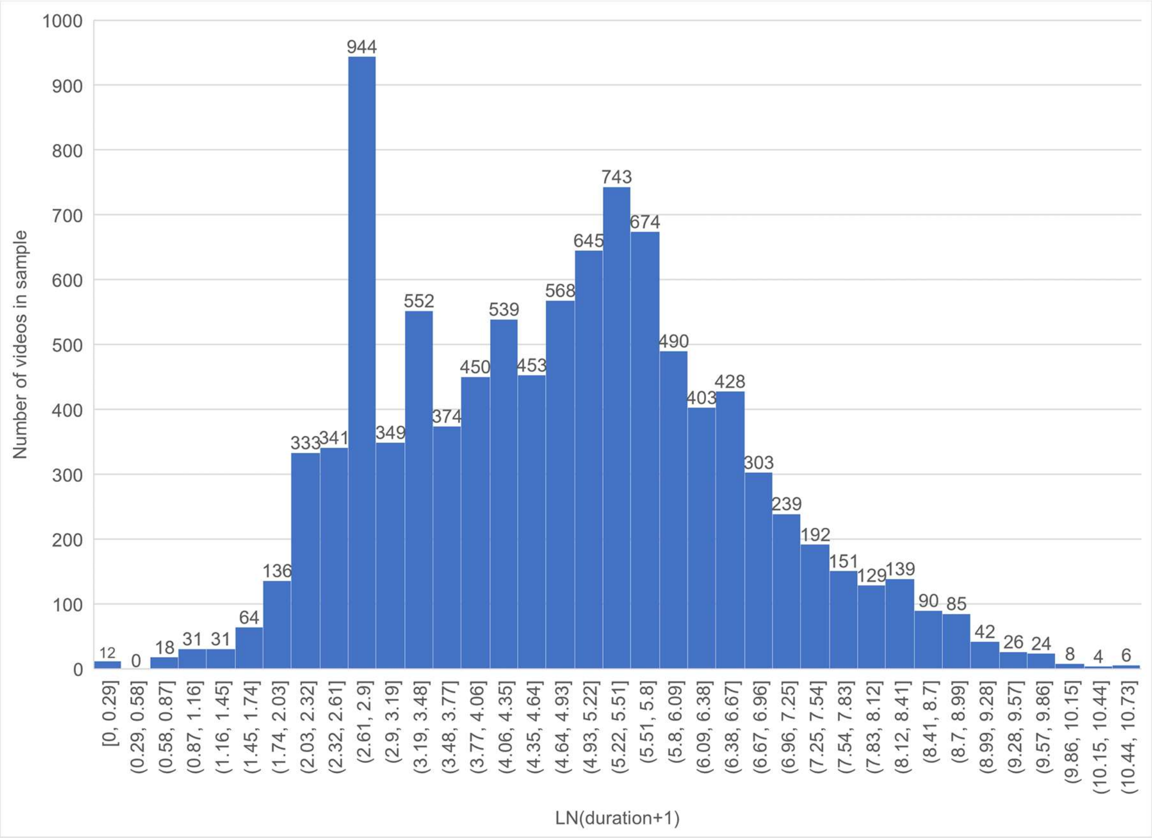


Figure 12. Log transformed distribution of video duration.

Streaming video posed several challenges in the early years of the web. Video takes up hard drive space, broadband internet was still uncommon in most places, and computers could not process video as quickly as today. Shorter video has been the norm for much of the history of the internet for these reasons. YouTube and other sites further contributed to the tendency for online video to be short by making it easy to share the content with others, cultivating the desire to "go viral" and encouraging users to watch many short videos rather than a few long ones. Figure 13 is a breakdown of videos a minute or shorter, showing a high number of 15-second videos**.** The concentration at 15 seconds may reflect the influence of TikTok, which uses 15 seconds as a default length. Our sample contains many examples of TikTok videos reuploaded to YouTube, some of which appear to have been uploaded by the creators themselves as a form of cross-platform engagement or archival, and some by fans, remixers, aggregators, or other users. Driven by competition with TikTok, YouTube also rolled out its own short video platform called YouTube Shorts, with a beta version released in India in 2020, following the country's ban of TikTok (Hern, 2020). It became available in the US in March 2021, and released to the public in July 2021.

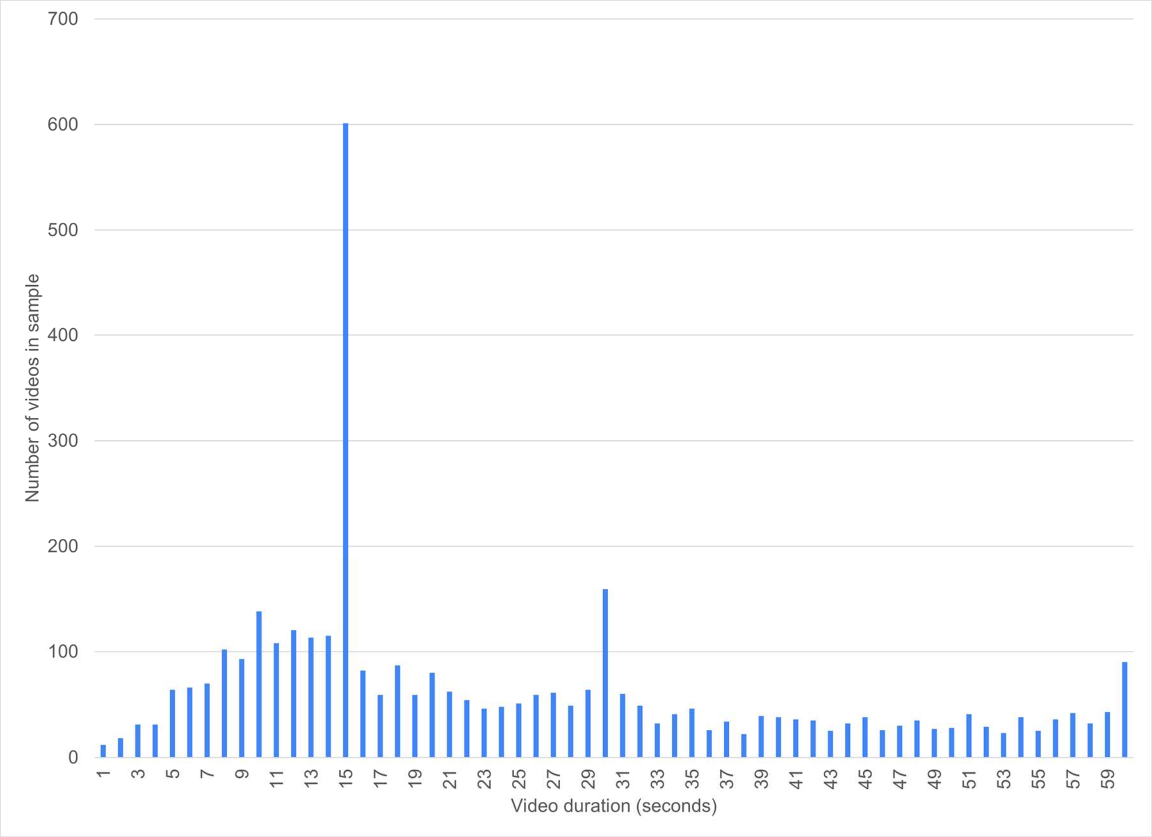


Figure 13. Distribution of videos 60 seconds or less.

In our sample, 81.77% of videos were 600 seconds or less. This is a decrease from when Cheng, et al. were writing in 2008, when 97.9% of videos were in that range. One development which may account for some of the longer videos is live-streaming, alongside the emerging popularity of long-form video game and podcast content (Smith, et al., 2013). As with TikTok, there are examples in our sample of Twitch content cross-posted to YouTube, and as with TikTok, YouTube has worked to expand its features in the niches where its competitors excel. YouTube first released its own live-streaming functionality in April 2011, but it was only available to certain partners until May 2013, when it was opened to anyone with more than 1,000 subscribers. A few months later the threshold lowered to 100 subscribers, and was removed altogether in December 2013, although 50 subscribers are still required to stream from the mobile application (Salgar, 2013a; Salgar, 2013b; YouTube, 2023).

583 (5.82%) of the videos in our sample were live-streamed according to the metadata, which is higher than we anticipated, and does not include videos which were live-streamed on other platforms and then uploaded to YouTube. The mean duration of live-streamed videos was significantly longer than other videos, at 4,681 seconds (78 minutes), with a median of 3,135 seconds (52 minutes).

Categories and tags

YouTube maintains a set of categories which uploaders can select as a default category for their channel but also choose from when uploading a video. At the time our sample was generated, the available categories were: Autos & Vehicles, Comedy, Education, Entertainment, Film & Animation, Gaming, Howto & Style, Music, News & Politics, Nonprofits & Activism, People & Blogs, Pets & Animals, Science & Technology, Sports, and Travel & Events.

Cheng, et al. published an early breakdown of category usage in 2008, based on a very large set of videos, gathered mostly through spidering “related” videos: Music 22.9%, Entertainment 17.8%, Comedy 12.1%, Sports 9.5%, Film & Animation 8.3%, People & Blogs 7.5%, Gadgets & Games 7.4%, News & Politics 4.4%, Autos & Vehicles 2.6%, Travel & Places 2.2%, Howto & DIY 2.0%, Pets & Animals 1.9%, Unavailable 0.9%, and Removed 0.5% (Cheng, et al., 2008). Category titles have changed since then, but it is noteworthy for the popularity of music at the time.

Bartl (2018) was especially interested in categories and genres in his analysis of his random set. He noted that between 2006-2009, the “Music" category was dominant, with 20% of new channels primarily uploading music videos in 2009. Starting in 2010, however, the "People & Blogs" category exploded with popularity, amounting to nearly 75% of uploads in his sample in 2016. In 2012, "Gaming" was the second most popular category, and remained there through the end of his sample period (2018). It is worth contextualizing these findings with YouTube setting People & Blogs to the default category around the time Bartl noted its popularity. The very existence of a default makes the category system difficult to rely upon for research purposes.

Figure 14 shows the distribution of categories for our random sample. The People & Blogs category accounts for more videos than the other fourteen categories combined (55.82%). Most categories account for less than two percent of videos each. Only Gaming (the distant second place to People & Blogs), Entertainment, and Music are more than 5%.

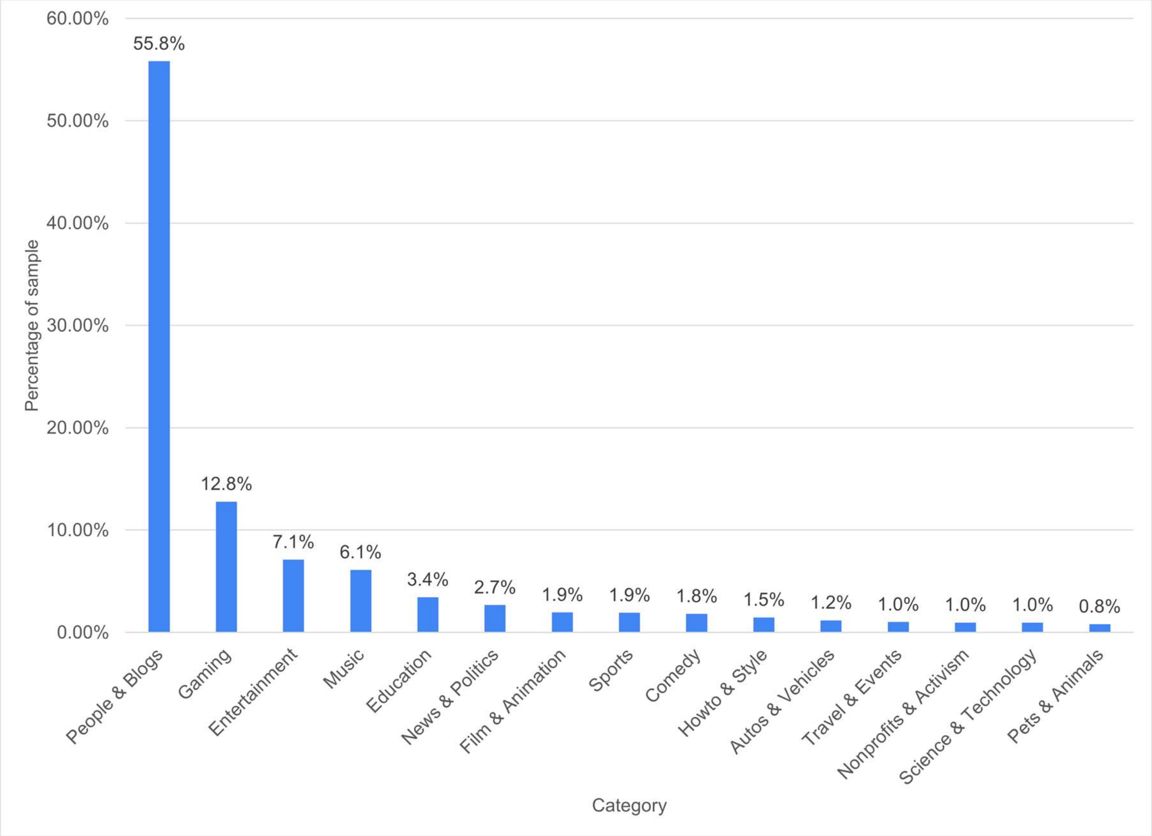
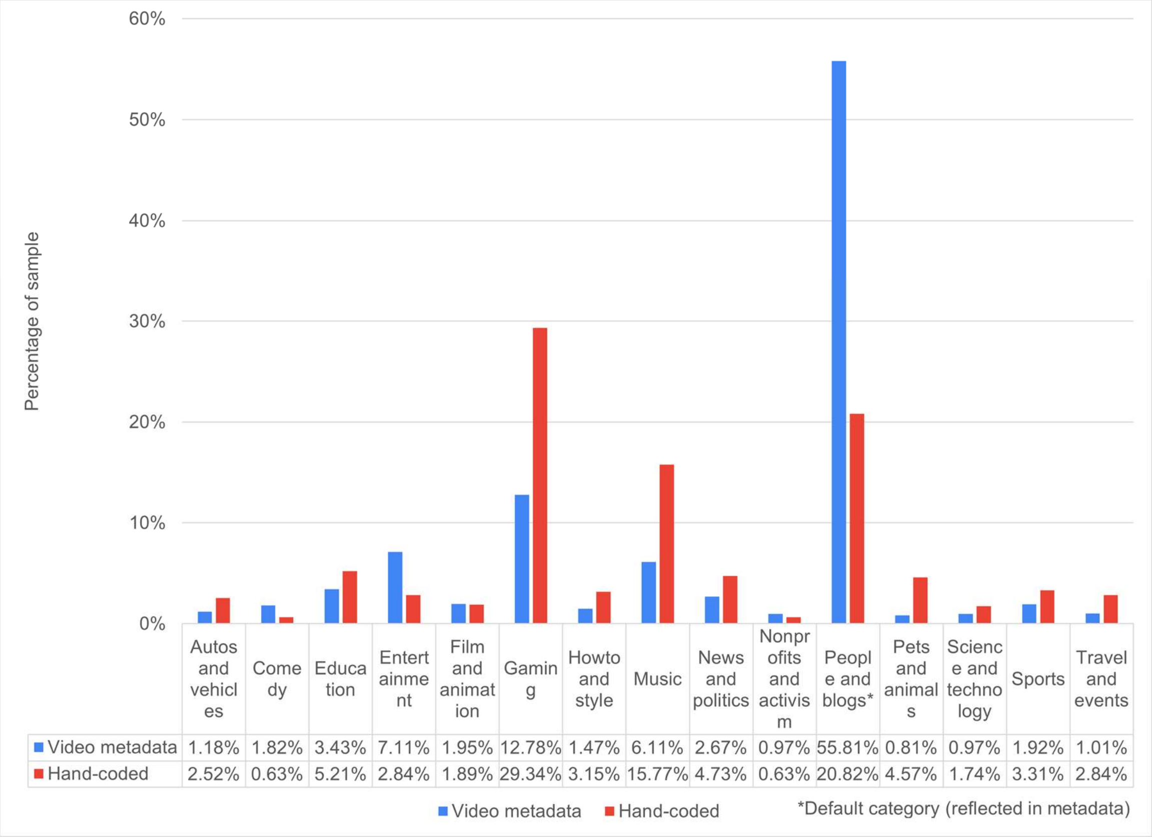


Figure 14. Distribution of categories.

The process of hand coding videos reveals just how difficult categorization can be, as well as how rudimentary YouTube's category set is. For example, there are many short videos of people dancing for the camera, in part perhaps due to the influence of TikTok. To call all of these "Music" or "People & Blogs" does not capture the performance that is central to the content. In fact, there is no category for "Arts" which might capture dance, painting, graphic art, or sculpting.

In the subset of random videos which were hand-coded, coders were asked to categorize the video using YouTube's category system in order to see how frequently coders' judgments aligned with the uploaders' (as well as with each other). Of the 940 videos coded for category, coders agreed with each other only 67.45% of the time. In those cases where coders agreed, the category they agreed upon matched the video metadata only 28.39% of the time. Framed another way, at least one coder disagreed with the uploader's choice of category 80.85% of the time. Whereas YouTube videos in our sample were predominantly People & Blogs, the distribution of categories by our coders favored Gaming followed by People & Blogs and Music. Those three categories accounted for 65.93% of the hand-coded agreements. The more even distribution of categories in the hand-coded set may be a result of the lack of a default category. Where some uploaders might not take the time to select a category and thus be placed in People & Blogs by default, coders had to make a decision. The full comparison between categorization in the metadata and the hand-coded videos is in Figure 15.



	Autos and vehicles	Comedy	Education	Entertainment	Film and animation	Gaming	Howto and style	Music	News and politics	Nonprofits and activism	People and blogs*	Pets and animals	Science and technology	Sports	Travel and events
🟦 Video metadata	1.18%	1.82%	3.43%	7.11%	1.95%	12.78%	1.47%	6.11%	2.67%	0.97%	55.81%	0.81%	0.97%	1.92%	1.01%
🟥 Hand-coded	2.52%	0.63%	5.21%	2.84%	1.89%	29.34%	3.15%	15.77%	4.73%	0.63%	20.82%	4.57%	1.74%	3.31%	2.84%

*Default category (reflected in metadata)

Figure 15. Comparison of metadata and hand-coded categorization.

Whereas every video has a category, even if just the default People & Blogs, the use of tags is optional. Tags are optional keywords or phrases intended to aid search and navigation. Users can add as many tags as they wish, but YouTube is clear that overtagging will result in deprioritization or removal of a video from search. Unlike the personal organizational function of tagging on sites like Del.icio.us or the expressive function like tags attached to a personal profile, the tags attached to individual YouTube videos serve to aid discoverability. Early research of tagging on YouTube found that while other platforms like Flickr use tags for categorization purposes, YouTube uploaders see them as basic descriptors of the video like the title and description, often using very specific terms rather than more generic categories (Greenaway, et al., 2009). Paolillo found the use of tags as a social tool, identifying groups of users and enabling them to share content with each other (2008). As is typical of tagging and folksonomies, they can be difficult to use meaningfully apart from the broadest terms and are inherently tied to language and regional linguistic variations. They may also be used with awareness of YouTube's recommendation algorithms, and thus prioritize tags which attract attention over the most accurate descriptors (Simonsen, 2011).

Several studies have relied upon tagging to identify relevant datasets, but our random sample indicates most videos (62.57%) do not even use them. In our sample, there were 31,667 unique tags used in 3,749 videos. The most common tags were: PlayStation 4, Sony Interactive Entertainment, #PS4Live, #PS4share, Fortnite, news, video, game, Gameplay, Music, and SHAREfactory. Some of these tags appear multiple times not because they were manually input but because of the way users uploaded them. Specifically, the Sony Playstation video game consoles can be linked directly to YouTube to facilitate sharing clips, and doing so includes a set of tags about Playstation and the game being played by default, although we do not have a way to distinguish automatically added tags from user-input tags. Figure 16 is a word cloud of the most common 250 words.

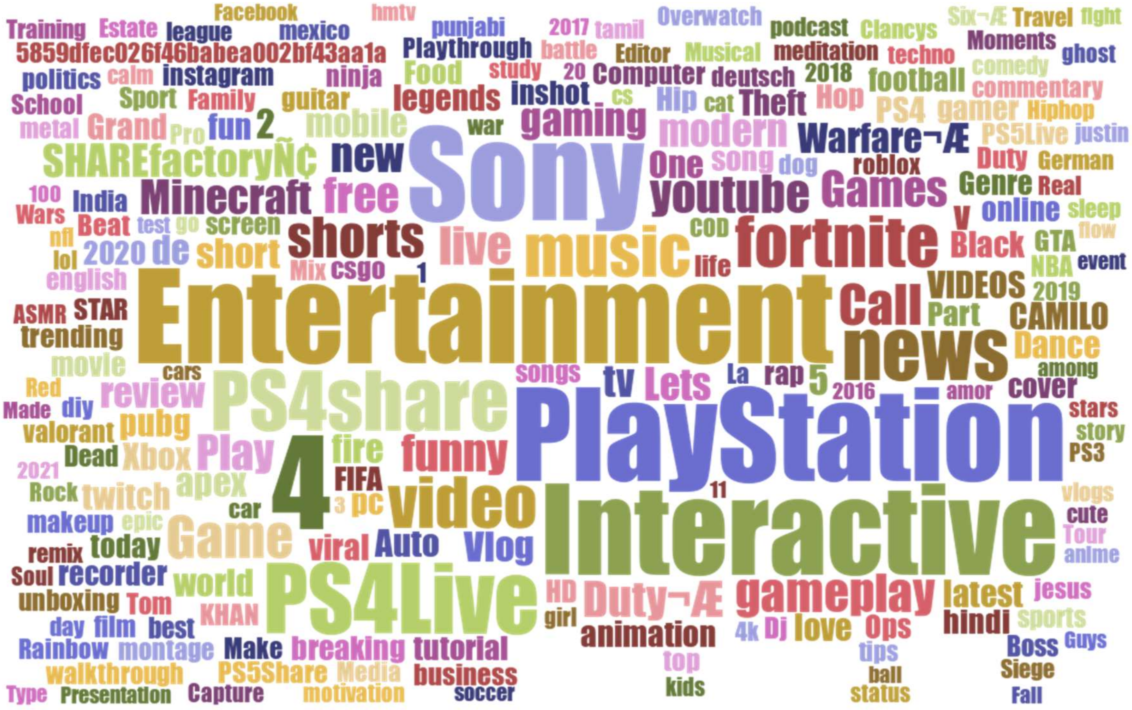


Figure 16. Word cloud of the most common 250 words in tag metadata.

This project to better understand YouTube as a whole began with discussions about an "International Hate Observatory", a project which aims to understand issues of hateful and harmful speech from a high level, bringing together research on and across a number of media and platforms. In order to find, measure, detect, and understand such speech at scale, we need to address denominator and distribution problems. That means being able to answer fundamental questions like "how many videos are on YouTube" but also "how much news, political, or current events-related content is on YouTube." Alongside the big picture questions in our hand-coding task, we included two which asked about specific subjects. We asked whether there is anything having to do with news, politics, or current events, and whether there is anything to do with religion. Both of our measurements will be undercounts due limitations on the languages our coders are able to understand. When coders could agree (96.06% of the time), they only found 2.99% of videos to have anything to do with news, politics, or current events. Slightly more, 3.84%, had anything to do with religion (based on 96.91% agreement). Both of these numbers were somewhat surprising. We know that many news companies use YouTube to distribute content, and that there are many politics and current events-related YouTube creators. To read the research and journalism about harmful content on YouTube (e.g. Lewis, 2018; Tufekci, 2018; Ribeiro, et al., 2021), one might be led to think it is ubiquitous on the site, but in fact it represents a tiny piece of YouTube as a whole. Religion was not initially a category we intended to look for, but in our preliminary coding task we noticed a surprising number of religious videos, including full-length religious services, vlog-style worship, and religious music.

Video characteristics

Metadata provides only a small amount of information about what a user sees when they watch a video, but data like horizontal and vertical resolution, links to thumbnails, and whether it is divided into chapters are of little use when looking to characterize video content. To gain a better sense of random videos' visual data, we included several questions about video production in our hand-coding task. The first question concerned whether the video was entirely or almost entirely still images or slightly modified still images. The proliferation of third-party phone applications has enabled users to turn their pictures into slideshows or add filters to make a static image appear more dynamic (for example, using sparkle effects). Our hand coders found 15.54% of videos in our sample are primarily still images (based on the 95.85% of videos for which coders agreed).

Given the popularity of different types of video game-related content on YouTube, we suspected that a random sample would include a lot of not only video game-related content, but videos where the only visual material is a video game. We asked coders to indicate if a video were entirely or almost entirely a video game, not including videos where a live person is also superimposed on the screen or when video games are intercut with other media. In our sample, 19.52% of videos fit this description (based on 98.09% agreement). Additionally, 8.35% of videos were entirely or almost entirely graphical (based on 81.70% agreement), meaning computer-generated graphics other than video games such as phone screens, PowerPoint presentations, or animated music visualizations, although that question had α < 0.667, the threshold Krippendorff recommends for tentative conclusions (Krippendorff, 2004).

When most people think of YouTube, they think of popular YouTube channels which exhibit a higher level of professionalism than random uploaders. In an effort to characterize the amount of effort and polish in our random set, coders answered several questions dealing with video production. Coders indicated 38.26% of videos included some form of editing, although there was less agreement among coders as to what constituted an edit (84.26% agreement). A larger percentage of videos than we anticipated had some form of set or background design (14.29%, based on 87.87% agreement) or a custom graphical introduction (13.37%, based on 91.49% agreement), both of which require greater consideration for production quality or brand identity than we believed a random sample would include. Some form of text, graphics, or other video overlaying the main video was present in 36.69% of videos (based on 80.32% agreement), but many of those were automatically added by third party applications or websites rather than an indication of production capabilities. Both the overlay and set/background design questions had α < 0.667.

More than half of the sample (55.86%) had some recorded video (apart from still images, graphics, or video games), with 35.67% primarily recorded indoors, 18.14% primarily outdoors, and 2.06% unclear (based on 87.98% agreement). Recorded video, where present, was considered noticeably shaky 52.33% of the time (based on 84.47% agreement). A human talking to the camera was seen in 18.32% of videos (based on 91.17% agreement) and 9.13% of videos include footage of a public or semi-public event (based on 95.53% agreement). Finally, when coders were asked if the video was obviously the work of someone else, such as a clip of a movie, television broadcast, or music video uploaded to a channel which clearly does not own that clip, coders indicated it was true only 4.81% of the time (based on 95.11% agreement, but with α < 0.667).

Advertisements and monetization

Advertisements form the backbone of Google's business model, but the algorithms which determine when a viewer sees an ad or which ad is displayed are kept private. Prior to 2020, uploaders could decide whether they wanted to be part of the advertising program, but a November 2020 Terms of Service update allowed Google to display ads on any video (YouTube Team, 2020). Qualifying uploaders could still decide not to participate in the advertising program, but that just means they would not receive a portion of the advertising revenue.

In our preliminary coding task, we found low intercoder reliability for questions asking whether a video or text ad was shown before or during a video. We retained the questions for our hand coding task for this paper because it is worth noting just how inconsistent the results are, even between coders working in "private" or "incognito" browsers. The question about whether a coder saw a text-based ad, such as a banner ad, had overall agreement of 91.91%, but a low intercoder reliability ($\alpha = 0.34$). The high agreement was because most of the time, neither coder saw a text-based ad, and the low Krippendorff's alpha figure is due to the ratio of answers and disagreement between coders when someone did see such an ad. Both coders saw a text-based ad 2.55% of the time, but at least one coder saw a text-based ad in an additional 8.09% of videos. Video ads were somewhat more consistent, as well as somewhat more common. Agreement on the video ad question was 93.3% with $\alpha = 0.65$. Both coders saw a video ad 7.45% of the time, and at least one saw a video ad an additional 6.7% of the time. Again, there are more instances when one person saw an ad and the other did not than there were videos where both coders saw an ad. It is unclear to what extent this is based on a certain amount of randomness or other variables such as location or browser information.

Coders were also asked whether the video was itself an advertisement for something other than the video or channel itself. Our random sample included footage of cars and home goods for sale, for example, as well as movie trailers, but coders only agreed that the video was an advertisement 3.09% of the time, with at least one indicating the video was an advertisement an additional 5.85% of the time. Again, the high degree of agreement that the video did not serve as an advertisement combined with the disagreement yielded a low reliability figure ($\alpha = 0.48$).

Monetization and calls to action are ubiquitous among the most popular YouTube channels, which might lead one to expect them to be common across YouTube broadly. We found that these are both exceedingly rare in our hand-coded sample. We defined monetization by the creator or uploader as including sponsorships, promotion, and advertising beyond the video and text-based advertisements that YouTube displays. We defined a call to action as a request for engagement outside of advertising, such as an invitation to subscribe, like, comment, visit a website, or visit a social media account. Coders only agreed that monetization was present 0.21% of the time, with at least one person indicating there was monetization an additional 2.77% of the time, and only agreed on the presence of calls to action in 3.83% of videos, disagreeing an additional 13.51% of the time. Both of these questions had very low reliability scores (monetization $\alpha = 0.12$; call to action $\alpha = 0.29$), which may be due in part to unforeseen gray areas in the definitions as well as the need to be able to understand the language in the video to distinguish these features. It is also possible that these features, especially calls to action, are more common at the ends of videos, which would not be captured in our methodology for videos over 90 seconds.

Audio characteristics

In our hand coding project, we dedicated more questions to audio than video. Audio is the primary medium for linguistic content on YouTube, and thus represents the greatest opportunity for researching the topics discussed on the site. While video content is also important, it presents a much wider range of technical challenges to parsing its meaning, whereas the availability of YouTube captions and third-party transcription software means we can process linguistic content from audio using tools that already exist. Most of the questions we posed to coders are thus concerned with the quality and character of the audio – the sort of information that could affect our ability to reliably understand or transcribe it.

First, we asked whether there was any audio at all. Of the videos that existed, coders agreed that 96.81% of videos had some audio. The next question separated videos which contain nothing but music from the rest. Any speech can be turned into music, but there is a practical distinction in that transcription software is not very good at picking up sung language. Coders agreed on this question 93.74% of the time, and among the agreed upon videos, 40.56% were judged to be entirely or almost entirely music. Although it is well documented that music is popular on YouTube, this number is higher than expected. Not all of the videos which are entirely music are categorized as such, however, as many of them just use music as background or context for a performance, video game clip, or slideshow. A follow-up question asked whether the audio obviously belongs to someone other than the uploader, such as when a small channel uses a popular song or a clip from a movie. Of the 85.60% of the time coders agreed, 32.35% of videos were obviously the work of someone else.

The ability to determine, at scale, whether videos are music is important to the transcription process as well as for a general understanding of the content of a given sample. In pursuit of a way to automate the process, we cross-referenced our sample with videos available through YouTube Music, the company's music streaming service. YouTube Music is built on top of YouTube, allowing subscribers to access additional features for a subset of music-based videos on the site, such as the ability to download songs and listen to them offline. YouTube Music does not contain all music on YouTube, but knowing how reliably music is included in the service would let us determine how useful it would be for research purposes.

Among the videos in our random sample categorized as Music, only 66.67% were part of YouTube Music, and only 32.96% of those which were part of YouTube Music were in the Music category. In other words, most of YouTube Music is not categorized as

Music, and a third of what is categorized as Music is not part of YouTube Music. Among the subset of our sample hand-coded as being in the Music category, 75.00% were part of YouTube Music. For the separate question of whether a video's audio was entirely or almost entirely music, only 34.20% were part of YouTube Music. So in our samples, categories are not reliable for distinguishing music from other types of videos, and while YouTube Music could be used to identify *some* of the music-based videos in a sample, it does not include enough to be a reliable differentiator.

Among the videos where coders agreed there is audio, 53.87% had spoken language in the first minute (95.05% agreement). The following figures are based just on the videos which had spoken language. Where coders agreed (83.69% of the time), 28.97% had spoken language on top of audio (22.05% on top of quiet music; 6.92% on top of loud or distracting music). Though there was some disagreement (2.58% of the time), coders did not agree on any videos containing text-to-speech language. In 11.11% of videos where coders agreed, there were at least two seconds where multiple people were speaking at once (based on 84.98% agreement). We also asked coders to provide an overall level of audio quality: high quality, medium quality, or low quality. There was considerable disagreement (only 58.80% agreement), but where there was agreement, 18.61% were high quality, 55.11% were medium quality, and 26.28% were low quality. Finally, we asked whether the audio quality varied significantly over the course of the video (6.06% said yes among the 84.98% of times coders agreed). The last four questions all had $\alpha < 0.667$.

Languages

When a user searches for something on YouTube, shares a link to a video, or finds a video through any of the other discovery mechanisms, they are likely to do so using their native language or another language they understand well. Most people's experience of YouTube is shaped by these languages. Of course, why would someone watch a video they cannot understand, and why would a company that profits from maximizing user attention risk pointing someone to a video they are unlikely to watch? Language is a critical factor in determining a person’s perception of what YouTube is and the kinds of content it includes, and thus an important consideration for researchers trying to understand YouTube as a whole. Unfortunately, the distribution of spoken languages on YouTube is not public, and there is only so much we can glean from the metadata, which is incomplete when it comes to languages.

There is some existing work attempting to discover language statistics on YouTube. For example, the Pew Research Center analyzed videos from just the most popular channels (greater than 250,000 subscribers) during a one-week period of 2019. 56% of the 43,770 channels on the list published a video during the studied period, with a total of 243,254 videos. The researchers examined what proportion of videos included a language, in either text or audio, other than English and found that only 17% of videos were *only* in English while 72% of channels published one or more videos that contained a language other than English (Van Kessel, et al., 2019).

We believe many methods for collecting YouTube videos are likely to oversample English-language videos, either starting with English-language seed videos or operating data collection tools in countries where English is widely spoken. Our method allows us to estimate the linguistic distribution of videos on YouTube by detecting the primary language spoken in a video.

When retrieving caption data from YouTube, it distinguishes automatic captions from user-submitted captions and labels the original. The “automatic_captions” field could have a null value if YouTube does not have an automatic transcription for a given video, or will contain a subfield for each language that YouTube has an automatic transcript for. The language of the audio track is labeled with the suffix “-orig,” so a video in English that has an automatic transcription will have an “en-orig” subfield within the “automatic_captions” metadata field. The “subtitles” field could be null or contain at least one subtitle file provided by the video’s uploader. Apart from being limited to Dutch, English, French, German, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Turkish, and Vietnamese (YouTube Help, 2022), the necessary conditions for YouTube to produce automatic captions are unclear. Only 38.41% of videos in our sample include any captions, and there are several examples of obvious language detection errors.

To produce a language distribution for our sample that is as complete as possible, we built our own language detection pipeline by running each video's audio file using the VoxLingua107 ECAPA-TDNN spoken language recognition model (Valk & Alumnae, 2021). Although this method is significantly more computationally expensive than a relying on metadata, directly classifying the audio allows for the classification of YouTube videos without preexisting subtitle files, creates a consistent system for language classification that can be applied to any audio file, has the possibility of detecting a much wider range of languages, and removes the ambiguity associated with user-selected language values and YouTube's own approach to language identification.

We divided audio files into 30-second chunks and used VoxLingua to assign a confidence score to each chunk. After processing all of our audio files, we compared videos based on the 30-second block with the highest confidence in each. We found this to be a practical way to maximize confidence and standardize units of comparison. An examination of videos with low maximum confidence scores revealed some inaccuracies, due in part to the software assigning a language to instrumental music and ambient sounds, as well as inconsistent accuracy overall. Looking at the mean confidence across videos for each language, we found that Russian (0.97), English (0.95), French (0.95), Ukrainian (0.95), and Indonesian (0.95) had the highest mean confidence while Occitan (0.78), Scots (0.82), Interlingua (0.84), Cebuano (0.84), and Macedonian (0.84) had the lowest average confidence. Given the low confidence of many videos, we created a subset of our sample limited to videos with a top confidence score of 0.90 or greater. The full lists of languages by average confidence and the most commonly spoken languages with a confidence greater than or equal to 0.90 are in Appendix B. The distribution including the top 20 most common languages (confidence >= 0.90) is visualized in Figure 17.

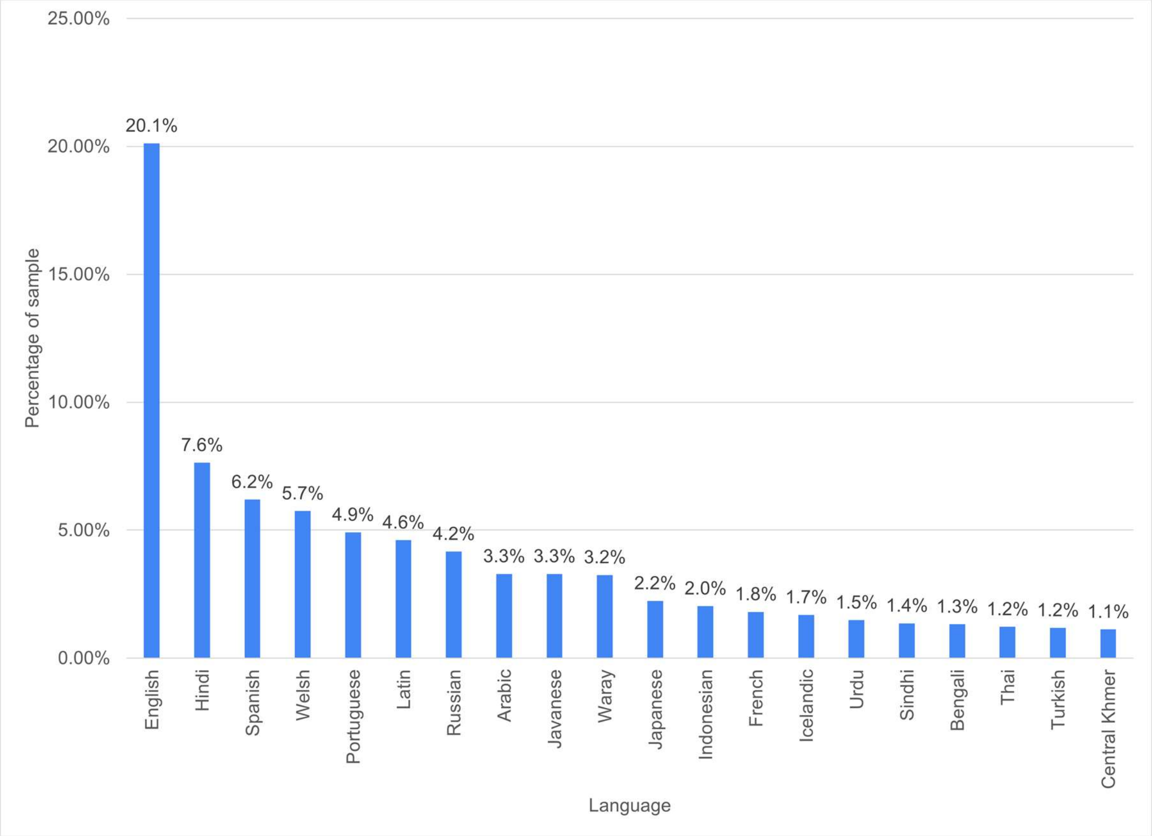


Figure 17. Language distribution (confidence >= 0.90), top 20.

Some unexpected languages should be immediately apparent. While many of the languages listed line up with our experience with the sample, others do not. It is unlikely that Welsh is the fourth most common language on YouTube, for example, or that Icelandic is spoken more often than Urdu, Bengali, or Turkish. More startling still is that, according to this analysis Latin is not a "dead language" but rather the sixth most common language spoken on YouTube. Of the top 20, Welsh, Latin, Waray-Waray, and Icelandic are not in the top 200 most spoken languages, and Sindhi and Central Khmer are not in the top 50 (Ethnologue, 2022). The VoxLingua107 documentation notes a number of languages which are commonly mistaken for another (Urdu for Hindi, Spanish for Galician, Norwegian for Nynorsk, Dutch for Afrikaans, English for Welsh, and Estonian for Finnish), but does not account for the other unusual results we have seen (Valk & Alumnae, 2021). We thought that some of the errors may be because of the amount of music in our sample, but removing the videos that are part of YouTube Music (which, as explained above, does not include all music) did not yield significantly different results (Figure 18).

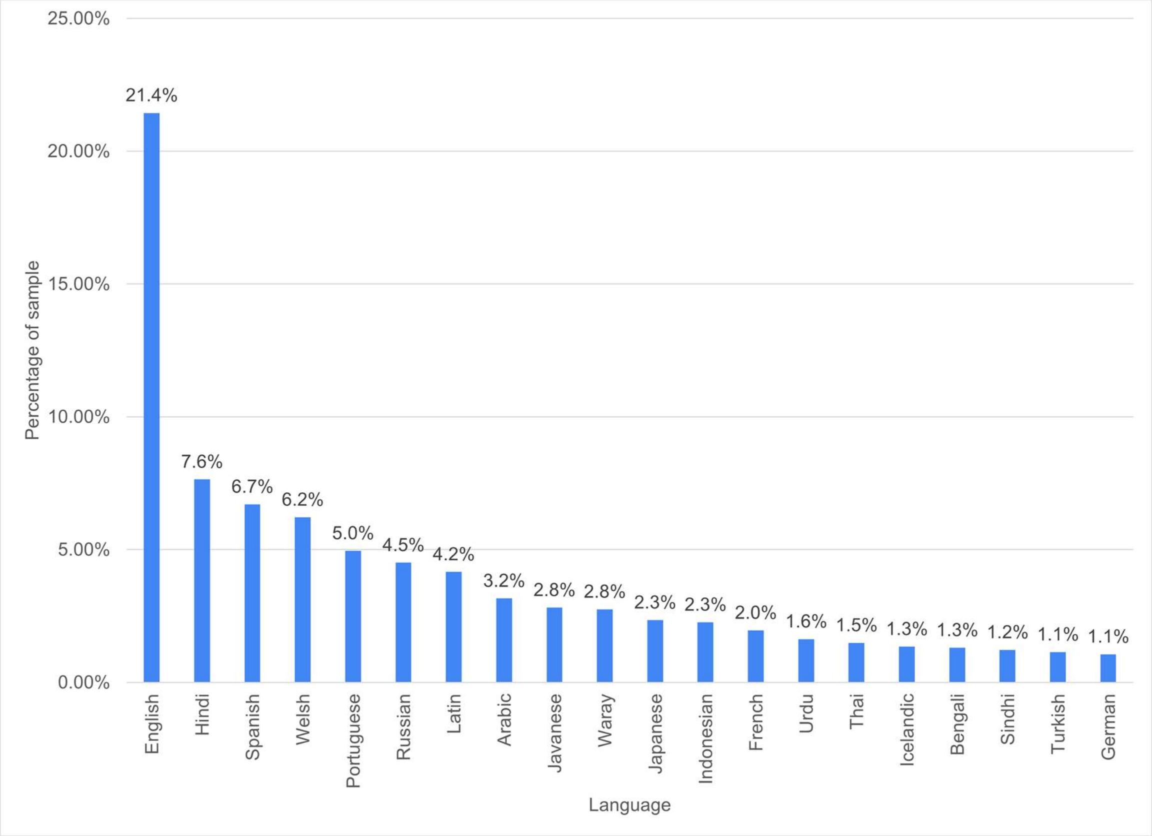


Figure 18. Language distribution (confidence >= 0.90, without "music"), top 20.

Even putting aside the unlikely prominence of languages like Welsh, Latin, and Icelandic, and the likelihood that some languages spoken by large populations are being mistaken for similar languages with a smaller number of speakers, it is worth highlighting just how many of the most popular languages are not among the languages available in the YouTube autocaptioning system: Hindi, Arabic, Javanese, Waray-Waray, Urdu, Thai, Bengali, and Sindhi.

We believe language detection is an essential part of researching YouTube. Knowing what language a video is in is a necessary sampling consideration and fundamental to processing linguistic content. We eventually want to be able to transcribe speech in YouTube videos at scale, and language detection is a fundamental first step. We have already begun experimenting with other software and other techniques to allow us to process language more effectively. While we are not confident in the results of our language distribution, we believe it provides a good indication of the extent to which YouTube, despite the experience of most users in English-speaking countries, is not predominantly English. To the contrary, while it may be the most common language, most of YouTube is not in English.

Comparison with Random Prefix Sampling

We believe our "Dialing for Videos" method is the best way to produce a truly random sample of videos within the constraints of the YouTube API and current computational capacity, but the amount of time and effort it took to perform makes it impractical for researchers to reproduce every time they want a random sample of their own. For that reason, we compared a newly produced "Random Prefix Sampling" collection of videos and compared it to our sample. As described above, Random Prefix Sampling takes advantage of a bug in the YouTube search function which provides videos with IDs starting with the search string, followed by a dash (Zhou, et al., 2011). While there is a risk of both overstudying this part of YouTube, as well as researchers influencing view counts, fundamental characteristics of the Random Prefix sample were similar to our Dialing for Videos sample.

Figure 19 compares the percentage of videos uploaded each year in the two samples. The most noticeable difference is due to when they were collected. Our sample was collected between October and December 2022 while the Random Prefix sample was collected on January 11, 2023. This accounts for our sample's zero value for 2023, significantly smaller value for 2022, and proportionally greater values before 2022. The full breakdown is in Table 2.

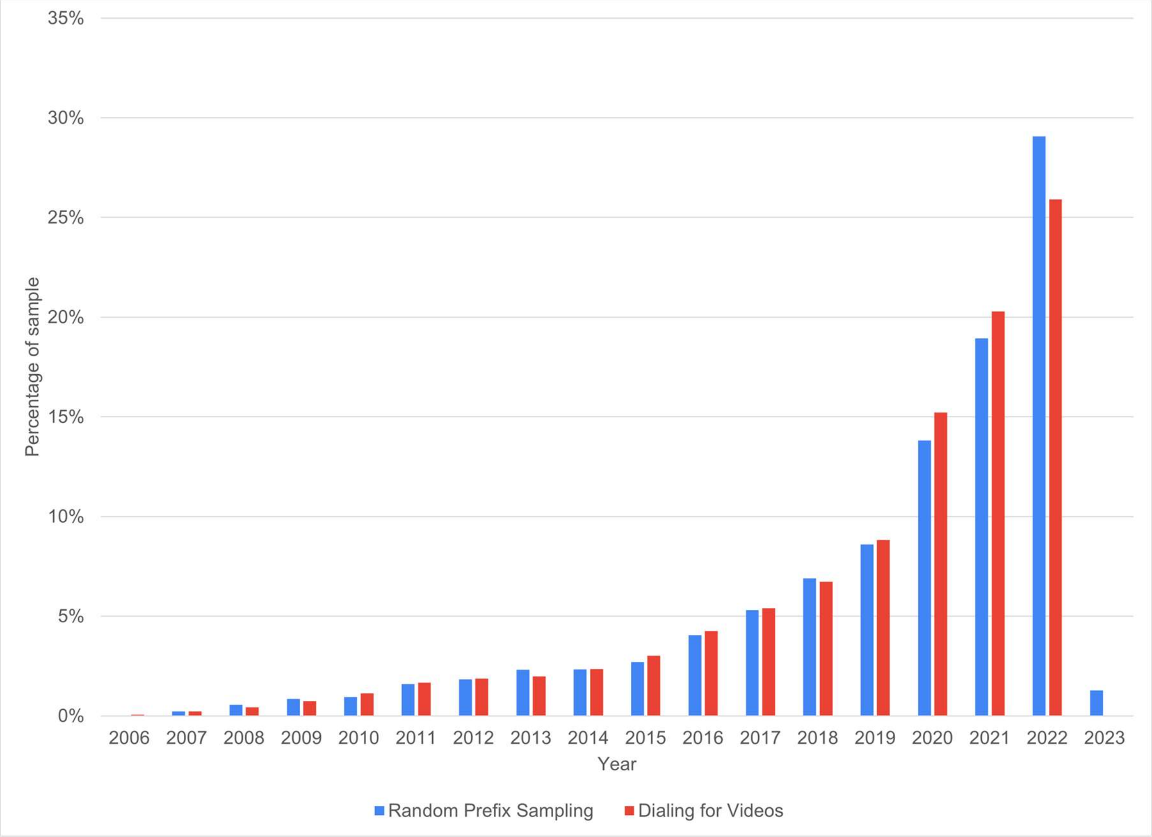


Figure 19. Comparison of Random Prefix and Dialing for Videos samples by percentage of videos uploaded each year.

Table 2. Comparison of Random Prefix and Dialing for Videos samples by percentage of videos uploaded each year.

Year	**Percentage of Random Prefix sample**	**Percentage of Dialing for Videos sample**
2006	0.02%	0.05%
2007	0.23%	0.22%
2008	0.56%	0.43%
2009	0.85%	0.74%
2010	0.95%	1.13%
2011	1.58%	1.67%
2012	1.83%	1.86%
2013	2.31%	1.97%
2014	2.32%	2.34%
2015	2.70%	3.02%
2016	4.04%	4.25%
2017	5.30%	5.39%
2018	6.89%	6.73%
2019	8.60%	8.81%
2020	13.81%	15.22%
2021	18.94%	20.29%
2022	29.07%	25.91%
2023	1.27%	0.00%

Key figures in the Random Prefix sample, alongside the same figures from our sample, are shown in Table 3**.**

Table 3. Comparison of key figures for Random Prefix and Dialing for Videos samples.

	Random Prefix sample	Dialing for Videos sample
Views: mean	15,900	5,868.02
Views: median	37	35
Views: minimum	0	0
Views: maximum	98,624,773	4,586,494
Views: zeros	4.75%	4.88%
Views: ones	4.12%	4.44%
Views: <100	65.31%	65.44%
Views: <1,000	86.45%	86.93%
Views: >10,000	3.42%	3.67%
Views: >10,000 as percentage of all views	97.44%	93.60%
Comments: mean	4.73	5.32
Comments: median	0	0
Comments: minimum	0	0
Comments: maximum	983	856
Comments: zeros	73.11%	72.64%
Comments: >100	0.99%	1.04%
Comments >100 as percentage of all comments	52.37%	54.60%
Likes: mean	2.83	16.48
Likes: median	0	0
Likes: minimum	0	0
Likes: Maximum	14,598	17,517
Likes: zeros	98.11%	88.71%
Subscribers: mean	61,991.27	55,100.04
Subscribers: median	64	61
Subscribers: minimum	0	0

Subscribers: maximum	92,700,000	33,100,000
Subscribers: >0	90.17%	90.52%
Unique channels	99.78%	99.81%

While the means and maximums of these figures indicate a difference in the size of outliers, the medians and distributions are very similar. The most noticeable difference is in the number of videos with no likes, with the Random Prefix sample including nearly 10% more. A comparison of log transformed view counts, comments, likes, and subscribers are in Figures 20, 21, 22, and 23.

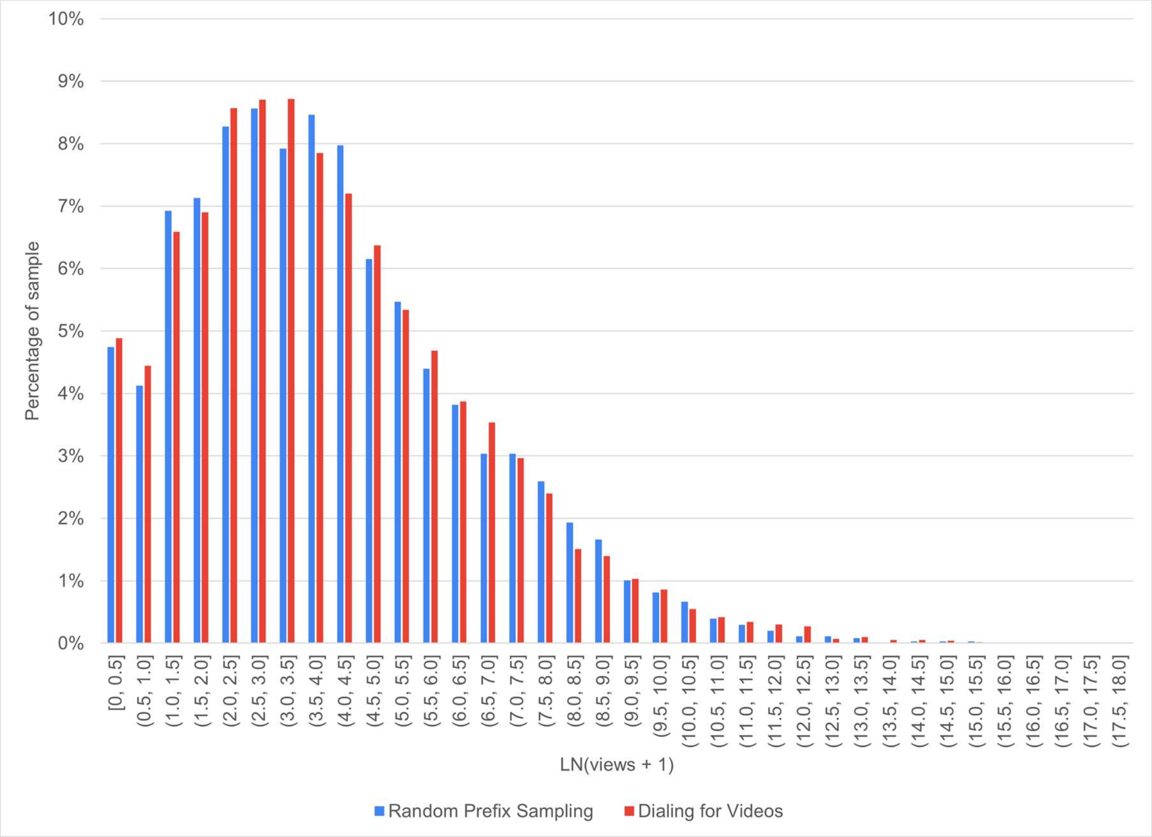


Figure 20. Comparison of log transformed view count distributions for Random Prefix and Dialing for Videos samples.

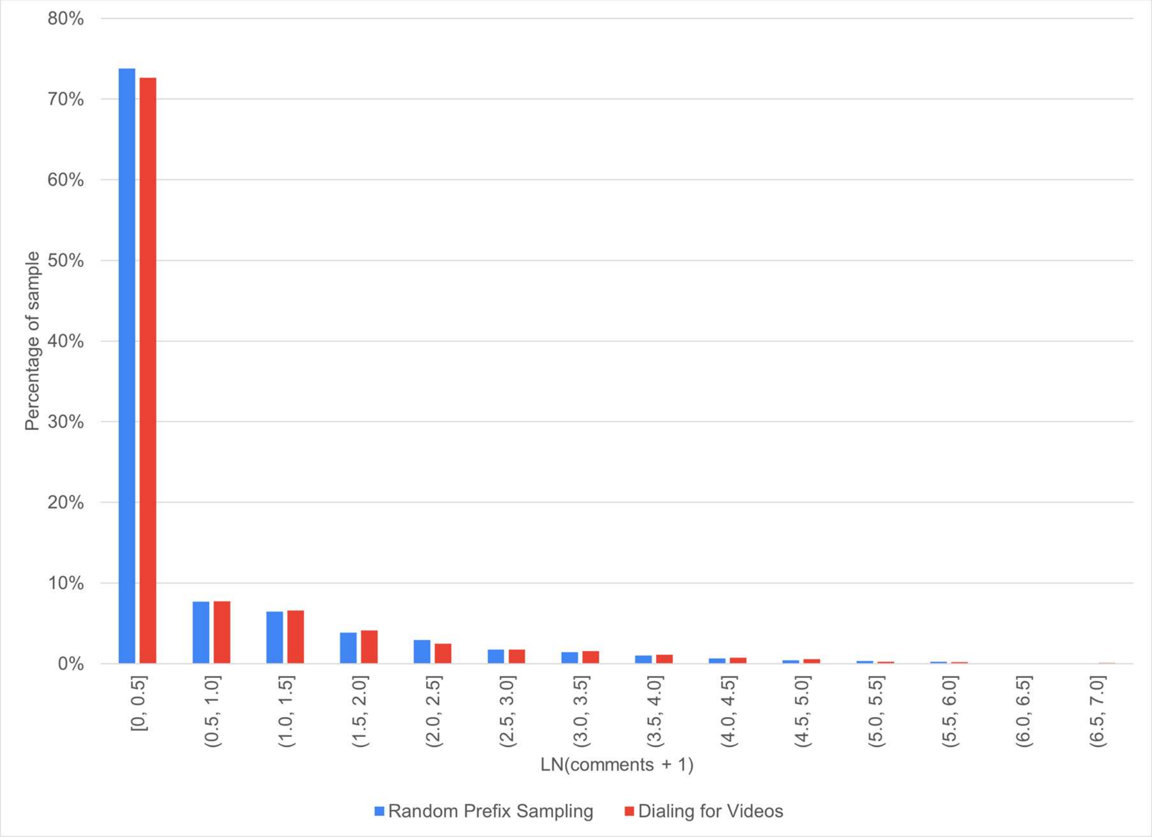


Figure 21. Comparison of log transformed comment count distributions for Random Prefix and Dialing for Videos samples.

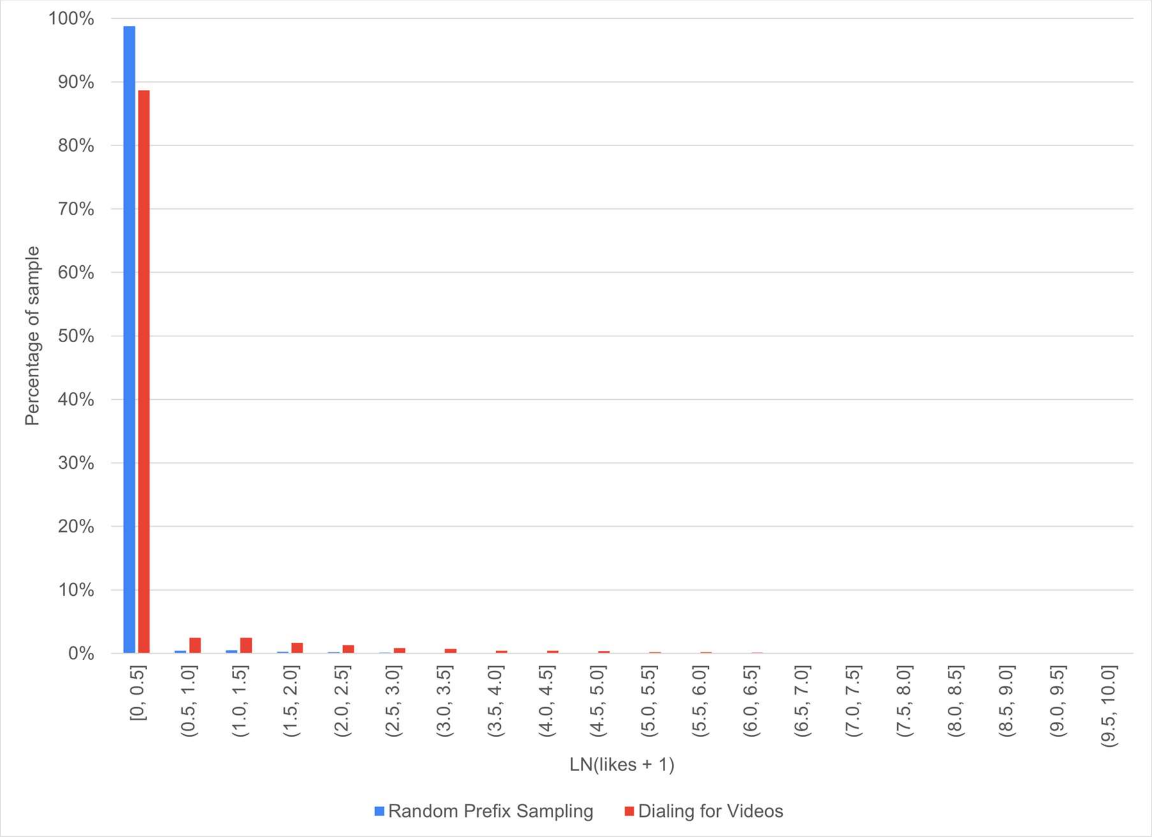


Figure 22. Comparison of log transformed like count distributions for Random Prefix and Dialing for Videos samples.

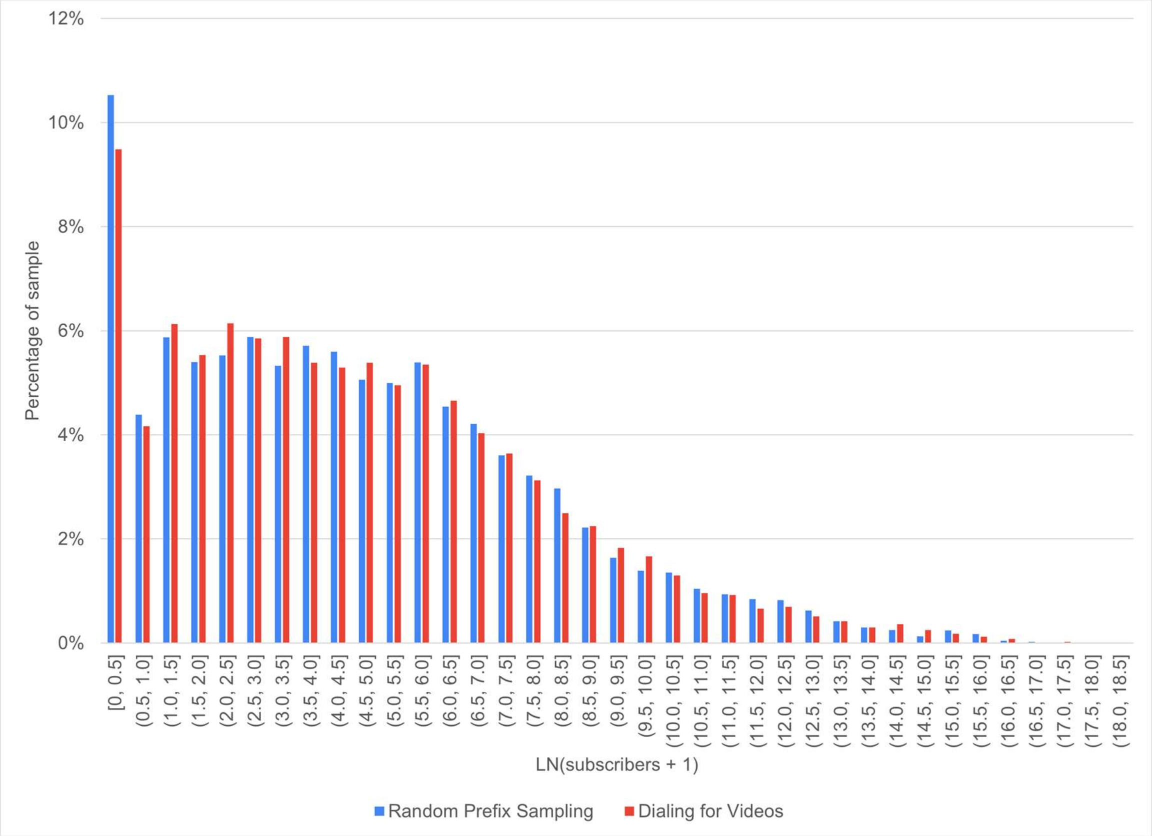


Figure 23. Comparison of log transformed subscriber count distributions for Random Prefix and Dialing for Videos samples.

Categorization was also similar between the two samples. Of note, in the time since we began collecting our Dialing for Videos sample, YouTube added a separate category for Shorts, which applies to a few videos in the Random Prefix. Table 4 lists the category breakdown, which is similar across samples.

Table 4. Comparison of categorization for Random Prefix and Dialing for Videos samples.

Category	Random Prefix sample	Dialing for Videos sample
Autos & Vehicles	1.25%	1.18%
Comedy	1.80%	1.82%

Education	3.17%	3.43%
Entertainment	7.65%	7.11%
Film & Animation	1.78%	1.95%
Gaming	12.14%	12.78%
Howto & Style	1.54%	1.47%
Music	6.23%	6.11%
News & Politics	2.51%	2.67%
Nonprofits & Activism	0.78%	0.97%
People & Blogs	56.42%	55.81%
Pets & Animals	0.79%	0.81%
Science & Technology	1.11%	0.97%
Shorts	0.01%	0.00%
Sports	1.96%	1.92%
Travel & Events	0.85%	1.01%

Finally, we used VoxLingua to analyze languages spoken in videos in the Random Prefix sample. We used the same method as the Dialing for Videos sample, which is subject to all of the same limitations described above. The results are very similar, with a comparison of the top 20 most spoken languages in each sample listed in Table 5.

Table 5. Comparison of twenty most spoken languages in the Random Prefix and Dialing for Videos samples.

Language	**Random Prefix sample**	**Dialing for Videos sample**
English	19.00%	20.12%
Hindi	7.45%	7.63%
Latin	5.76%	4.61%
Spanish	5.26%	6.19%
Welsh	5.24%	5.75%
Portuguese	4.81%	4.91%
Waray-Waray	3.81%	3.25%
Arabic	3.66%	3.29%
Russian	3.42%	4.16%
Javanese	3.28%	3.29%
Indonesian	2.10%	2.03%
Icelandic	1.92%	1.68%
Japanese	1.90%	2.23%
Sindhi	1.76%	1.36%
Urdu	1.54%	1.48%
French	1.45%	1.81%
Thai	1.36%	1.22%
Gujarti	1.16%	Not in top 20
Bengali	1.07%	1.32%
Sundanese	1.07%	Not in top 20
Turkish	Not in top 20	1.18%
Central Khmer	Not in top 20	1.12%

Although we have reservations about the Random Prefix Sampling described above, we believe it is sufficiently similar to our sample that it can be used as a more practical method to produce a random sample for most purposes. We strongly encourage researchers to use the dash method to generate their own pseudo-random YouTube samples, and we will make our random sample available to researchers on request so they can compare the characteristics of their video sample to a reference random sample. We will publish a subsequent paper outlining a method for comparison between a set of YouTube videos and various reference sets.

Discussion: What we can learn from a random sample of YouTube videos

YouTube is a profoundly important platform with an impact that reaches beyond the internet and into many dimensions of social and political life. It is also understudied, in part due to a lack of fundamental information about the site as a whole. This paper attempts to help researchers move beyond biased YouTube samples and to find answers to questions about what percentage of the total YouTube collection their videos are drawn from (denominator questions) and what subset of videos they represent (distribution questions), giving researchers context for their claims. To achieve our goal, we developed a method for producing a random sample of YouTube videos which we called Dialing for Videos and described what we found in an analysis of metadata, hand-coding, and a language detection pipeline.

Our approach to analyzing YouTube began with an informal hypothesis: that the YouTube most people have come to know is not representative of YouTube as a whole. YouTube is a multilingual space, with an enormous diversity of videos that generally get very few views. It is commonly an archive or alternative distribution channel for companies as well as individuals who cross-post their social media. To the extent that YouTube is a portrait of global popular culture - with obvious geographic biases - there are an awful lot of people playing games, and an awful lot of music. It has seen explosive growth in the last few years, and now totals almost 10 billion publicly visible videos. It is not primarily English, not predominantly professionalized, and is not easily mapped to a short list of categories.

Aspects of “YouTube culture”, notably the importance that prominent YouTube creators put on converting viewers into subscribers, likely characterize only a tiny fraction of extremely popular videos, like the 16 in our set that accounted for more than 50% of views. That tiny subset of videos may be focused on monetization, increasing subscriber counts, and urging viewers to “smash that like button,” but the vast majority of videos we encountered do not seem to participate in this “creator economy.” This suggests that there is “another” YouTube – likely many other YouTubes – where its use to store video or share content with a small audience (from classroom assignments to video greetings for friends to simulcast religious services) is far more common than creating videos seeking an audience of millions. We hope our work can help other scholars begin to explore these different YouTubes and understand the complexity of a system used by many different groups of users for disparate reasons.

YouTube is many things, and we hope that by answering many of the questions about YouTube’s fundamentals, we can help to provide context for researchers to dig deeper into more specific parts of the site. We are also excited about the research this will allow us to do into YouTube content. Along the way, we have also identified a number of areas that are perhaps areas that researchers should be cautious about relying on. YouTube’s category system, for example, is too simple, incomplete, and inconsistently applied to be useful for most research, especially due to a heavy skew towards the default category. YouTube’s captions are only available for certain languages, with a number of errors, although they appear to be better than many other sites’ approaches to the problem of captions and transcription. Perhaps most importantly, we have shown that any study that relies on a sample of popular or English-language videos should not be generalized to make claims about YouTube as a whole. We believe a distribution of view counts, languages, upload dates, and duration may go a long way to providing a sample’s “fingerprint,” and a good practice for any study using a non-random sample may be to compare those distributions to a random sample.

In addition to the limitations described in the section above, it is worth highlighting two. First, it is likely that our language detection system is misidentifying a significant number of videos, characterizing them as languages that are unlikely to appear on YouTube. As a result, our findings are more useful in understanding the wide range of languages spoken and their relative prevalence on YouTube than in making an accurate estimate of any specific language's size. We hope to improve on our work here by integrating OpenAI's Whisper system to transcribe content in 99 languages (OpenAI, n.d.). Second, because our sample is relatively small, and because very popular YouTube videos are quite uncommon, our description of the characteristics of YouTube videos likely does not well describe the characteristics of very popular videos. It is likely that our sampling method includes a great deal of variance in the summary statistics of very popular videos due to sensitivity to the particularities of a small number of very popular videos.

Conclusion

We developed a method, Dialing for Videos, to generate a random sample of YouTube videos and used it to produce a set of 10,016 videos. We retrieved and analyzed the metadata, downloaded the audio track and ran it through a language detection pipeline, and hand-coded a subset of 1,000 videos, answering 32 questions about aspects of the videos' content, including audio and video characteristics.

Extrapolating from our sampling methods, we can offer an estimate of the size of YouTube (9,881,141,822 publicly searchable videos at the time of our late-2022 sampling window), a growth curve for the number of videos hosted by the site (46.20% uploaded in 2021 and 2022) and descriptive statistics for video views, length, likes, comments and channel subscriptions. The "long tail" of YouTube is very, very long: the 16 most popular videos in our set are responsible for more than half of all views. The vast majority of YouTube videos (86.93%) are seen by fewer than 1,000 people. A significant number (4.88%) are never watched at all, and most videos have no comments (72.64%) and no likes (88.71%). Some of the research questions we asked – about the professionalism of

YouTube creators and how often they monetize their videos – are simply not applicable to most videos in a sample like this.

Rather than finding evidence that most YouTube users wanted to be part of a creator economy, we found surprising and unusual uses for the platform:

- Just over 40% (40.56%) of videos featured music and no speech (and only 34.2% of that subset were in the YouTube Music database).
- Just over 15% (15.54%) were primarily still images.
- Just under 20% (19.52%) of videos were people playing video games.
- Almost 4% (3.84%) of videos were religious in nature, with about 3% (2.99%) focused on news and current events.

YouTube's metadata is often unhelpful in understanding this obscure long tail of videos. Only 38.41% have captions available, and YouTube only transcribes in 13 languages, which led us to develop our own pipeline for language identification. User-entered categories in the metadata have very little alignment with categories human coders sorted the videos into, with the default category, People & Blogs, including more videos than all other categories combined (55.82%).

We compared our random sampling method, Dialing for Videos, to an older method that relies on autocompletion behavior, Random Prefix Sampling. The similarity between these randomly generated sets suggests that the Random Prefix method, which is much faster to implement, should be used by scholars generating random YouTube video sets, while being mindful of its opacity and potential for overstudying the subset of total videos available to it.

Our examination of "ordinary" long-tail videos suggests that there is an enormous scholarly opportunity to examine less popular YouTube videos as a way of understanding cultural production and different uses for shared video beyond the most obvious and popular uses. We encourage researchers to use the tools we have shared here to both investigate these less-explored – though numerically dominant – corners of YouTube and to provide additional context for the samples of YouTube videos used to study phenomena like hate speech and mis- and disinformation.

References

Arthurs, J., Drakopoulou, S., & Gandini, A. (January 10, 2018). Researching YouTube. *Convergence: The International Journal of Research into New Media Technologies, 24*(1).

Barrett, P. M. & Hendrix, J. (June 2022). A Platform 'Weaponized': How YouTube Spreads Harmful Content–And What Can Be Done About It. *NYU Stern Center for Business and Human Rights*. https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/62a38fc022745a7274601da0/1654886337000/NYU+CBHR+YouTube_Final_June10.pdf

Bartl, M. (2018). YouTube channels, uploads and views: A statistical analysis of the past 10 years. *Convergence, 24*(1): 16-32.

Biggs, J. (February 20, 2006). A Video Clip Goes Viral, and a TV Network Wants to Control It. *The New York Times.* Retrieved January 25, 2023 from https://www.nytimes.com/2006/02/20/business/media/a-video-clip-goes-viral-and-a-tv-network-wants-to-control-it.html

Brouwer, B. (July 26, 2015). YouTube Now Gets Over 400 Hours of Content Uploaded Every Minute. *Tubefilter*. Retrieved January 25, 2023 from https://www.tubefilter.com/2015/07/26/youtube-400-hours-content-every-minute/

Bryant, L. V. (2020). The YouTube Algorithm and the Alt-Right Filter Bubble. *Open Information Science, 4*(1): 85-90.

Burgess, J. & Green, J. (2018). *YouTube: Online Video and Participatory Culture*. Cambridge: Polity Press.

Cayari, C. (July 8, 2011). The YouTube Effect: How YouTube Has Provided New Ways to Consume, Create, and Share Music. *International Journal of Education & the Arts, 12*(6). Retrieved January 23, 2023 from http://www.ijea.org/v12n6/

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., & Moon, S. (2007). I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. *IMC'07*. Doi: 10.1145/1298306.1298309

Chen, A. Y., Nyhan, B., Reifler, J., Robertson, R. E., & Wilson, C. (2021). Exposure to Alternative & Extremist Content on YouTube. *Anti-Defamation League*.

De Saa, E. & Ranathunga, L. (2020). Self-Reflective and Introspective Feature Model for Hate Content Detection in Sinhala YouTube Videos. *2020 From Innovation to Impact (FITI)*, Colombo, Sri Lanka.

Duffy, P. (2008). Engaging the YouTube Google-Eyed Generation: Strategies for Using Web 2.0 in Teaching and Learning. *The Electronic Journal of e-Learning, 6*(2): 119-130.

Ethnologue. (2022). What are the top 200 most spoken languages? *Ethnologue*. Retrieved May 13, 2023 from https://www.ethnologue.com/insights/ethnologue200/

Fisher, K. (March 29, 2006). YouTube caps video lengths to reduce infringement. *Ars Technica.* Retrieved January 25, 2023 from https://arstechnica.com/uncategorized/2006/03/6481-2/

Fisher, M. & Bennhold, K. (2018, September 7). As Germans Seek News, YouTube Delivers Far-Right Tirades. *The New York Times.* Retrieved April 26, 2023 from https://www.nytimes.com/2018/09/07/world/europe/youtube-far-right-extremism.html

Hatmaker, T. (June 30, 2022). YouTube will disable hidden subscriber counts to fight comment spam. *TechCrunch.* Retrieved March 1, 2023 from https://techcrunch.com/2022/06/30/youtube-subscriber-count-impersonator-spam-special-characters/

Hern, A. (September 15, 2020). YouTube Shorts launches in India after Delhi TikTok ban. *The Guardian*. Retrieved March 8, 2023 from https://www.theguardian.com/technology/2020/sep/15/youtube-shorts-launches-in-india-after-delhi-tiktok-ban

Hráček, F. (2009). *Audiovisual Style of User-Generated YouTube Videos* [Master's thesis, Masaryk University].

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring Misinformation in Video Search Platforms: An Audin Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1): 1-27.

Jung, S., Salminen, J., & Jansen, B. J. (2022). The Effect of Hiding Dislikes on the Use of YouTube's Like and Dislike Features. *WebSci '22*. Doi: 10.1145/3501247.3531546

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

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Second Edition. Thousand Oaks, CA: Sage.

Ledwich, M. & Zaitsev, A. (2020, March 2). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday, 25*(3). https://firstmonday.org/ojs/index.php/fm/article/view/10419/9404

Lewis, P. (2018, February 2). 'Fiction is outperforming reality': how YouTube's algorithm distorts truth. *The Guardian*. Retrieved April 26, 2023 from https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth

Lewis, R. (2018, September 18). Alternative Influence: Broadcasting the Reactionary Right on YouTube. *Data & Society*. Retrieved April 26, 2023 from https://datasociety.net/library/alternative-influence/

Li, H. O., Bailey, A., Huynh, D., & Chan, J. (2020). YouTube as a source of information on COVID-19: a pandemic of misinformation? *BMJ Global Health.*

Martin, R. (April 13, 2021). Exploring YouTube and the Spread of Disinformation. *Morning Edition*. Retrieved January 23, 2023 from https://www.npr.org/2021/04/13/986678544/exploring-youtube-and-the-spread-of-disinformation

McCrosky, J. & Geurkink, B. (2021). *YouTube Regrets: A Crowdsourced Investigation into YouTube's Recommendation Algorithm*. Mozilla. Retrieved December 25, 2022, from https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf

Mowlabocus, S. (2018). 'Let's get this thing open': The pleasures of unboxing videos. *Convergence, 23*(4): 564-579.

Munn, L. (2019, June 3). Alt-right pipeline: Individual journeys to extremism online. *First Monday, 24*(6). Retrieved April 26, 2023 from https://firstmonday.org/ojs/index.php/fm/article/view/10108/7920

Nashmi, E. A., North, M., Bloom, T., & Cleary, J. (2017). 'Boots on the Ground?': How international news channels incorporate user-generated content into their YouTube presence. *The International Communication Gazette, 79*(8): 746-768.

OpenAI. (n.d.). Openai/whisper-large-v2. *Hugging Face*. Retrieved May 13, 2023 from https://huggingface.co/openai/whisper-large-v2

Ørmen, J. & Gregersen, A. (2022). Towards the engagement economy: interconnected processes of commodification on YouTube. *Media, Culture & Society*.

Paolillo, J. C. (2008). Structure and Network in the YouTube Core. *Proceedings of the 41st Hawaii International Conference on System Sciences.* Doi: 10.1109/HICSS.2008.415

Reuters. (July 16, 2006). YouTube Serves Up 100 Million Videos a Day. *NBC News.* Retrieved 26 December, 2022 from https://www.nbcnews.com/id/wbna13890520

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2021). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 131-141).

Rieder, B., Matamoros-Fernandez, A., & Coromina, O. (2018). From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results. *Convergence: The International Journal of Research into New Media Technologies, 24*(1): 50-68.

Rodriguez, S. (April 7, 2021). YouTube is social media's big winner during the pandemic. *CNBC.* Retrieved December 25, 2022 from https://www.cnbc.com/2021/04/07/youtube-is-social-medias-big-winner-during-the-pandemic.html

Salgar, S. (May 15, 2013). WE'LL DO IT LIVE: YouTube live streaming expanding to more channels. *YouTube Official Blog.* Retrieved January 24, 2023 from https://blog.youtube/news-and-events/well-do-it-live-youtube-live-streaming/

Salgar, S. (December 12, 2013). Now you can live stream on YouTube. *Official YouTube Blog.* Retrieved January 24, 2023 from https://blog.youtube/news-and-events/now-you-can-live-stream-on-youtube/

Salsman, J. E. (January 10, 2018). YouTube's AI Is the Puppet Master Over Most of What You Watch. *Cnet*. Retrieved December 25, 2022, from https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan/

Selinger, E. & Hartzog, W. (2018). Obscurity and Privacy. In Pitt, J. C. & Shew, A. (Eds.). *Spaces for the Future: A Companion to Philosophy of Technology.* New York: Routledge.

Siegel, J. (July 29, 2010). Upload limit increases to 15 minutes for all users. *YouTube Official Blog.* Retrieved January 25, 2023 from https://blog.youtube/news-and-events/upload-limit-increases-to-15-minutes/

Siegler, M. G. (April 28, 2009). Hulu Now the Number Three U.S. Web Video Site. Soon to Be Number Two. *TechCrunch*. Retrieved December 24, 2022 from https://techcrunch.com/2009/04/28/as-youtube-passes-a-billion-unique-us-viewers-hulu-rushes-into-third-place

Smith, T., Obrist, M., & Wright, P. (June 2013). Live-streaming changes the (video) game. *EuroITV '13: Proceedings of the 11th European Conference on Interactive TV and Video*. Doi: 10.1145/2465958.2465971

Snickars, P. & Vonderau, P. (Eds.). (2010). *The YouTube Reader.* New York: Columbia University Press.

Staff. (April 21, 2020). YouTube sees surge in subscriber base, views due to Covid-19 lockdown. *Business Standard.* Retrieved December 25, 2022 from https://www.business-standard.com/article/technology/youtube-sees-surge-in-subscriber-base-views-due-to-covid-19-lockdown-120042100710_1.html

Tolson, A. (2010). A new authenticity? Communicative practices on YouTube. *Critical Discourse Studies, 7*(4): 277-289.

"Top Websites Ranking." (2022). *SimilarWeb*. Retrieved December 25, 2022 from https://www.similarweb.com/top-websites/

Tufekci, Z. (2018, March 10). YouTube, the Great Radicalizer. *The New York Times.* Retrieved April 26, 2023 from https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Valk, J. & Alumnae, T. (2021). VoxLingua107: A Dataset for Spoken Language Recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)* (pp. 652-685). Retrieved April 4, 2023 from https://arxiv.org/pdf/2011.12998.pdf

Van Kessel, P., Toor, S., & Smith, A. (July 25, 2019). A Week in the Life of Popular

YouTube Channels. *Pew Research Center.* Retrieved April 4, 2023 from https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/07/DL_2019.07.25_YouTube-Channels_FINAL.pdf

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

Wesch, M. (March 17, 2008). YouTube Statistics. *Digital Ethnography @ KSU.* Retrieved December 26, 2022 from https://web.archive.org/web/20130215021954/http://ksudigg.wetpaint.com/page/YouTube+Statistics

YouTube Team. (November 18, 2020). Updates to YouTube's Terms of Service. *YouTube Official Blog.* Retrieved April 6, 2023, from https://blog.youtube/news-and-events/updates-to-youtubes-terms-of-service/

YouTube. (n.d.). Create a live stream on mobile. *YouTube Help.* Retrieved January 24, 2023 from https://support.google.com/youtube/answer/9228390

YouTube. (n.d.). Use automatic captioning. *YouTube Help.* Retrieved May 13, 2023 from https://support.google.com/youtube/answer/6373554

YouTube Team. (January 23, 2012). Holy Nyans! 60 hours per minute and 4 billion views a day on YouTube. *YouTube Official Blog*. Retrieved January 25, 2023 from https://blog.youtube/news-and-events/holy-nyans-60-hours-per-minute-and-4/

Zhou, J., Li, Y., Adhikari, V. K., & Zhang, Z. (2011). Counting YouTube Videos via Random Prefix Sampling. *IMC'11: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement*: 371-380.

Zuckerman, E. (November 2, 2021). Facebook has a misinformation problem, and is blocking access to data about how much there is and who is affected. *The Conversation*. Retrieved January 25, 2023 from https://theconversation.com/facebook-has-a-misinformation-problem-and-is-

blocking-access-to-data-about-how-much-there-is-and-who-is-affected-164838

Appendixes

Appendix A: Survey questions, instructions, agreement, and reliability

All questions, possible answers, and provided instructions are listed below. After coding their first few videos, coders were invited to ask questions about the codebook or process. Answers to these questions were gathered in a shared frequently asked questions document (FAQ) which all coders were then asked to read before continuing. The answers to these questions are presented below the relevant instructions, indicated by "FAQ."

Table 6. Questions, possible answers, and instructions provided to hand coders.

Question	Possible answers	Instructions
Does the video exist?	Yes No	In some rare cases, the video may have been taken down for some reason. The list was generated recently, so please double-check that you've copied the video id/url correctly before marking "no". Marking "no" will immediately take you to the end of this form. FAQ: In the rare case you come across a video marked private, which should only happen if the uploader switched its privacy setting from public to private since we collected the data, treat it as though it does not exist.
Is there any audio at all?	Yes No	Only choose "no" if no audio track exists. If there is ambient sound, static, faint music, etc., still choose "yes".
Is the video entirely still images (or slightly modified still images)?	Yes No	"Slightly modified still images" includes sparkle filters, images with text overlay, image zooms, images which move across the screen ("the Ken Burns effect"), and similar effects.

		FAQ: As this section only concerns the visual portion of the video, it does not matter if there is audio content. In other words, a slideshow is still “entirely still images” even if there is narration or background music.
Is the video entirely (or almost entirely) a video game?	Yes No	FAQ: If there is video overlaying the video game, such as streaming video of the person playing the game, select “no.”
Is the video entirely (or almost entirely) graphical (other than video games)?	Yes No	Phone screens, desktops, websites, custom graphics, etc. If the graphics generally include photos or video, choose "no".
Does the video contain edits?	Yes No	This question concerns edits in the sense of combining, adding, or removing video clips. For example, does it have cuts, fades, dissolves, or transitions? FAQ: A graphical introduction does count as an edit because of the transition from the introduction to the rest of the video. FAQ: Edits within a video game, such as when a player moves from one level to another, are not the kind of edits we are looking for. This question is attempting to gauge an aspect of visual production.
Is there apparent set/background design?	Yes No	Choose yes only if it looks like effort was put into a set or someone's background. If there is no background/foreground distinction (e.g. if it's entirely graphical), choose "no". If you are unsure, choose "no". FAQ: Professional visual arts productions like a film trailer or stage performance will typically have some

		attention to the background, and should not be excluded.
Does the video include a custom graphical introduction?	Yes No	A still image or video which appears at the beginning or near the beginning of a video, introducing the video or channel, visibly distinct from the rest of the video. If you are unsure, select "no".
Does the video include any text, graphics, or other video overlaying the main video?	Yes No	This specifically concerns overlays, not simply whether there is textual/graphical content. Overlays place text or graphics on top of the primary video footage.
Does the video include a human talking to the camera?	Yes No	If the person talking is using filters or avatars to mask their appearance, still answer "yes". If you are unsure, answer "no". FAQ: This question, and others in this section, are only about the visual portion of the video. As such, voice-over alone does not constitute "human talking to the camera."
If there is recorded video, is it shot primarily indoors or primarily outdoors?	Yes No Unclear N/A	"Recorded video" here means live/in-person video as opposed to video that is entirely textual, entirely graphical/animated, or entirely still images. If it is textual, graphical, still images, or otherwise does not have recorded video, select "N/A". Use "unclear" when there is recorded video but it is unclear if it's indoors or outdoors.
If there is recorded video, is the video noticeably shaky?	Yes No N/A	This question is intended to be about production, with use of phones or inexpensive handheld cameras vs. mounted cameras or more specialized equipment. Recorded video here means live/in-person video as opposed to video that is entirely textual, entirely

		graphical/animated, or entirely still images. If it is textual, graphical, still images, or otherwise does not have recorded video, select "N/A".
Does the video include footage of a public or semi-public event?	Yes No	If you're unsure, choose "no". FAQ: This aims to include events like concerts, church services, town halls, and other situations where people gather for a specific, planned purpose. Something does not qualify as being an “event” in this context just by virtue of being public or semi-public (like a person filming themselves dancing in a mall). FAQ: A video of an in-person classroom lecture which would take place even if it were not recorded should count as an event. If students record themselves having fun in the back of a classroom, the fact that it is a classroom does not mean the recording depicts an event.
Is the video obviously the work of someone else?	Yes No	Movies, music videos, television, video montages of clips likely found on TikTok/YouTube/Instagram, etc., published on a channel that is unlikely to own those clips. If you're unsure, choose "no".
Did you see a YouTube video ad?	Yes No	
Did you see a YouTube text-based ad (such as a banner)?	Yes No	This is about text-based ads outside of video ads (sometimes a text-based banner/pop-up will appear on top of a video ad -- that is not what we are looking for with this question).
Is there evidence of in-video monetization?	Yes No	This is monetization by the creator/uploader, such as sponsorships, promotion, or advertisements. Do not consider YouTube ads in answering

		this question. Sometimes YouTube indicates the inclusion of paid promotion a little box on top of the video that says "includes paid promotion". If you see that box, select "yes". If you are unsure, choose "no". FAQ: The distinction between monetization and advertisements is between people trying to make money from their YouTube channels/videos and people who use YouTube to distribute advertising for a product/service outside of YouTube. The second group may or may not also be trying to make money from the video, too, but that's not the primary intention. There will be some gray area here, and if you're unsure, choose "no."
Is there a direct call to action (other than advertising)?	Yes No	This is primarily about engagement with the video, channel, or YouTuber. For example, an invitation to like, subscribe, leave a comment, visit the channel's website, visit another social media account, or visit Patreon or other funding site. If you're unsure, choose "no".
Does the video obviously serve as an advertisement for a product or service other than the video/channel itself?	Yes No	This is about products or services outside of YouTube. If someone creates a video to promote their car dealership, uploads a television advertisement, or shows off their custom jewelry for sale, they are advertising through YouTube (as opposed to monetizing YouTube videos). If you're unsure, select "no". FAQ: A typical film trailer is a form of advertising. A fan-made trailer is not. A straightforward clip from a movie or television show typically will not be an

		advertisement unless it also contains promotional material.
Which of the YouTube categories best fits?	Autos and vehicles Comedy Education Entertainment Film and animation Gaming Howto and style Music News and politics Nonprofits and activism Pets and animals People and blogs Science and technology Sports Travel and events	There will be many videos that do not fit cleanly into one of these categories. Choose what feels closest. Base your choice only on the video, and not on any other knowledge about the channel/uploader.
Is there anything related to politics, news, or current world events?	Yes No	Do not include "pop culture" news like entertainment or sports unless there is a political element. If you're unsure, choose "no".
Is there anything related to religion?	Yes No	Church services, televangelism, religious music, etc. If you're unsure, choose "no".

Table 7. Intercoder agreement and reliability.

Question	**Percent Agreement**	**Krippen-dorff's Alpha**	**N Agreements**	**N Disagree-ments**	**N Cases**
Does the video exist?	97.98	0.73	969	20	989
Is there any audio at all?	99.36	0.89	934	6	940

Is the video entirely still images (or slightly modified still images)?	95.85	0.85	901	39	940
Is the video entirely (or almost entirely) a video game?	98.09	0.94	922	18	940
Is the video entirely (or almost entirely) graphical (other than video games)?	91.70	0.60	862	78	940
Does the video contain edits?	84.26	0.67	792	148	940
Is there apparent set/background design?	87.87	0.60	826	114	940
Does the video include a custom graphical introduction?	91.49	0.69	860	80	940
Does the video include any text, graphics, or other video overlaying the main video?	80.32	0.59	755	185	940
Does the video include a human talking to the	91.17	0.74	857	83	940

camera?					
If there is recorded video, is it shot primarily indoors or primarily outdoors?	87.98	0.82	827	113	940
If there is recorded video, is the video noticeably shaky?	84.47	0.76	794	146	940
Does the video include footage of a public or semi-public event?	95.53	0.77	898	42	940
Is the video obviously the work of someone else?	95.11	0.63	894	46	940
Did you see a YouTube video ad?	93.30	0.65	877	63	940
Did you see a YouTube text-based ad (such as a banner)?	91.91	0.34	864	76	940
Is there evidence of in-video monetization?	97.23	0.12	914	26	940

Is there a direct call to action (other than advertising)?	86.49	0.29	813	127	940
Does the video obviously serve as an advertisement for a product or service other than the video/channel itself?	94.15	0.48	885	55	940
Which of the YouTube categories best fits?	67.45	0.63	634	306	940
Is there anything related to politics, news, or current world events?	96.06	0.57	903	37	940
Is there anything related to religion?	96.91	0.69	911	29	940
Is the audio entirely (or almost entirely) music?	93.74	0.87	853	57	910
Is the audio obviously the work of someone else?	85.60	0.68	779	131	910

Is there spoken language in the first minute?	95.05	0.90	865	45	910
Is at least some of the spoken language on top of music?	83.69	0.68	390	76	466
Is there text-to-speech (computer-generated) language?	97.42	-0.01	454	12	466
Is there at least one instance of multiple concurrent speakers for more than two seconds?	84.98	0.47	396	70	466
Rate the audio quality (on the whole).	58.80	0.33	274	192	466
Does the audio quality vary significantly?	84.98	0.32	396	70	466

Appendix B: VoxLingua language data

Table 8. Most frequently detected languages with confidence of at least 0.90 and the top languages by average confidence.

Language	**Percentage of videos with confidence**	**Language**	**Average confidence**

	>= 0.90		
English	20.12%	Russian	0.97
Hindi	7.63%	English	0.95
Spanish	6.19%	French	0.95
Welsh	5.75%	Ukrainian	0.95
Portuguese	4.91%	Indonesian	0.95
Latin	4.61%	Kazakh	0.95
Russian	4.16%	Spanish	0.95
Arabic	3.29%	Hebrew	0.94
Javanese	3.29%	Urdu	0.94
Waray-Waray	3.25%	Czech	0.94
Japanese	2.23%	Hindi	0.94
Indonesian	2.03%	German	0.93
French	1.81%	Tagalog	0.93
Icelandic	1.68%	Tamil	0.93
Urdu	1.48%	Arabic	0.92
Sindhi	1.36%	Bosnian	0.92
Bengali	1.32%	Portuguese	0.92
Thai	1.22%	Thai	0.92
Turkish	1.18%	Latin	0.92
Central Khmer	1.12%	Lao	0.92
Telugu	1.08%	Japanese	0.92
Vietnamese	0.99%	Georgian	0.92
German	0.97%	Galician	0.92
Gujarati	0.95%	Bulgarian	0.91
Tagalog	0.91%	Vietnamese	0.91
Tamil	0.89%	Belarusian	0.91
Sundanese	0.89%	Slovak	0.91
Korean	0.75%	Gujarati	0.91

Sanskrit	0.69%	Azerbaijani	0.90
Maori	0.69%	Turkish	0.90
Chinese	0.63%	Malayalam	0.90
Malay	0.61%	Luxembourgish	0.90
Panjabi	0.59%	Dutch	0.90
Italian	0.55%	Telugu	0.90
Malayalam	0.47%	Welsh	0.90
Yoruba	0.47%	Bengali	0.90
Burmese	0.45%	Yoruba	0.90
Marathi	0.41%	Panjabi	0.90
Polish	0.41%	Polish	0.90
Norwegian Nynorsk	0.39%	Marathi	0.90
Galician	0.37%	Italian	0.90
Dutch	0.37%	Greek	0.89
Belarusian	0.35%	Hungarian	0.89
Breton	0.30%	Chinese	0.89
Kannada	0.26%	Nepali	0.89
Yiddish	0.26%	Javanese	0.89
Romanian	0.24%	Kannada	0.89
Shona	0.24%	Danish	0.89
Nepali	0.24%	Korean	0.89
Basque	0.24%	Assamese	0.89
Hawaiian	0.22%	Central Khmer	0.88
Lao	0.22%	Pushto	0.88
Ukrainian	0.20%	Sundanese	0.88
Sinhala	0.20%	Sindhi	0.88
Swahili	0.18%	Estonian	0.88
Assamese	0.18%	Icelandic	0.88
Norwegian	0.18%	Shona	0.87

Azerbaijani	0.16%	Swahili	0.87
Faroese	0.16%	Sinhala	0.87
Hungarian	0.16%	Lithuanian	0.87
Guarani	0.16%	Tibetan	0.87
Tibetan	0.16%	Romanian	0.87
Danish	0.16%	Faroese	0.87
Hausa	0.14%	Waray-Waray	0.87
Kazakh	0.12%	Maori	0.87
Bosnian	0.12%	Malay	0.87
Persian	0.12%	Finnish	0.87
Somali	0.12%	Tatar	0.87
Czech	0.12%	Breton	0.87
Greek	0.10%	Tajik	0.87
Tatar	0.10%	Lingala	0.87
Lingala	0.10%	Basque	0.87
Hebrew	0.10%	Burmese	0.87
Swedish	0.10%	Yiddish	0.87
Croatian	0.10%	Hawaiian	0.87
Haitian	0.08%	Hausa	0.87
Pushto	0.08%	Swedish	0.87
Bashkir	0.08%	Uzbek	0.87
Catalan	0.08%	Norwegian	0.87
Georgian	0.08%	Sanskrit	0.87
Bulgarian	0.06%	Afrikaans	0.86
Lithuanian	0.06%	Malagasy	0.86
Afrikaans	0.06%	Croatian	0.86
Slovak	0.06%	Manx	0.86
Manx	0.04%	Norwegian Nynorsk	0.86
Finnish	0.04%	Serbian	0.86

Slovenian	0.04%	Latvian	0.86
Macedonian	0.04%	Somali	0.86
Luxembourgish	0.04%	Bashkir	0.86
Uzbek	0.04%	Persian	0.86
Serbian	0.02%	Slovenian	0.86
Mongolian	0.02%	Catalan	0.86
Albanian	0.02%	Guarani	0.85
Malagasy	0.02%	Haitian	0.85
		Esperanto	0.85
		Mongolian	0.85
		Albanian	0.84
		Amharic	0.84
		Macedonian	0.84
		Cebuano	0.84
		Interlingua	0.84
		Scots	0.82
		Occitan	0.78
		no language	0.00
		unknown	0.00

Appendix C: Table of key values

Table 9. Key values from the Dialing for Videos sample.

Size of random sample (unique videos)	10,016
Unique channels represented	9,997
Estimated size of YouTube (public videos)	9,881,141,822
Mean views	5,868.02
Median views	35.00

Mean comments	5.32
Median comments	0.00
Mean likes	16.48
Median likes	0.00
Mean subscribers	55,100.04
Median subscribers	4.00
Pearson correlation (r) - views and comments	0.46
Pearson correlation (r) - likes and comments	0.35
Pearson correlation (r) - views and likes	0.30
Pearson correlation (r) - likes and subscribers	0.03
Pearson correlation (r) - comments and subscribers	0.10
Pearson correlation (r) - views and subscribers	0.11
Mean duration (seconds)	615.14
Median duration (seconds)	126.00
Live-streamed videos	5.82%
Videos with tags	37.43%
Videos with chapters	1.87%
Videos with captions	38.41%
Category: Autos & Vehicles	1.18%
Category: Comedy	1.82%
Category: Education	3.43%
Category: Entertainment	7.11%

Category: Film & Animation	1.95%
Category: Gaming	12.78%
Category: Howto & Style	1.47%
Category: Music	6.11%
Category: News & Politics	2.67%
Category: Nonprofits & Activism	0.97%
Category: People & Blogs	55.81%
Category: Pets & Animals	0.81%
Category: Science & Technology	0.97%
Category: Sports	1.92%
Category: Travel & Events	1.01%