Media and Communication
2026 • Volume 14 • Article 12085
https://doi.org/10.17645/mac.12085

cogitatio

ARTICLE

Open Access Journal

TubeStats and TokStats: Research Tools for Random Samples of YouTube and TikTok

Kevin Zheng [1], Reagan Keeney [2], Ryan McGrady [2], Vikramaditya Jaisingh [2], and Ethan Zuckerman [2]

[1] School of Information, University of Michigan, USA
[2] Manning College of Information and Computer Sciences, University of Massachusetts Amherst, USA

Correspondence: Ryan McGrady (rmcgrady@umass.edu)

Submitted: 1 February 2026 **Accepted:** 28 May 2026 **Published:** 2 July 2026

Issue: This article is part of the issue "Open Research Infrastructures and Resources for Communication and Media Studies" edited by Silke Fürst (University of Zurich), Johannes Breuer (Center for Advanced Internet Studies / University of Duisburg-Essen), Erik Koenen (University of Bremen), Dimitri Prandner (Johannes Kepler University of Linz), Christian Schwarzenegger (University of Bremen), and Christian Strippel (Weizenbaum Institute), fully open access at https://doi.org/10.17645/mac.i504

Abstract

YouTube and TikTok are two of the most popular digital communications platforms in the world, playing a disproportionately large role in global communications infrastructure in general and the consumption and dissemination of information in particular. As neither platform provides adequate mechanisms to produce representative samples of the content they host, researchers largely depend on opportunistic samples of popular, recommended, or otherwise known content. In this article, we present two dashboard-based tools, TubeStats and TokStats, built upon our recent research into random sampling techniques for each platform. These tools provide platform-wide statistics such as the number of hosted videos, view count distributions, linguistic distributions, and growth over time, which researchers can use to quantify and contextualize their research. We explain the architecture and sampling pipeline of each tool as well as the unique technical and methodological affordances and constraints involved with each. We document how these related techniques and tools have been applied by our lab, other scholars, and journalists to contextualize non-representative samples, compare platform use across languages and regions, and examine quotidian uses of the platforms that attention-optimized samples may obscure, as well as the broader range of methodological possibilities that representative sampling opens for platform research. Not to be taken for granted, we also explain the many challenges we face in developing and maintaining such tools, with implications for the practical development of open research infrastructures.

Keywords

open research infrastructure; platform studies; random sampling; social science tools; TikTok; YouTube

1. Introduction

YouTube and TikTok are two of the most successful internet communication platforms. They are stages for self-expression, drivers of popular culture, and essential communications infrastructure. They have given rise to new forms of entertainment (e.g., Burgess & Green, 2018) and new business models and economies (e.g., Ørmen & Gregersen, 2023), while simultaneously transforming existing media (e.g., Cayari, 2011). As much of what people use the internet for is concentrated into fewer and fewer platforms owned by for-profit companies, they have come to own and have influence over large parts of the global public sphere and play an active role in learning, beliefs, and opinions (e.g., Bryant, 2020; Duffy, 2008; Rieder et al., 2018). Since 2020, conversations about these platforms have frequently been critical, for example, concerning the mental health of young users (e.g., Haidt, 2024) or anxieties about the national security implications of applications developed by geopolitical adversaries (e.g., Zeng & Kaye, 2022).

Despite their importance and prevalence in discourse, the content that platforms host and the ways they operate remain mostly unknown. This is in large part due to the difficulty in accessing platform data, which has only intensified after the "APIcalypse" (Bruns, 2019). In the "post-API age" (Freelon, 2018; Freelon et al., 2024), companies are eliminating or heavily restricting access to tools capable of retrieving sufficient data to answer key questions about these influential sociotechnical systems. In particular, representative samples of platform content and the ways people use platforms have been especially difficult to produce and access. Studies are thus more often based on opportunistic samples, relying on keyword searches or sets of known videos or channels either as the sample or a starting point. While valuable insights can be gained from this approach, it lacks the statistical power to be generalized to the whole and may oversample popular content in a way that emphasizes consumer-side impact rather than how the platforms are used by uploaders (Hargittai, 2020; Tufekci, 2014). Access to data is a persistent challenge and limitation in studies of communication and media. As sanctioned methods for platform data access all but disappear, researchers have called for strengthened data disclosure regulations to expand access to platform data and ensure open scientific research (e.g., Davidson et al., 2023; Rathje, 2024). However, in the lax regulatory environment for-profit technology platforms often enjoy—especially in the US—they simply have little motivation to disclose information that could be used by competitors or critics, even if it would serve the public interest. Recent regulatory developments, most notably the EU's Digital Services Act (DSA), take meaningful steps towards requiring data access for researchers, although the implementation has been uneven and practicalities remain unclear (Darius et al., 2026).

Until platforms provide sufficient data to independent third parties to conduct transparency and accountability research, the onus is on researchers to develop methods to collect data from the platforms and share their findings. In this article, we describe two systems to produce random samples of YouTube and TikTok, sharing aggregate statistics through a pair of web-based tools called TubeStats and TokStats. The links to access these tools are listed below, along with the repositories containing the code used to build the dashboards and generate random samples of their respective platforms:

- TubeStats: https://tubestats.org
  - TubeStats front-end code: https://github.com/iDPI-Umass/tubestats-www
  - YouTube sampling code: https://github.com/iDPI-Umass/youtubescripts

- TokStats: https://tokstats.org (in development, planned release by September 2026)
  - TokStats front-end code: https://github.com/iDPI-Umass/tokstats-www
  - TikTok sampling code: https://github.com/iDPI-Umass/tiktokstats

Through regular updating of these samples and allowing comparisons between them, we show a gradual change in our digital communications infrastructure while replicating our own methods. In accordance with principles of open science in communication (Dienlin et al., 2021), we make these tools publicly available, as well as the code behind them. We also share our samples, but, as explained in Section 6, we require agreement to privacy principles and do not make them fully public.

2. Platform Data Access in the Post-API Age

Social media sites host billions of pieces of content, posted by hundreds of millions of users. Although the most common term we use to describe them, "platforms," implies they are neutral, optional stages on which to stand (Gillespie, 2010), they have become essential digital communication infrastructures (Plantin et al., 2018; van Dijck et al., 2018). Many questions researchers have about these platforms concern the nature of this content and the platform's response to the posting of this content. Researchers might want to know whether YouTube is used for political speech or activism, and whether original political content is more common than posting clips from speeches or news reports. Regulators in the US are especially interested in seeing whether political content from the right is being moderated at different rates than content from the left (Jackson, 2025).

Researchers can conduct experiments by using keyword searches to retrieve social media content on a particular topic, or they can study influential social media content by capturing the output of the most popular content producers via tools like SocialBlade. There is important work based on these methods, studying subjects like public library use of short-form video (Soelseth et al., 2025), communication of Covid-19 information (Li et al., 2021), and political radicalization (Ribeiro et al., 2020). But researchers using opportunistic samples must be careful about the claims they make, as they typically cannot be generalized to the whole or even large parts of the whole.

Content posted by the most popular creators is furthermore not representative of content posted by most users of the platform. Not all users are aspiring influencers, and extrapolating from the behavior of popular users will miss a wide range of unanticipated uses of social tools (Khan, 2017; Zuckerman & McGrady, 2026). Similarly, content retrieved via keyword search or other mechanisms processed by platform filtering systems is likely also non-representative (Annabell et al., 2025; Lin, 2026). Platforms have an incentive to maximize engagement, so returning popular content for any given search is a better strategy than presenting a random sample of content that matches a particular keyword. Without insight into the operations of a search engine, it is dangerous to assume that the results of a keyword search are exhaustive, as a platform might only return search results that have been vetted or exceed a certain threshold of popularity.

Studies focused on influence, recommendations, or user exposure must use samples that are not representative, but many of them would benefit from the contextualization made possible through random sampling. For example, searching for mentions of prominent anti-vaccine advocates on Facebook will surface thousands of mentions that collectively total billions of views (Avaaz, 2020), but it is unclear whether this information is common or uncommon on Facebook without knowing how much content is posted to the

platform as a whole and how likely a user is to see a particular piece of content. Similarly, when a study reports on a mean or median view count of a sample of YouTube videos, the number is useful to quantify impact, but how does it compare to the mean or median views of videos on YouTube in general? These missing platform statistics are "denominator problems" and "distribution problems" (Zuckerman, 2021), and broad claims based on data without denominators can be deceptive or misleading (Leetaru, 2019).

Our two tools, TubeStats and TokStats, are designed in part to address denominator and distribution problems in social media research. First, by estimating the total content size of each platform based on our random sampling techniques, we provide a defensible denominator to contextualize research findings. The videos found by searching for keywords can now be understood as a specific fraction of all content on YouTube. Similar contextualization is possible within time ranges, languages (YouTube), geographic regions (TikTok), or other large subsamples, with the caveat that the more one drills down to particulars within a representative sample, the more uncertainty increases.

Second, by providing a distribution curve of views, likes, comments, and subscribers, we make it possible for researchers to contextualize engagement metrics found in narrower samples. Should a researcher retrieve a set of YouTube videos through a keyword search with view counts between 1,000—10,000, our data makes it clear that those videos are in the top 10% of all videos in terms of view count, suggesting that a keyword may be associated with popular videos, or pointing to a possible methodological issue where YouTube's search engine might surface popular videos rather than a random sample.

3. Methods for Collecting Representative Samples of YouTube and TikTok

While both YouTube and TikTok have APIs that researchers can use to retrieve data for academic research, these systems impose restrictive technical limitations on the type and quantity of data that can be retrieved. What data is retrieved is furthermore opaque and unverifiable for its comprehensiveness or representativeness. Verifiability is crucial for creating a representative sample, where each data point within a sample could be representative of millions of videos on the platform, and descriptions of the sample overall aim to be descriptive of the platform as a whole. Past work has documented the limitations of the sanctioned methods for obtaining data. For example, an audit of TikTok's Research API revealed substantial temporal bias of their "random" API endpoint, with over 55% of the videos returned being posted on Saturdays, indicating that TikTok's "random" endpoint does not operate as expected (Corso et al., 2024). Another audit found discrepancies between the view count and other data returned by the TikTok Research API and the web interface (Pearson et al., 2025).

Beyond technical limitations, users of official research APIs are also subject to contractual obligations that regulate the kinds of questions researchers can ask and methods they can use. These obligations also commonly require researchers to seek authorization for study specifics and to pre-submit research artefacts to the platform owners prior to publication. Such policies may be intended to protect the privacy of their users, but they may also influence otherwise critical research for fear of losing future data access. Fundamental public interest transparency research, such as random sampling, then, needs alternative, independent, reproducible pathways. We address ethical issues in Section 6.

Random sampling of online platforms like YouTube and TikTok is theoretically possible due to the architecture of the internet. Every video on a platform must have a unique URL path or identifier to allow a user to retrieve a specific video from the platform's servers. While the form and stochasticity of these URLs vary by platform, we can take advantage of this indexing scheme, querying a very large number of randomly generated but potentially valid URLs to obtain a large enough sample to make statistically confident claims about the whole. We can do this using a combination of third-party tools and libraries, without the use of a research-specific API.

However, when every video that has ever been created on a platform, and every video that could conceivably ever be uploaded, needs a unique ID, the search space must be incredibly large, with a lot of room to grow. Both YouTube and TikTok have quintillions of possible IDs. Searching across the full space would be prohibitively computationally expensive, so to make random sampling more realistic, we must find ways to narrow the search space wherever possible. By working to understand the structure of a platform's ID schema, we can combine queries or hold sections of IDs constant to more efficiently find extant videos with fewer requests. Depending on the features of the platform being sampled, we can query the platform either with an existing third-party software tool or with a programmatically controlled web browser.

3.1. YouTube: Strategically Using the Search Engine

After "youtube.com/," YouTube video URLs contain "watch?v=," followed by a case-sensitive 11-character ID. While YouTube URLs are case-sensitive, the platform's search engine is not, and IDs are case-insensitively indexed and searchable. That allowed us to query purely random IDs until we had built a large enough sample, a technique we call dialing for videos (McGrady et al., 2023). Dialing for videos produces a true random sample, but is still slow due to randomizing the full ID. To address this inefficiency, we used our sample to validate an older method called random prefix sampling (Zhou et al., 2011), which leverages a quirk of the search engine. Video IDs that contain a hyphen are indexed and made searchable in the search engine in three ways: by the full ID, prefix (left of the hyphen), and postfix (right). Narrowing the search space to videos that include a hyphen and begin with alphabetic characters (to take advantage of case insensitivity) greatly improves the efficiency of our search. While there has been some doubt as to the true randomness of this approach, we found that it produced results sufficiently similar to our much less efficient dialing for videos sample. Consequently, we use random prefix sampling for TubeStats and recommend the method to other researchers.

Many independent and well-maintained tools exist to scrape and download YouTube data. We search using the Python package Innertube (Bulled, 2021). The audio and metadata of extant videos are downloaded with yt-dlp (yt-dlp, 2020). The metadata does not consistently include language or location information, so we process audio using OpenAI's Whisper multilingual language model for spoken language identification (Radford et al., 2022), then immediately discard the audio. We do not archive either audio or video.

3.2. TikTok: Timestamp Plus Random IDs

While most TikTok users likely access the platform on their mobile devices, the platform is also accessible on desktop web browsers. Visiting a TikTok video on desktop reveals that, similar to YouTube, each video has a unique identifier in the URL. However, unlike YouTube, these URLs also contain the uploader's alphanumeric username, a non-random and hard-to-guess value. But TikTok videos are retrievable with knowledge of the unique ID alone, making it safe to ignore the uploader portion of the URL (Zheng et al., 2026).

From previously published forensic work (Benson, 2020), we can further narrow the 10 quintillion possible decimal values that fit in 19-digit TikTok IDs by converting the decimal ID into its 64-bit binary representation. This sequence is further divided into two distinct subsections. The first 32 bits are associated with the Unix timestamp for post creation (Benson, 2020). In the second half, a number of bits can be held constant while other bits vary. We begin by selecting a random second between the creation of TikTok and the time of execution to sample. This is used as the first 32 bits. For each random second, we randomly generate tens of thousands of combinations of characters for the second 32 bits and query each generated ID using Selenium, retrieving metadata when a video exists at the address. We iterate this process until we produce a large enough random sample to make confident statistical claims about the platform as a whole (Zheng et al., 2026).

4. System Description

4.1. TubeStats

Our data dashboard for YouTube's platform characteristics is publicly available at TubeStats.org (Figure 1) and demonstrates the kinds of analysis that random sampling makes possible. At a high level, TubeStats aims

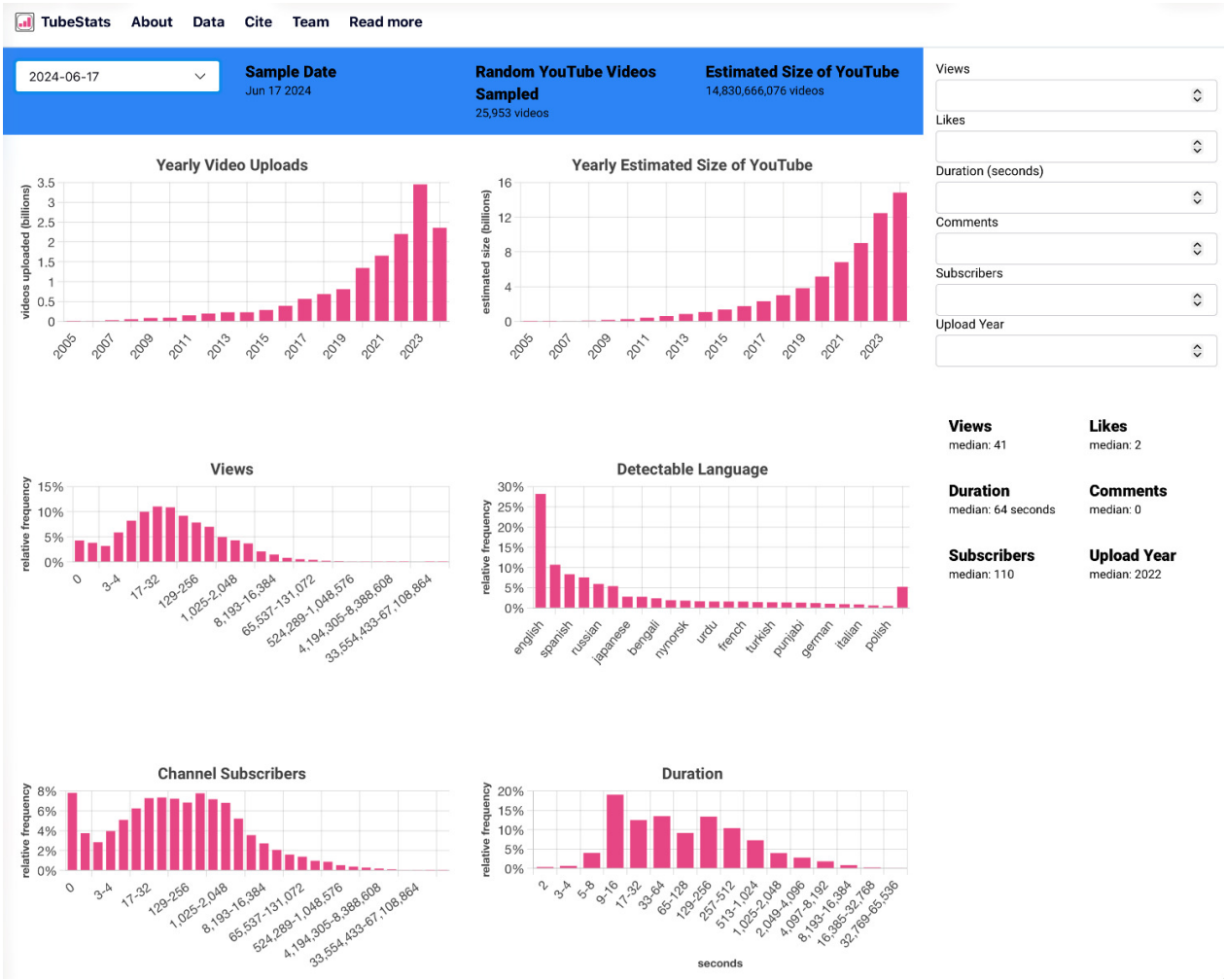

Figure 1. TubeStats dashboard, displaying a sample from June 2024.

to provide critically necessary transparency data to researchers, regulators, journalists, educators, and any member of the public looking for up-to-date data about YouTube. Our dialing for videos article, published at the end of 2023, was based on data collected in 2022, making the data outdated at publication time (McGrady et al., 2023). After demonstrating and validating the random sampling method for YouTube, we developed this website to provide data at a faster cadence than a standard academic publication process could allow for. The website was developed using the SvelteKit framework, and the static data is hosted in an S3 bucket.

The statistical summaries of our samples provided on the dashboard help to model the platform along a number of metadata fields, while preserving the privacy of users whose videos we randomly sampled. Visitors to TubeStats will encounter three main sections: a dated sample selector and selected sample information in the top blue section, a number of charts modeling the distributions of various metadata fields in the selected sample, and text-input boxes to test where a given quantitative value would rank, measured in percentiles of the sample.

The sample selector allows TubeStats users to select the various samples we have collected over time, allowing for time-based comparisons and analysis of how the platform has changed or grown over time. We developed TubeStats with the intent of providing regular updates, so long as the sampling method and our software packages were still functional. As discussed in Section 6, recent changes to YouTube's access policies pose short-term impediments to our regular sampling, though we are working to find workarounds.

We provide charts of the estimated yearly uploads to YouTube and cumulative size, distributions of Whisper-detected languages, user-selected video categories, accessibility in YouTube music, proportion of videos that were livestreamed, proportion of videos with age restrictions, and logarithmic distributions of view counts, subscriber counts, video durations, like counts, and comment counts. We implement these charts with the Chart.js JavaScript package, which provides a flexible framework for displaying these varied distributions. Chart.js also allows users to hover their mouse cursor to get the specific value of each bar.

The text-input boxes allow users to test views, likes, durations, comments, subscribers, and upload year values against our random sample. While most of the charts on TubeStats are presented in logarithmic distributions due to the wide range of values for these metadata fields, these text input boxes calculate the percentile ranking of an entered numeric value. For example, if you were interested in seeing where a video with 150 views would rank relative to the entire platform, typing "150" into the views input box reveals the video to be in the 68th percentile of all videos on YouTube.

Behind the bar charts and percentile calculators for each sample is a JSON file that contains the summary statistics that are displayed to users in the three main sections of the website. We generate these summary statistics for our charts with a script we developed that uses functions from the pandas Python data analysis library to logarithmically bin or categorically count the videos in our random samples. The percentile ranges are effectively a linear interpolation of our entire sample which we have divided into 100 segments. The input tool calculates the segment that an entered value would fall in and reports that back to the user.

Since these summary statistics contain no identifiable information about the actual videos contained in each sample, we safely anonymize the contents of our samples while still providing useful information about the overall contours of the samples. One limitation of this approach is that because we have pre-processed our

sample and only deliver to users the data required to generate the charts and percentile calculations, users cannot customize the provided charts (for example, changing the scale of the x-axis) or add new charts for other metadata fields.

Beginning in mid-2026, TubeStats will also provide links to access our datasets. Each sample will have two corresponding datasets: one public and one restricted. The public dataset has all fields removed that could conceivably be used to tie the data to a specific video. The restricted dataset will be shared with researchers upon application and agreement to privacy terms for reasons explained in Section 6.

4.2. TokStats

Once we developed our methodology for collecting random samples of TikTok, we got to work on developing TokStats to provide transparent data for this platform that has seen rapid adoption since the late 2010s. Building on the ethos and framework we established with TubeStats, TokStats (Figure 2) adapts our existing SvelteKit-based template to present the characteristics of a new random dataset collected from a different platform context. As with our other tool, we obscure personally identifiable information which may be contained within our data by only reporting results in aggregate. While visually similar to our previous tool, we adapted to the differing metadata fields that TikTok provides, including a distribution of the location of uploaders. TokStats will be publicly available in mid-2026.

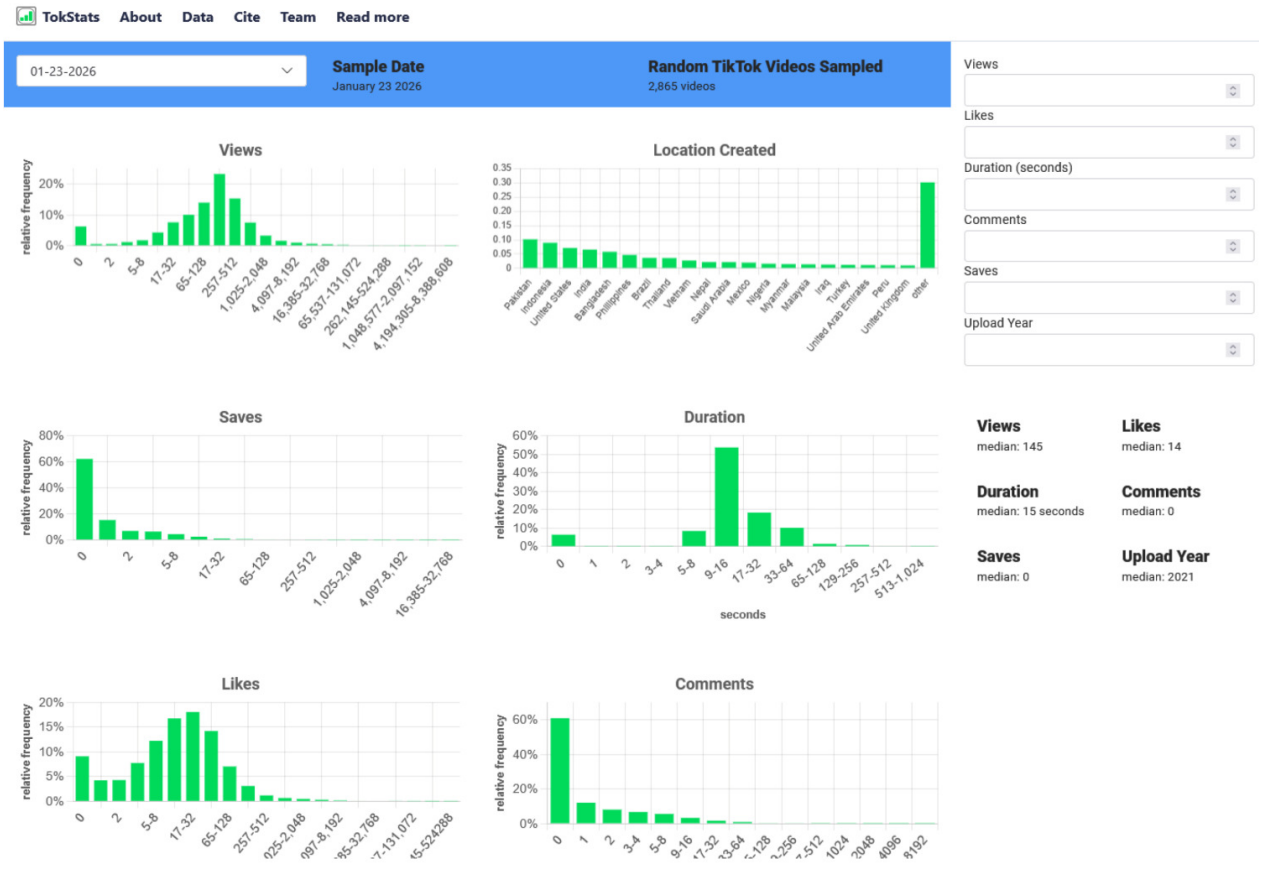

Figure 2. TokStats dashboard, showing data from a partial sample in January 2026.

Like TubeStats, we also aim to provide regular updates and new samples through TokStats to ensure researchers and the public have access to fundamental, current transparency data, so long as our methods remain usable to measure the platform. However, one difference with TikTok compared to YouTube is that even with our methodological efficiencies and parallelized sampling processes, sample generation takes on the order of months, compared to just a few hours for YouTube. This time limitation means that our updates will only be as frequent as our sampling methods will allow. Pairs of public and gated datasets will likewise be shared on TokStats.

5. Use Cases

Our datasets and findings have been used to provide insight into cultural uses of the platforms, to surface quotidian use cases obscured by attention-optimized engagement, to contextualize other non-random samples and ground methodological claims, and as the starting point for qualitative studies. This section provides an overview of these uses, carried out by our lab and other scholars, as well as other potential use cases for journalism and scholarship.

There is an assumption built into much platform research that platform use patterns in one culture are generalizable to people in other cultures (Matassi & Boczkowski, 2023). While any researcher would, if pressed, admit this is absurd, it is nonetheless common to extend claims based on studies of users in one geographic area, or speakers of one language, to platforms as a whole. Equipped with a large enough representative sample, it is possible to subdivide to create representative samples of constituent parts, with the caveat that uncertainty grows as the sample gets smaller. Once we published our initial random sample of YouTube, we expanded our sample and improved our language detection pipeline to compare representative samples of English, Hindi, Russian, and Spanish YouTube. What stood out was just how different Hindi YouTube was from the other three, with videos that are much newer, much shorter, and more often categorized as education and entertainment, with a different pattern of liking in relation to views (McGrady, Zheng, & Zuckerman, 2025).

Representative samples provide a clearer picture of how these platforms are actually used by their diverse stakeholders. Much discourse about TikTok in the 2020s focuses on the US, which passed a law intended to ban it in 2024 (Bolton, 2024). But most TikTok videos are not uploaded from the US. In the late 2010s, TikTok was exceptionally popular in India, but today it is most popular in other Asian countries, not in the US or Europe (Zheng et al., 2026). Representative samples are thus the key to understanding these global phenomena.

YouTube and TikTok are vast "accidental archives" of human communications and culture (Zuckerman & McGrady, 2026), but most of what we know about the videos they contain is based on a tiny portion of the whole: the popular content. That leaves a vast quantity of overlooked, unpopular material that is often very different from its more popular counterpart. As implied by the terms used to describe unpopular social video content, like "digital obscura" (Berliner, 2024) and "Deep YouTube" (McGrady, 2024), the chief impediment is that it is difficult to access. By casting a wider net of videos to study, we provide a window to see how people use YouTube not as watchers but as uploaders, and the wide range of quotidian functions it plays. BBC journalist Thomas Germain (2025) used our sample to explore "the hidden world beneath the shadows of YouTube's algorithm," painting a picture of "shaky camera work and voices meant for no one in particular" (para. 37).

The most common use of our work is to contextualize other non-random samples. Scholars use a wide range of methods to generate the samples they study, but keyword-based or search-based sampling systematically oversamples popular, algorithmically amplified content. A researcher studying YouTube videos about antidepressants, for example, might build a sample with a median view count that is orders of magnitude above the platform median, which reframes their findings as claims about visible public discourse rather than typical uploading behavior around that topic. Our dataset provides a representative sample for use as a baseline or benchmark to compare non-representative datasets against.

For a survey of how social media platforms design their algorithmic content feeds, Edelson, Haugen, and McCoy (2025) used our representative YouTube view count distributions as a cross-platform comparison of attention concentration on X/Twitter, finding the latter "to be significantly more imbalanced than YouTube" (p. 15). Speaking to the rarity of this work, the authors went on to comment that "ideally, we would have more than one comparison point but we are unaware of published analyses of random samples of other platforms" (p. 15). Similarly, our samples have been used to contextualize the skewness of engagement on Facebook (Edelson, Kovba, et al., 2025), the mean views of a sample of abortion-related YouTube content (Herold et al., 2025), videos in the YouTube-Commons corpus of freely licensed videos (La Rocca et al., 2025), and the amount of educational (Venegas Mejía et al., 2024) or how-to content on YouTube (Perraud, 2025). Addressing some of the long-standing difficulties with conducting cross-platform comparative research (Blank & Lutz, 2017), our methods and tools provide the necessary baseline statistics needed to situate comparative analyses of platforms.

Tonneau et al. (2025) combined our YouTube and TikTok datasets with platform transparency data recently mandated by the European DSA to quantify cross-lingual disparities in the allocation of moderation workers. In other words, because our datasets allow the quantification of video content by language and the DSA data includes the number of moderators by language spoken, it was possible to compare how much video content an English-speaking moderation worker would be responsible for, versus a Spanish- or Hungarian-speaking worker, for example.

Habib and Nithyanand (2025) used a sock-puppet audit method to study how YouTube's recommendations might reinforce negative emotions. They created topic-specific seed videos and supplemented them with a random sample of videos using the method we validated, acting as a control—an "unbiased baseline that better represents the platform's content library" (p. 5).

When not used as a concrete part of the methodology, individual findings serve to anchor argumentative points similar to those we make here: that the true size of YouTube dwarfs any particular form of content (Samaranayake, 2025), that most studies of platforms like YouTube only examine a tiny portion of the whole (Amico-Korby et al., 2026), that more content exists on YouTube than could possibly be evaluated by human moderators (Skrobisz, 2025), or that YouTube's recommendation algorithm makes it difficult to understand what is on YouTube (Linscott, 2024).

Journalists have used and discussed this work, too. When, in NetChoice, LLC v. Paxton, Supreme Court Justice Samuel Alito asked how much YouTube would weigh if it were a newspaper, Philip Bump at *The Washington Post* referenced our article and the TubeStats tool to provide an answer while commenting on the value of the analogy: about 350,000 pounds of newspaper each day (Bump, 2024).

Our methods are largely computational and quantitative, which means we cannot make broad claims about the content of videos. In the spirit of collaboration in open science in communication (Dienlin et al., 2021), we partner with qualitative researchers and scholars with deep cultural knowledge to analyze videos after identifying interesting patterns or anomalies in metadata. For example, to follow up on our language comparison, we are undertaking a qualitative study to better understand the reasons for these differences. Preliminary findings indicate a larger proportion of content meant for friends, family, or small audiences, likely tied to India's unique sociotechnical history with short-form video.

Moving forward, we anticipate that economists studying digital labor could use our view count distributions to estimate what percentage of video creators could plausibly earn income from their content (not very many). Any study that uses YouTube's or TikTok's recommendation system to generate a sample can evaluate the extent to which that biases the data versus a representative sample. Communications scholars studying virality have comparable datasets to compare videos that get a lot of views and those that do not. There are many possibilities for longitudinal analyses: shifts in types of content hosted, cultural trends, regions or languages represented, and even inferences into what the recommendation algorithms are lifting up.

6. Discussion

6.1. Challenges and Limitations

YouTube and TikTok are two of the most popular sites of human expression, discourse, and documentation in the world; yet they are notoriously difficult to research. Methods that produce results capable of describing platforms as a whole are especially challenging, requiring random sampling techniques outside of the provided research structures. We have undertaken this work and believe the findings can not just inform discourse about the platforms, but also help other researchers contextualize their studies of narrower samples and identify sites for more focused qualitative work.

Maintaining an open research infrastructure with up-to-date samples is challenging for a combination of technical, structural, and financial reasons. Even after significant efficiency gains through our methods, our search space is still large and computationally expensive to sample. Changes in platform architecture are often not well documented, leading to a range of data quality issues. Metadata fields, search operators, data locations, and formats can all change with little or no public notice, and such changes may be introduced inconsistently. Continuous testing and auditing are therefore necessary, requiring both time and resources.

Sometimes these platform changes are intentional, to make access more difficult. Platforms, particularly YouTube, have become increasingly intolerant of automated queries and are more prone to throttle or block based on IP addresses that they deem to be querying in excess, a phenomenon which we attribute at least in part to the recent increase in for-profit scraping of YouTube to collect data to train commercial AI models (McGrady, Zuckerman, & Zheng, 2025). Although public interest researchers are not the likely target of these interventions, it nonetheless forces us into a game of cat and mouse to update libraries and scripts to comply with (or work around) new technical requirements. For example, yt-dlp users on Github recently noticed that YouTube shifted its streaming delivery away from the standard DASH protocol toward a new, custom protocol referred to as SABR (server-side adaptive bitrate). While the old DASH protocol had easily accessible video and audio streaming URLs, SABR is more session-oriented and less transparent. Tools like

yt-dlp, which we use to download audio files and metadata from YouTube, can no longer reliably retrieve data in the same way. Newer versions of yt-dlp additionally require users to set up an external JavaScript runtime (EJS) to work around SABR, adding additional complexity to random sampling. While the success of repository maintainers for yt-dlp should not be discounted, frequent periods of non-functionality and constant updates can interrupt this kind of transparency research.

In addition to these technical challenges, the limited availability of financial resources to support this kind of work will also challenge the maintenance of our research tools. Funding for scientific research in general is declining (McKie, 2026), especially in the US, where our project is based. Even before these declines, grant support is often more focused on projects and papers than on long-term support of research infrastructure. Tool grants tend to be tied to broader research projects with concrete outputs or rely on the addition of features, which can make overall project sustainability more difficult. For example, Media Cloud, a database and research platform for studying news, had to be completely reengineered after continued expansion led it to become brittle due to feature creep (Bermejo et al., 2026).

The code we use to generate our samples and dashboards is available through repositories hosted on GitHub, and we allow others to remix and reuse our code for their own contexts and projects. Ensuring that our public web-based dashboards remain useful to researchers will require continued development effort, and while our team will continue to be the primary stewards of the project's assets, we hope that others who depend on these tools can offer feedback, suggestions, and expertise in sustaining this work.

We recognize that the resulting data and analyses from these methods, where the intended level of analysis is the entire platform, are inherently limited in their explanatory power for small-scale individual or community behaviors. While metadata cannot adequately describe or analyze the content of videos, we believe that patterns and trends found within representative samples can help guide future directions for inquiry. By quantitatively surveying the field of videos contained within a platform, we can find interesting patterns and anomalies ripe for investigation using other methods. Heeding Humphreys et al.'s (2021) call for stronger collaborations across the quantitative and qualitative divide in communications research, we have several ongoing projects with social media scholars working in their particular methodological, disciplinary, and cultural contexts, and continue to seek new collaborators interested in digging into this data.

6.2. Ethics

The methods we use to produce our random samples of YouTube and TikTok differ from one another, but are similar in that they are both examples of "unpermissioned research." In the "post-API age" (Freelon, 2018), it has become increasingly difficult—often impossible—for researchers to obtain sufficient access to platform data through officially sanctioned channels. As described above, neither YouTube nor TikTok have adequate mechanisms to produce representative samples of videos. Even if they did, we believe methods that do not seek explicit permission from companies are essential tools for social scientists. Importantly, we believe such methods can be used ethically and are often the only way to retrieve data that is in the public interest. This position is not without precedent—the EU's DSA recognizes that public data should be accessible by researchers (Darius, 2024). Following the distinction made by Breuer et al. (2025), our approach is platform-centered rather than user-centered: Instead of recruiting participants to share their data, we sample platforms directly, centering ethical questions on privacy and data handling rather than informed consent.

YouTube and TikTok are commonly understood through their professional creators, viral content, and videos lifted up by their recommendation systems. Something that became clear when studying everyday uses of these platforms through representative samples is just how often uploaders do not appear to be reaching for a broad audience. Many appear to be "publicly private" (Lange, 2007) or "accidental vlogs" (McGrady & Snehi, 2025), whereby uploaders disclose aspects of their lives with a small audience, but do so in public. Helen Nissenbaum and danah boyd described this phenomenon in terms of "contextual integrity" (Nissenbaum, 2010) and "collapsed contexts" (boyd, 2008), referring to the way material shared with one audience on a social platform is easily viewed or appropriated for other audiences. Rich media like video can furthermore leak incidental data via personal property or other people in the background (Kutschera et al., 2024). To be sure, most uploaders do not anticipate their videos becoming part of an academic reference dataset. This poses complex privacy questions about the handling of our datasets.

The aggregate statistics reflected in our TubeStats and TokStats dashboards represent a compromise. We want to ensure that researchers have easy access to fundamental characteristics of our digital video communications infrastructure that companies do not readily disclose, but we also want to respect the privacy of video creators who did not anticipate their uploads becoming part of an academic reference dataset. By focusing on providing denominators and distributions based on metadata, we allow comparison without risking exposure of personally identifiable information in individual videos.

Starting in mid-2026, TubeStats and TokStats will include access to the datasets themselves, but in two versions. The first version will be the full random samples we use in our own research. These will only be shared with researchers who apply for access and agree to strict privacy requirements including a commitment not to distribute the data, not to disclose any personally identifiable information about uploaders, not to disclose data that would allow others to extract personally identifiable information, and not to try to connect videos with identities of uploaders. We will share these datasets on a case-by-case basis upon application review, and plan to explore additional hosting options to allow for scaling access.

To maximize the usefulness of the data, we will also distribute a public version of the datasets. These will be identical except all fields that could conceivably be used to identify an uploader or a video—identifier, title, description, uploader name, channel name, etc.—will be removed. We view this public version as a counterpart to the TubeStats and TokStats dashboards, presenting raw statistics about undisclosed videos for a range of uses, not limited to researchers who go through an application process. In line with best practices for sharing research data outlined by Bowman and Spence (2020), these datasets take the form of simply presented CSVs, processed for confidentiality, and are indeed valuable for our own continued work in addition to helping advance other scholars' research.

Funding

Support for this work was provided by the Ford Foundation, the John D. and Catherine T. MacArthur Foundation, and the John S. and James L. Knight Foundation.

Conflict of Interests

The authors declare no conflicts of interest.

Data Availability

Data is available through the TubeStats and TokStats websites, and through the Github repositories as detailed in Section 1.

References

Amico-Korby, D., Harrell, M., & Danks, D. (2026). Do it yourself content and the wisdom of the crowds. *Erkenntnis*, *91*(2), 609–637. https://doi.org/10.1007/s10670-024-00919-z

Annabell, T., Gorwa, R., Scharlach, R., van de Kerkhof, J., & Bertaglia, T. (2025). *TikTok search recommendations: Governance and research challenges*. arXiv. https://doi.org/10.48550/arXiv.2505.08385

Avaaz. (2020). *Facebook's algorithm: A major threat to public health*. https://secure.avaaz.org/campaign/en/facebook_threat_health

Benson, R. (2020, August 11). Tinkering with TikTok timestamps. *Dfir.Blog*. https://doi.org/10.21428/b0ac9c28.fcc2b9c2

Berliner, L. S. (2024). ...Like no one is watching: Taking digital obscura seriously. *JCMS: Journal of Cinema and Media Studies*, *63*(3), 164–169. https://doi.org/10.1353/cj.2024.a927692

Bermejo, F., Bhargava, R., Budne, P., Gulley, P., Leon, E., McGrady, R., Ndulue, E. B., & Zuckerman, E. (2026). Media Cloud 2.0: An updated open web news archive. *Proceedings of the International AAAI Conference on Web and Social Media*, *20*(1), 2735–2746. https://doi.org/10.1609/icwsm.v20i1.42778

Blank, G., & Lutz, C. (2017). Representativeness of social media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, *61*(7), 741–756. https://doi.org/10.1177/0002764217717559

Bolton, A. (2024, April 26). GOP's inclusion of TikTok ban is secret weapon against Biden. *The Hill*. https://thehill.com/homenews/senate/4622455-gops-inclusion-of-tiktok-ban-is-secret-weapon-against-biden

Bowman, N. D., & Spence, P. R. (2020). Challenges and best practices associated with sharing research materials and research data for communication scholars. *Communication Studies*, *71*(4), 708–716. https://doi.org/10.1080/10510974.2020.1799488

boyd, d. (2008). *Taken out of context: American teen sociality in networked publics*. SSRN. http://doi.org/10.2139/ssrn.1344756

Breuer, J., Stier, S., Lukito, J., Mangold, F., Wieland, M., & Radovanović, D. (2025). *Overview of ethical considerations when working with digital behavioral data* (GESIS guides to digital behavioral data, no. 14). GESIS—Leibniz-Institute for the Social Sciences. https://doi.org/10.60762/GGDBD25014.1.0

Bruns, A. (2019). After the "APIcalypse": Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Bryant, L. V. (2020). The YouTube algorithm and the alt-right filter bubble. *Open Information Science*, *4*(1), 85–90. https://doi.org/10.1515/opis-2020-0007

Bulled, T. (2021). *Innertube* [Computer software]. https://github.com/tombulled/innertube

Bump, P. (2024, February 27). Here's how much a YouTube newspaper would weigh, Justice Alito. *The Washington Post*. https://www.washingtonpost.com/politics/2024/02/27/heres-how-much-youtube-newspaper-would-weigh-justice-alito

Burgess, J., & Green, J. (2018). *YouTube: Online video and participatory culture* (2nd ed.). Polity Press.

Cayari, C. (2011). The YouTube effect: How YouTube has provided new ways to consume, create, and share music. *International Journal of Education & the Arts*, *12*(6). http://www.ijea.org/v12n6

Corso, F., Pierri, F., & De Francisci Morales, G. (2024). What we can learn from TikTok through its research API.

In R. Heiberger, U. Gadiraju, M. Spaniol, K. Kinder-Kurlanda, A. Faleńska, A. Mashhadi, J. Sun, S. Kaiser, & S. Staab (Eds.), *Companion proceedings of the 16th ACM Web Science Conference* (pp. 110–114). ACM. https://doi.org/10.1145/3630744.3663611

Darius, P. (2024, September 24). Researcher data access under the DSA: Lessons from TikTok's API issues during the 2024 European elections. *Tech Policy Press*. https://www.techpolicy.press/-researcher-data-access-under-the-dsa-lessons-from-tiktoks-api-issues-during-the-2024-european-elections

Darius, P., Breuer, J., Kruschinski, S., Loecherbach, F., Riedl, J., & Stier, S. (2026). Election research in the age of regulated data access under the EU Digital Services Act. *Internet Policy Review*, *15*(1). https://doi.org/10.14763/2026.1.2080

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J. F., Ayravainen, L., & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, *7*(12), 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., . . . de Vreese, C. (2021). An agenda for open science in communication. *Journal of Communication*, *71*(1), 1–26. https://doi.org/10.1093/joc/jqz052

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. https://academic-publishing.org/index.php/ejel/article/view/1535

Edelson, L., Haugen, F., & McCoy, D. (2025). A comparative survey of algorithmic feed recommendation system designs. *ACM Transactions on Recommender Systems*. Advance online publication. https://doi.org/10.1145/3757327

Edelson, L., Kovba, B., Yershova, H., Botelho, A., McCoy, D., & Lauinger, T. (2025). Measurement and metrics for content moderation: The multi-dimensional dynamics of engagement and content removal on Facebook. *Journal of Online Trust and Safety*, *2*(5). https://doi.org/10.54501/jots.v2i5.220

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Freelon, D., Monzer, C., Jeon, G., Moy, C., & Williams, N. (2024). The post-API age of social media data access: Past, present, and future. *The ANNALS of the American Academy of Political and Social Science*, *715*(1), 16–37. https://doi.org/10.1177/00027162251372557

Germain, T. (2025, April 23). The hidden world beneath the shadows of YouTube's algorithm. *BBC*. https://www.bbc.com/future/article/20250306-inside-youtubes-hidden-world-of-forgotten-videos

Gillespie, T. (2010). The politics of "platforms." *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Habib, H., & Nithyanand, R. (2025). *YouTube recommendations reinforce negative emotions: Auditing algorithmic bias with emotionally-agentic sock puppets*. arXiv. https://doi.org/10.48550/ARXIV.2501.15048

Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Penguin Press.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Herold, S., Narendorf, P., & Hoffman, B. (2025). Reading the comments: An exploratory quantitative analysis of YouTube comments in response to abortion plotlines on fictional television programs. *Sexual & Reproductive Healthcare*, *45*, Article 101135. https://doi.org/10.1016/j.srhc.2025.101135

Humphreys, L., Lewis, N. A., Jr., Sender, K., & Won, A. S. (2021). Integrating qualitative methods and open

science: Five principles for more trustworthy research. *Journal of Communication*, *71*(5), 855–874. https://doi.org/10.1093/joc/jqab026

Jackson, D. (2025, March 25). Scientists respond to FTC inquiry into tech censorship. *Tech Policy Press*. https://www.techpolicy.press/scientists-respond-to-ftc-inquiry-into-tech-censorship

Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, *66*, 236–247. https://doi.org/10.1016/j.chb.2016.09.024

Kutschera, S., Slany, W., Ratschiller, P., Gursch, S., Deininger, P., & Dagenborg, H. (2024). Incidental data: A survey towards awareness on privacy-compromising data incidentally shared on social media. *Journal of Cybersecurity and Privacy*, *4*(1), 105–125. https://doi.org/10.3390/jcp4010006

Lange, P. G. (2007). Publicly private and privately public: Social networking on YouTube. *Journal of Computer-Mediated Communication*, *13*(1), 361–380. https://doi.org/10.1111/j.1083-6101.2007.00400.x

La Rocca, L., Corso, F., & Pierri, F. (2025). *Evaluating AI capabilities in detecting conspiracy theories on YouTube*. arXiv. https://doi.org/10.48550/ARXIV.2505.23570

Leetaru, K. (2019, February 27). Whatever happened to the denominator? Why we need to normalize social media. *Forbes*. https://www.forbes.com/sites/kalevleetaru/2019/02/27/whatever-happened-to-the-denominator-why-we-need-to-normalize-social-media

Li, Y., Guan, M., Hammond, P., & Berrey, L. E. (2021). Communicating Covid-19 information on TikTok: A content analysis of TikTok videos from official accounts featured in the Covid-19 information hub. *Health Education Research*, *36*(3), 261–271. https://doi.org/10.1093/her/cyab010

Lin, Y. (2026). Algorithmic bias in recommender systems: Implications for consumer choice and information diversity. *Advances in Economics and Management Research*, *16*(1), 364–371. https://doi.org/10.56028/aemr.16.1.364.2026

Linscott, W. (2024). *Algorithmic capture* [Unpublished doctoral dissertation]. University of Auckland. https://hdl.handle.net/2292/74027

Matassi, M., & Boczkowski, P. J. (2023). *To know is to compare: Studying social media across nations, media, and platforms*. The MIT Press.

McGrady, R. (2024, January 26). What we discovered on "Deep YouTube." *The Atlantic*. https://www.theatlantic.com/technology/archive/2024/01/how-many-videos-youtube-research/677250

McGrady, R., & Snehi, H. (2025). The ethics of accidental vlogs. *M/C Journal*, *28*(4). https://doi.org/10.5204/mcj.3201

McGrady, R., Zheng, K., Curran, R., Baumgartner, J., & Zuckerman, E. (2023). Dialing for videos: A random sample of YouTube. *Journal of Quantitative Description: Digital Media*, *3*, 1–85. https://doi.org/10.51685/jqd.2023.022

McGrady, R., Zheng, K., & Zuckerman, E. (2025). One platform, four languages: Comparing English, Spanish, Hindi, and Russian YouTube. *Social Media + Society*, *11*(3). https://doi.org/10.1177/20563051251363216

McGrady, R., Zuckerman, E., & Zheng, K. (2025, January 30). AI companies threaten independent social media research. *Tech Policy Press*. https://www.techpolicy.press/ai-companies-threaten-independent-social-media-research

McKie, A. (2026, January 21). More than half of authors of leading research say funding is declining. *Nature Index*. https://doi.org/10.1038/d41586-026-00054-5

Nissenbaum, H. (Ed.). (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford Law Books.

Ørmen, J., & Gregersen, A. (2023). Towards the engagement economy: Interconnected processes of

commodification on YouTube. *Media, Culture & Society*, *45*(2), 225–245. https://doi.org/10.1177/01634437221111951

Pearson, G. D. H., Silver, N. A., Robinson, J. Y., Azadi, M., Schillo, B. A., & Kreslake, J. M. (2025). Beyond the margin of error: A systematic and replicable audit of the TikTok research API. *Information, Communication & Society*, *28*(3), 452–470. https://doi.org/10.1080/1369118X.2024.2420032

Perraud, R. (2025). *Investigating how users perceive interface differences and similarities across analogous graphical user interfaces* [Unpublished doctoral dissertation]. University of Lille. https://hal.science/tel-05423029

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*. arXiv. https://doi.org/10.48550/ARXIV.2212.04356

Rathje, S. (2024). To tackle social-media harms, mandate data access for researchers. *Nature*, *633*(8028), 36–36. https://doi.org/10.1038/d41586-024-02853-0

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. In M. Hildebrandt & C. Castillo (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 131–141). ACM. https://doi.org/10.1145/3351095.3372879

Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (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. https://doi.org/10.1177/1354856517736982

Samaranayake, R. (2025). *Exploring the use of digital technology to support mathematics learning in pre-service teacher education* [Unpublished master thesis]. University of Prince Edward Island. https://islandscholar.ca/islandora/object/17731

Skrobisz, N. J. (2025). *The evolution of morality and the problem of AI value over-alignment*. ResearchGate. https://doi.org/10.13140/RG.2.2.29296.72964

Soelseth, C. H., Bøyum, I., Colbjørnsen, T., Pharo, N., & Tallerås, K. (2025). Public libraries on TikTok: Emerging platform vernaculars of communication and distribution. *Information, Communication & Society*, *28*(14), 2521–2540. https://doi.org/10.1080/1369118X.2025.2461644

Tonneau, M., Liu, D., McGrady, R., Zheng, K., Schroeder, R., Zuckerman, E., & Hale, S. (2025). *Language disparities in moderation workforce allocation by social media platforms*. OSF. https://osf.io/preprints/socarxiv/amfws_v1

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 505–514. https://doi.org/10.1609/icwsm.v8i1.14517

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Venegas Mejía, V. L., Esquivel-Grados, J., Venegas-Mejía, C., & Gonzáles-Benites, M. (2024). Audiovisual content of YouTube videos on the research problem statement: A didactic analysis of their relevance. *Visual Review: International Visual Culture Review*, *16*(8), 235–249. https://doi.org/10.62161/revvisual.v16.5697

yt-dlp. (2020). *yt-dlp* [Computer software]. https://github.com/yt-dlp/yt-dlp

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to *visibility moderation*: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

Zheng, K., McGrady, R., Keeney, R., & Zuckerman, E. (2026). *TokStats: A longitudinal view of TikTok's multinational growth through a randomized, stratified sample*. Manuscript submitted for publication.

Zhou, J., Li, Y., Adhikari, V. K., & Zhang, Z.-L. (2011). Counting YouTube videos via random prefix sampling. In P. Thiran & W. Willinger (Eds.), *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference* (pp. 371–380). ACM. https://doi.org/10.1145/2068816.2068851

Zuckerman, E. (2021, November 2). Facebook has a misinformation problem, and is blocking access to data about how much there is and who is affected. *The Conversation*. https://theconversation.com/facebook-has-a-misinformation-problem-and-is-blocking-access-to-data-about-how-much-there-is-and-who-is-affected-164838

Zuckerman, E., & McGrady, R. (2026). The quotidian web and the accidental archive. *International Journal of Communication*, *20*, 993–1012. https://doi.org/10.65476/ppzen716

About the Authors

Kevin Zheng is a PhD student in the School of Information at the University of Michigan, where he develops research methods and tools to study online social platforms and digital infrastructures. Kevin received a BS in computer engineering and BA in science, technology, and society from the University of Massachusetts Amherst.

Reagan Keeney is a computer science PhD student at the University of Massachusetts Amherst. Their research lies at the intersection of natural language processing, human–computer interaction, and social science. They are particularly interested in large-scale mixed methods research of online social systems.

Ryan McGrady is a senior research fellow with the Initiative for Digital Public Infrastructure at the University of Massachusetts Amherst and a researcher with Media Cloud. He uses mixed methods to study large internet platforms and their place in society, with a focus on YouTube, TikTok, and Wikipedia.

Vikramaditya Jaisingh recently graduated with a BS in computer science and BA in philosophy from the University of Massachusetts Amherst, where he is a research affiliate of the Initiative for Digital Public Infrastructure. His research interests include platform studies, critical digital media studies, and human–computer interaction.

Ethan Zuckerman is an associate professor of public policy, communication, and information at the University of Massachusetts Amherst. He directs the Initiative for Digital Public Infrastructure and is the co-author of *The Field Guide to Social Media*, forthcoming from MIT Press, with Chand Rajendra-Nicolucci.