Mapping YouTube

A quantitative exploration of a platformed media system

Rieder, B.; Coromina, Ò.; Matamoros-Fernández, A.

DOI
10.5210/fm.v25i8.10667

Publication date
2020

Document Version
Final published version

Published in
First Monday

License
CC BY-NC-SA

Link to publication

Citation for published version (APA):
Rieder, B., Coromina, Ò., & Matamoros-Fernández, A. (2020). Mapping YouTube: A quantitative exploration of a platformed media system. *First Monday*, *25*(8). https://doi.org/10.5210/fm.v25i8.10667

f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Mapping YouTube: A quantitative exploration of a platformed media system

by Bernhard Rieder, Òscar Coromina, and Ariadna Matamoros-Fernández

Abstract
Over the past 15 years, YouTube has emerged as a large and dominant social media service, giving rise to a ‘platformed media system’ within its technical and regulatory infrastructures. This paper relies on a large-scale sample of channels (*n* =36M+) to explore this media system along three main lines. First, we investigate stratification and hierarchization in broadly quantitative terms, connecting to well-known tropes on structural hierarchies emerging in networked systems, where a small number of elite actors often dominate visibility. Second, we inquire into YouTube’s channel categories, their relationships, and their proportions as a means to better understand the topics on offer and their relative importance. Third, we analyze channels according to country affiliation to gain insights into the dynamics and fault lines that align with country and language. Throughout the paper, we emphasize the inductive character of this research by highlighting the many follow-up questions that emerge from our findings.

Contents

1. Introduction
2. Methodology and data
3. Findings
4. Discussion and conclusion

1. Introduction

Social media platforms play important roles on a global scale and, consequently, have been studied empirically in many different ways. The overwhelming majority of studies, however, rely on issue or user samples to investigate particular slices of social media reality. Due to limited data access, painting an ‘overall’ picture of a platform was always difficult for independent researchers and the recent ‘APIcalypse’ (Bruns, 2019) has further exacerbated the situation. In the past, researchers could rely on a limited number of newspapers or television channels to assess what kind of media contents are on offer and how their production is structured in organizational terms. But the emergence of a ‘hybrid media system’ (Chadwick, 2013), where traditional actors and networked platforms enter into complex constellations, further adds to the difficulty to assess *what is out there* .

One of the actors that take a central position in this ‘high-choice media environment’ (van Aelst, *et al.* , 2017) is YouTube. Since launching as a Web site for sharing videos in 2005 and becoming part of Google one year later, YouTube has become a dominant platform that hosts millions of channels and billions of videos, reaching an audience of more than two billion active users every month [1]. Long in the shadow of Facebook and Twitter when it comes to data-driven research, YouTube has moved into the center of scholarly interest over the last years, most notably around questions such as extreme political content (Ribeiro, *et al.* , 2020) and misinformation (Bounegru, *et al.* , 2020). This literature is concerned with the implications of YouTube’s algorithms in matters of politics and culture (Airoldi, *et al.* , 2016; Rieder, *et al.* , 2018) and follows recent media controversies on the site’s role in processes of radicalization, mischief, and abuse. Scholars writing for lay audiences have called YouTube the ‘great radicaliser’ (Tufekci, 2018), a ‘far-Right propaganda machine’ (Lewis, 2020), and a platform that inflicts ‘infrastructural violence’ on children (Bridle, 2017). Beyond these qualifications of YouTube as a threat to democracy, qualitative research has historically offered a more positive side of the platform by examining the quotidian practices of its wide range of amateur and professional users (*e.g.* , Abidin, 2019, 2018; Bishop, 2019; Lange, 2007; Sayago, *et al.* , 2012). Broader theoretical takes (*e.g.* , Kessler and Schäfer, 2009; Gillespie, 2010), anthologies (*e.g.* , Burgess and Green, 2018, 2009; Lange, 2019; Lovink and Niederer, 2008; Snickars and Vonderau, 2009), and special issues (Arthurs, *et al.* , 2018) have added to a growing body of research focusing on the importance of the video platform for everyday life, entertainment, politics, and the economy. Given its worldwide reach as cultural mediator, YouTube has also attracted scholars researching ‘issues of globalization and cultural difference’ [2] and, in particular, how tensions between local and global structures challenge established ideas about media globalization (Cunningham and Craig, 2016).

While most empirical YouTube research has focused on specific content creators, genres, texts, and subcultures, an interest to ‘map’ the platform in order to account for what is on offer has driven research since the beginning. Paolillo (2008) examined the social network structure of YouTube early on and Burgess and Green (2009) provided the first broad picture of YouTube’s popular culture in the late 2000s by conducting a content analysis of the most popular videos. In the second edition of their book, however, the authors acknowledge that their empirical approach could not be replicated today since YouTube has evolved from being a Web site for sharing videos to a global media company whose commercial interests are centered on the monetization of channels (Burgess and Green, 2018). And with success: YouTube recently announced for the first time that it generated US$15 billion from advertising in 2019, roughly 10 percent of Google’s overall revenue (Statt, 2020). The quest for profitability has triggered important changes in terms of design, content, and audiences. Reacting to a series of scandals, the company has for example begun to set up stricter rules on what counts as ‘advertiser-friendly’ content, which function as a ‘deterrence to creating risky, edgy or experimental content’ [3]. The resulting ‘professionalization’ of YouTube’s content creators has attracted scholarly attention, with Cunningham and Craig (2017) coining the term ‘social media entertainment’ to describe the type of content that is popular on YouTube. At the same time, it is clear that videos are produced and uploaded by a wide variety of actors, ranging from amateurs engaging in intimate sharing of their everyday experiences, to star YouTubers with millions of subscribers, to established television networks and music labels that use the platform to distribute their content to mass audiences, and in particular younger viewers.

A series of recent papers (Bärtl, 2018; Paolillo, *et al.* , 2019) have attempted to provide overall characterizations of YouTube in its current complexity, including available content and user dynamics [4]. These attempts to broadly describe and analyze an online platform in empirical terms are not only producing interesting insights in their own right but also provide valuable scope and context to other researchers’ work. Our own project follows similar goals and this paper serves to document both our methodology and a number of key findings that take aim at forms of stratification and segmentation marking YouTube, that is, at the hierarchies and dividing lines that carve through the channel landscape. Relying on the Web-API that YouTube provides to external developers, we were able to implement a sampling strategy based on network crawling that resulted in a very large collection of channel data (*n* =36M+). Guided by the principles of ‘exploratory data analysis’ (Tukey, 1977), we rely on this sample to shed light on a broad question: What is on YouTube and how is the channel landscape structured? Following the platform’s increasingly important division into channel ‘tiers’ (*cf.* , Kumar, 2019), we focus on three analytical directions. *First* , we investigate stratification and hierarchization in broadly quantitative terms, connecting to well-known tropes on structural hierarchies emerging in networked systems, where a small number of elite actors often dominate visibility (*e.g.* , Hindman, 2009). While Pareto distributions and power laws are nothing new, we seek to present and discuss our findings in ways that provide concrete reference points for scholars interested in YouTube, rather than abstract assessments of inequality. *Second* , we inquire into YouTube's channel categories, their relationships, and their proportions as a means to better understand the topics on offer and their relative importance. *Third* , we analyze channels according to country affiliation to gain insights into the dynamics and fault lines that align with country and language. Throughout the paper, we emphasize the inductive character of this research by highlighting the many follow-up questions that emerge from our findings.

These three directions together seek to provide a broad picture of YouTube as host to an increasingly substantial 'platformed media system' in its own right, that is, to a large and complex media ecology that has developed *within* YouTube's technical and regulatory infrastructures. Enabled, guided, and coerced by the company, this media system has fermented a 'protoindustry of social media entertainment' [5] that can reach large audiences and build viable businesses in the process. The next section presents our methodology in detail and discusses the analytical possibilities and limitations it affords.

2. Methodology and data

Conceptually, we situate our work within the frame of what statistician John Tukey (1977) called 'exploratory data analysis', which rather 'seeks an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question' [6]. Much like qualitative approaches such as grounded theory (Glaser and Strauss, 1967), exploratory data analysis can be conceived as *inductive* , that is, rather than using a preset theoretical framework to formulate narrow research questions and hypotheses, it generates questions, ideas, and theories iteratively in conversation with the data. Tukey's goal was certainly not to promote carelessness or undirected data crunching, but to argue that many of the empirical phenomena we encounter are not yet sufficiently well understood to set them into the strictures of confirmatory hypothesis testing. This applies very much to large-scale platforms like YouTube, where our understanding is hampered by size and complexity. As we will see, a central outcome of our work is the formulation of *new* questions and problems that were hardly visible before, adding to the list of directions for follow-up research. Some of this research is already in preparation and this paper serves as a methodological introduction for forthcoming work as well as a broad attempt to investigate stratification and segmentation on YouTube.

2.1. Data collection

Looking at the basic setup or *information architecture* of YouTube, the two main ‘units’ that structure the platform are videos and channels. Unlike Twitter, which issues a random sample of tweets in real time (Morstatter, *et al.* , 2014; Gerlitz and Rieder, 2013), YouTube provides no easy way to create representative datasets. Most authors have focused on channels as entry points. Bärtl (2018) proposes an approach to creating such a sample of channels (*n* =19,025) that uses randomly generated search strings to retrieve channel data (*e.g.* , the author used random key word searches such as ‘why’ or ‘gol’ to collect channels that contain the letters of the queries in their name). However, the exclusion of non-Latin characters is only one reason why it seems doubtful that this method can achieve its goal of providing an accurate picture of what is on YouTube. Popularity bias and other vagrancies of YouTube’s API are another one, but the main problem is the non-random character of language itself: the distinct phonetic patterns of human languages mean that channel names or video titles are not spread equally over the alphabet and will cluster around particular character patterns. Another recent paper (Paolillo, *et al.* , 2019) uses a combination of searching, browsing, and crawling to create a large collection of channels (*n* =549,383), but their idiosyncratic method is neither systematic, nor do the authors discuss the biases it may contain. They erroneously state that ‘channel subscriptions are treated as private by the API’ [7], missing the most important source of connectivity in the network of channels. To be clear, both of these approaches yield interesting results in the spirit of iterative exploration, but also have clear limitations. In the following section, we propose an approach that seeks to deal with these limitations by constructing a much larger sample of channels through crawling. As with all strategies that lack privileged access to a platform’s database, this method comes with its own set of shortcomings.

In the late 1990s, attempts to crawl the Web in its entirety were highly popular (*e.g.* , Albert, *et al.* , 1999; Broder, *et al.* , 2000), but such efforts are now mostly [8] limited to commercial companies like Google, Exalead, or Ahrefs. Software like the Digital Methods Initiative’s *Issuecrawler* (Rogers, 2002) or *Hyphe* (Jacomy, *et al.* , 2016), developed at SciencePo’s Médialab, provides researchers with the capabilities to create networks of smaller sub-sections of the Web and similar approaches have been adopted for crawling YouTube, for example to detect extremist content (*e.g.* , Agarwal and Sureka, 2015). One of the perks of crawling is that the collected data have a relational component, making it possible to analyze them both as *populations* through traditional statistical methods and as *networks* with the help of concepts like density, centrality, or clustering coefficient. While our work is not part of a larger ‘science of networks’ (Watts, 2004), we will discuss certain network properties of our dataset and apply graph-based methods further down. The main problem with crawling, however, is to know whether a crawl is complete and, if not, what is missing. This question will be discussed in more detail in the following section.

On the broadest level, our approach is basically an attempt to crawl a significant portion of YouTube and it builds on years of experience with topic- and community-based channel crawling via the YouTube Data Tools (Rieder, 2015). We relied on two types of connections between channels as conduits: *featured channels* allow creators to highlight other channels on their profile and *channel subscriptions* point to the idea that creators are also users that watch content through the same account. Both can also be seen as means to gain visibility on the platform, either through ‘traditional’ networking or as input into algorithmic ranking and recommendation. While the latter is largely speculation, we will see throughout this paper that ‘algorithmic imaginaries’ (Bucher, 2017) and ‘algorithmic gossip’ (Bishop, 2019) may be important sources for explaining why channel owners are doing what they do. For both featured channels and subscriptions limitations apply: channels may simply not feature or subscribe to other channels and channel subscriptions may be set to private. An even more fundamental interrogation concerns the status of ‘channel’ itself since any user who chooses to activate their channel feature is technically a channel. For our project, which is particularly interested in ‘public-facing’ channels that seek or have already found a large audience, the question was thus how to define such a channel. One way to do so would be to exclude channels that have only uploaded a limited number of videos, but since we are interested in publicness and channel professionalization, we decided on subscriber count as criterion, not least because subscriber numbers are the central component of YouTube’s tiered governance [9]: when activating an account’s channel feature, one acquires ‘graphite’ status, which comes with access to basic tools such as Creator Studio; passing the threshold of 1,000 subscribers awards ‘opal’ credentials and, more importantly, access to monetization through advertisement; moving above 10,000 subscribers to the ‘bronze’ tier gives admission to the online and offline creator community ‘YouTube Space’, to pop-up events, and tailored training opportunities; ‘silver and up’ starts at 100,000 subscribers and these ‘elite’ channels get their own partner manager and awards. As Kumar (2019) has argued, these tiers also come with less visible perks, such as prioritization in appeals against demonetization. They are thus essential components of YouTube’s platformed media system. Rather than relying on arbitrary percentage groups, both our crawling strategy and our findings are organized around these four tiers and we will refer to them frequently throughout the text.

To collect data, we implemented a breadth-first crawler in Python that interfaces with YouTube's Web-API. The script started from a single seed [10] and followed connections until no new channels were discovered.

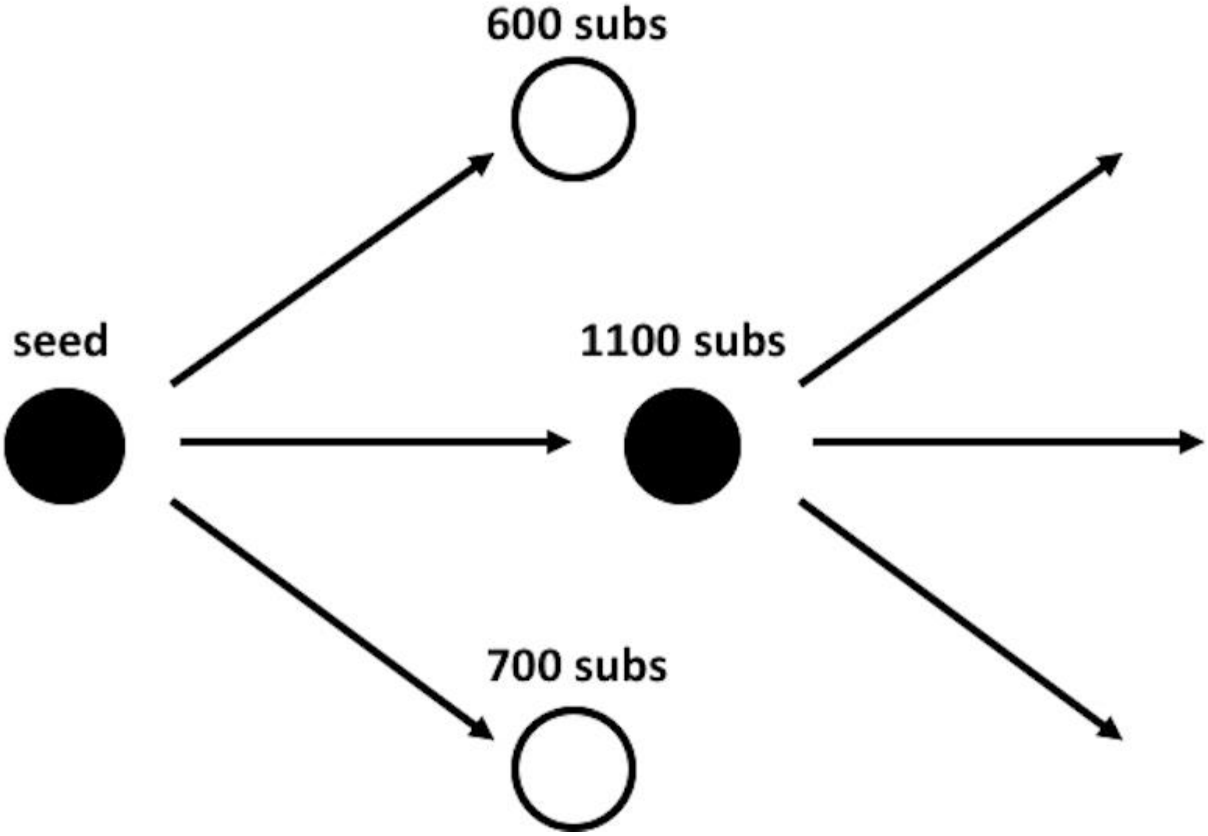

Figure 1: A schematic overview of our crawling method.

Figure 1 shows how the crawler uses featuring and subscribing connections to discover new channels and collect their metadata. If a new channel's subscriber count, which we retrieve directly from YouTube, is above the specified cutoff (1,000 subscribers), the crawler continues deeper into the network; if it is below the cutoff, the channel is added to our dataset, but its outgoing connections are not followed further. This method discovered 4,415,180 channels with more than 1,000 subscribers and a full total of 36,336,861 channels — well above existing estimates for the total number of channels on the platform (Funk, 2020).

For the three tiers making up what we call ‘monetizable’ YouTube (1k–10k, 10k–100k, and 100k+), we retrieved the listings of published videos, which include video ids and publication dates. We then gathered detailed metadata for all of the videos published by the elite 100k+ tier and took a one percent sample for the other two [11]. Table 1 provides an overview of all the collected data:

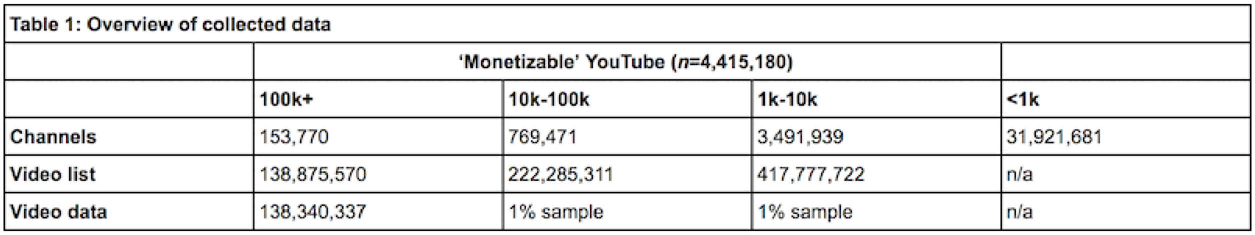

Table 1: Overview of collected data

	‘Monetizable’ YouTube (*n*=4,415,180)			
	100k+	**10k-100k**	**1k-10k**	**<1k**
Channels	153,770	769,471	3,491,939	31,921,681
Video list	138,875,570	222,285,311	417,777,722	n/a
Video data	138,340,337	1% sample	1% sample	n/a

The data collection process relied on the same API access token used for the YouTube Data Tools, which provides a quota of 50,000,000 units per day [12]. YouTube unfortunately no longer seems to issue similarly generous tokens for new research projects, which makes our research difficult to replicate. Despite the high number of requests we could make in a day, the data collection lasted from 26 November 2019 until 8 January 2020, two days before YouTube’s introduction of a special ‘made for kids’ flag [13] came into effect.

2.2. Sample qualification and limitations

Despite the substantial number of channels we were able to discover, there are serious questions about the character and coverage of our method. What did it capture and what was left out? One way to begin to answer these questions is to examine the distribution of featuring and subscribing connections. When looking at the full 4.4M channels above the 1k threshold, we find that 27.37 percent feature at least one channel, yielding an overall mean of 1.04. But it is mainly the subscription numbers that explain the density of the resulting channel network and broadly justify our approach: 37.03 percent made their subscriptions publicly available and subscribed to at least one channel, with a much higher overall mean of 53.34. Table 2 provides these numbers separately for our four tiers:

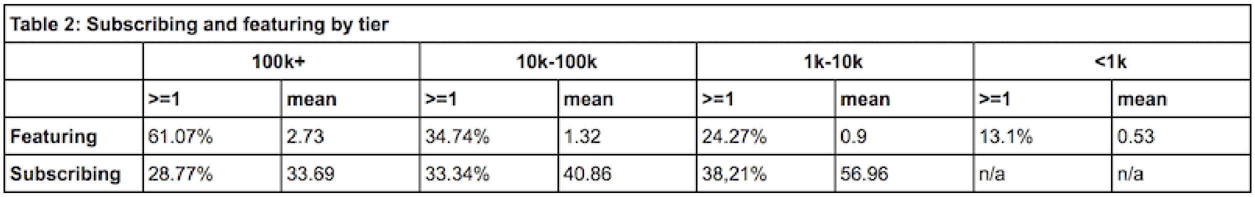

Table 2: Subscribing and featuring by tier

	100k+		10k-100k		1k-10k		<1k	
	>=1	mean	>=1	mean	>=1	mean	>=1	mean
Featuring	61.07%	2.73	34.74%	1.32	24.27%	0.9	13.1%	0.53
Subscribing	28.77%	33.69	33.34%	40.86	38,21%	56.96	n/a	n/a

Channels with higher subscriber numbers tend to feature other channels more often, but subscriptions are less readily made available. One interpretation for this is that these channels are indeed seeking to professionalize, investing more heavily in public networking and reducing the more 'private' or 'consumption' oriented practices that subscribing is indicative of. In fact, many professional creators run more than one channel to differentiate their offer and to cover broader advertising targets. These channels are then connected by featuring each other. Since motivations for featuring and subscribing to other channels are diverse and represent some mixture of strategic networking and private consumption, we do not distinguish between the two connection types in this paper, even though this may be an interesting direction for future research.

A second approach to understanding our sample consists in comparing crawls with different subscriber cutoffs. While all data used for analysis are based on the 1,000 cutoff mentioned above, we also performed crawls with a limit of 10,000 and 100,000. Table 3 lists the number of channels with more than 100,000 subscribers discovered in each crawl:

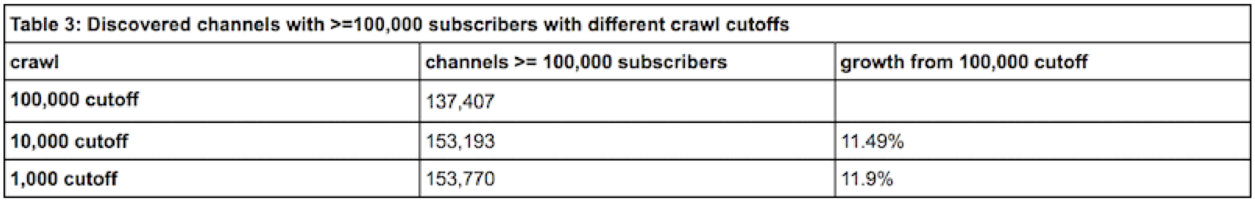

Table 3: Discovered channels with >=100,000 subscribers with different crawl cutoffs

crawl	channels >= 100,000 subscribers	growth from 100,000 cutoff
100,000 cutoff	137,407	
10,000 cutoff	153,193	11.49%
1,000 cutoff	153,770	11.9%

While moving from a 100k to a 10k cutoff grows the number of discovered elite 100k+ channels by 11.4 percent, lowering the cutoff further only adds very few channels with more than 100k subscribers to the list. This means that the 'structural holes' (Burt, 1992) that limit the crawl on one level are 'filled-in' by lowering the cutoff. The fact that lowering the cutoff further adds very little, together with the high subscribing count mentioned above, makes us confident that our method was indeed able to discover virtually all channels above 100k subscribers. If we repeat the same exercise for channels with more than 10k subscribers, a similar pattern emerges:

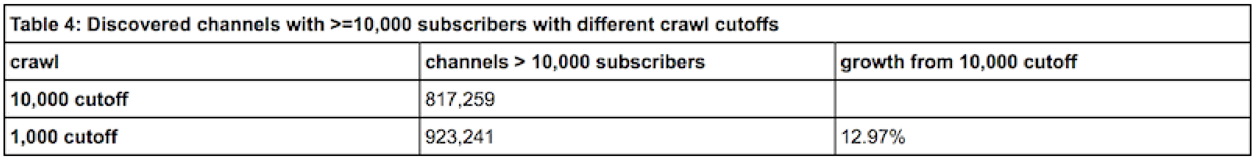

Table 4: Discovered channels with >=10,000 subscribers with different crawl cutoffs

crawl	channels > 10,000 subscribers	growth from 10,000 cutoff
10,000 cutoff	817,259	
1,000 cutoff	923,241	12.97%

While the growth in channels from lowering the cutoff is not exactly the same as before, it is sufficiently close to hypothesize that the structural proportions are similar, meaning that our 1k cutoff crawl was indeed able to discover close to all 10k+ channels. Extending this logic further, we can estimate that we are missing about 10–15 percent of channels in our 1k–10k dataset. We thus end up with a *hierarchy of confidence* in the results: while the 100k+ and 10k–100k datasets are near complete, the 1k–10k dataset needs to be interpreted with prudence and the almost 32M channels below 1,000 subscribers with even more caution. While we did attempt to run a crawl without cutoff, the resulting numbers prompted serious problems with quota limits and led further into spaces where ‘user’ and ‘channel’ become hard to distinguish. Since we are mostly interested in ‘public-facing’ YouTube, we settled on the 1,000 subscriber cutoff as a workable compromise.

Several limitations with our approach and dataset cluster around *time* . First, the fact that collecting the data took over six weeks means that our snapshot does not capture channel and video data at the same moment, manifesting in different ‘video list’ and ‘video data’ numbers in Table 2 . We made sure, however, that the start of our crawl serves as the cutoff for incoming videos and our macro-scale approach is less sensitive to small variations. Second, numbers are cumulative, and we lack historical data for virtually all variables. Just because a video came out years ago does not mean that this is when it was viewed: an older video may well find an audience years after initial publication. A similar caveat holds for data that may have been changed over the course of a video’s or channel’s timeline. For example, descriptions or keywords may have been edited one or several times during a video’s lifetime. Third, our sample is marked by what is often referred to as ‘survivorship bias’: the channels that made it above our thresholds are indeed those that ‘made it’. While this is not a problem for the snapshot-type analyses that follow, certain historical approaches need to take stock that today’s elite today is not (necessarily) the same as yesterday’s. Most of these problems could be solved or mitigated by an approach based on regular snapshots.

2.3. Analytical approach

One of the challenges of presenting an exploratory, ‘overall’ view of a very large dataset is the choice of what to highlight and what to omit. Many more analytical directions than what we present in the following sections would have been possible. As a basis for future work, we have focused on providing relatively broad overviews that will hopefully be useful for other researchers as a point of reference, for example to situate their own samples within the larger platform ecology. Section 3.1 addresses the question of stratification and hierarchization in YouTube and presents quantitative descriptions in two main ways: comparing the four YouTube tiers gives us a general idea of the differences within YouTube’s institutionalized creator hierarchization; and statistical description of variables like views and comments highlights hierarchization in more continuous terms. Section 3.2 examines channel categories as a fundamental means to segment channels along topic areas, providing insights into overall content distribution as well as the differences between them. Section 3.3 applies similar analytical means to channels’ country affiliations, investigating the complex forms of media globalization as they play out within YouTube. In both of these latter sections, we highlight differences in volume, but also investigate the subtle inequalities that manifest when intersecting these broad analytical directions with the different variables present in our dataset.

Taken together, these three sections seek to step closer to our overall research goal, that is, a broad quantitative characterization of the platformed media system emerging on YouTube, with a focus on the ‘protoindustry’ and its quest to professionalize. To keep the following at least somewhat accessible to our readers, we not always compare all four tiers (100k+, 10k–100k, 1k–10k, <1k), but often focus on the particularly interesting ‘elite’ (100k+) segment or, in other instances, analyze the full 4.4M channels that comprise the top three tiers and make up what we have called ‘monetizable’ YouTube. In the spirit of collegiality and reproducibility, we make our data partially available, allowing researchers to dig deeper themselves (cf. Appendix).

In terms of analytical methodology, we rely mostly on descriptive statistics and visualizations. This is not a quip against more mathematically involved forms of analysis, but a concession to both the complexity of our dataset and our exploratory goals. We feel that more complex forms of multivariate analysis would require a stronger subject focus and analytical directionality than we adopt in this paper. One of the outcomes of our inquiry — in line with our inductive outlook — is indeed a call for follow-up research, which we highlight at various points in the text. We hope to address some of these questions ourselves in future publications.

3. Findings

3.1. Channel overview and stratification

In this section, we provide a quantitative overview of YouTube channels along a number of standard and derived metrics. The guiding research interest, here, is to scope the channel landscape, to provide reference points for YouTube scholars (*e.g.* , to localize channels within the hierarchy), and to understand stratification in terms of views, subscribers, videos, and user reactions. While we cannot establish clear causalities, we pay attention to 'success factors', that is, characteristics that distinguish channels that are doing well from those that do not.

Channel overview

The main structuring unit on YouTube is the channel, which serves not only as a 'binder' for videos and playlists, but also to build more stable, non-algorithmic audiences through subscription. In line with common assessments of social media as highly unequal in terms of views, followers, shares, or other metrics, our dataset is heavily dominated by 'elite' channels (100k+ subscribers). As Table 5 shows, these channels accrue most of the views and subscribers despite generating only 8.9 percent of all the published videos by the four tiers. What emerges, here, is the dominance of the most popular channels in terms of visibility and popularity and the vigorous creative effort of the user base that does not obtain the same reward, a trend already observed in previous literature (Bärtl, 2018).

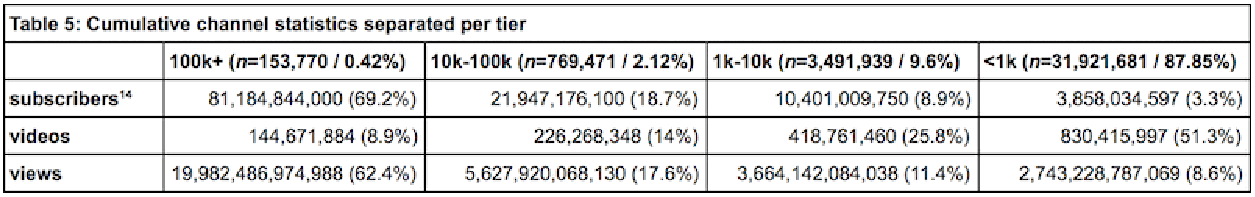

Table 5: Cumulative channel statistics separated per tier

	100k+ (*n*=153,770 / 0.42%)	10k-100k (*n*=769,471 / 2.12%)	1k-10k (*n*=3,491,939 / 9.6%)	<1k (*n*=31,921,681 / 87.85%)
subscribers[14]	81,184,844,000 (69.2%)	21,947,176,100 (18.7%)	10,401,009,750 (8.9%)	3,858,034,597 (3.3%)
videos	144,671,884 (8.9%)	226,268,348 (14%)	418,761,460 (25.8%)	830,415,997 (51.3%)
views	19,982,486,974,988 (62.4%)	5,627,920,068,130 (17.6%)	3,664,142,084,038 (11.4%)	2,743,228,787,069 (8.6%)

This pattern continues further up the top: looking at the 15,496 channels that have more than 1M subscribers (0,04 percent), we found that they account for 37 percent of the subscribers, 37.4 percent of the views, but only 2.3 percent of the videos published. This is a much heavier skew than the 80/20 Pareto principle [14] and a clear indication that YouTube's elite is central to the life — and earnings — of the platform. Overall, subscriber and view count correlate strongly (0.74) and we can easily imagine a mutually reinforcing dynamic that is further exacerbated by YouTube’s algorithmic visibility management.

Table 6 provides more detailed statistical descriptions that add nuance by highlighting the considerable internal variation *within* our four tiers. Subscribers and per-channel views drop significantly when moving down the subscriber tiers. This is particularly visible if we move below the monetization threshold, where the average subscriber number drops to 122. When it comes to published videos, however, we see that the elite may receive disproportionate levels of exposure but have also published, on average, many more videos per channel (940) than the next lower tier (294). And this is not simply an effect of having been around for a longer time: if we compare the average number of days active since the starting date, the numbers are surprisingly constant between tiers.

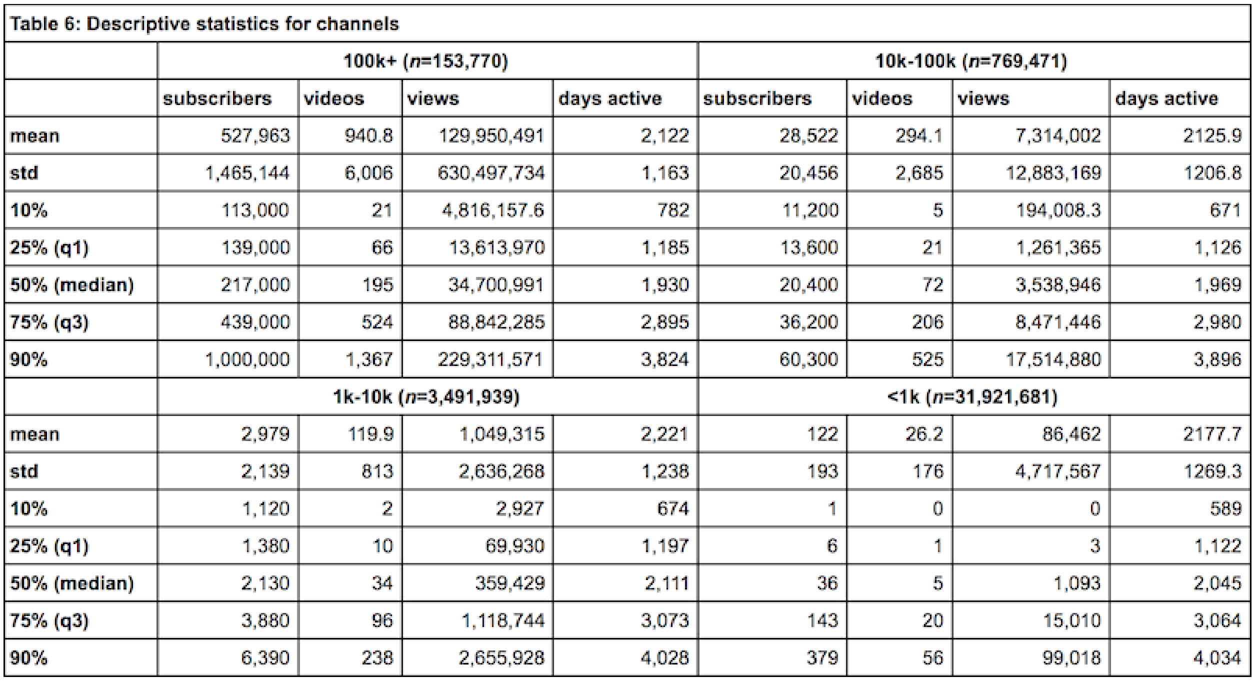

Table 6: Descriptive statistics for channels

	100k+ (*n*=153,770)				**10k-100k (*n*=769,471)**			
	subscribers	videos	views	days active	subscribers	videos	views	days active
mean	527,963	940.8	129,950,491	2,122	28,522	294.1	7,314,002	2125.9
std	1,465,144	6,006	630,497,734	1,163	20,456	2,685	12,883,169	1206.8
10%	113,000	21	4,816,157.6	782	11,200	5	194,008.3	671
25% (q1)	139,000	66	13,613,970	1,185	13,600	21	1,261,365	1,126
50% (median)	217,000	195	34,700,991	1,930	20,400	72	3,538,946	1,969
75% (q3)	439,000	524	88,842,285	2,895	36,200	206	8,471,446	2,980
90%	1,000,000	1,367	229,311,571	3,824	60,300	525	17,514,880	3,896
	1k-10k (*n*=3,491,939)				**<1k (*n*=31,921,681)**			
mean	2,979	119.9	1,049,315	2,221	122	26.2	86,462	2177.7
std	2,139	813	2,636,268	1,238	193	176	4,717,567	1269.3
10%	1,120	2	2,927	674	1	0	0	589
25% (q1)	1,380	10	69,930	1,197	6	1	3	1,122
50% (median)	2,130	34	359,429	2,111	36	5	1,093	2,045
75% (q3)	3,880	96	1,118,744	3,073	143	20	15,010	3,064
90%	6,390	238	2,655,928	4,028	379	56	99,018	4,034

The picture that emerges from these numbers is the existence of a preeminent and professionalized elite that has the resources to produce more content, succeeds at gaining views and subscribers, and, consequently, gains more income via monetization. This dynamic continues into the next section.

Channel network properties

Looking at the population of channels as a network produces further insights — and questions — into the dynamics of success. The relational structure emerging from our crawl, shows a relatively well-connected network. For the 4.4M channels above 1k subscribers, where we have complete linking data, there are 162,502,327 directed edges, an average of 36.8 per node. The channels with the most incoming links are not just YouTube stars like *PewDiePie* or *Eminem* , but also categories like Music that serve as video aggregators within YouTube's information architecture rather than as 'real' channels and therefore miss video and view numbers. Interestingly, the popularity of channels like *NoCopyrightSounds, TeamYouTube [Help]* , or *YouTube Creators* shows the relevance of material and advice for the established and budding creators in our sample. The prominence of these channels is an indicator for the importance of copyright on YouTube, but also for the aid the platform provides to creators seeking to professionalize. Table 7 shows the ten most linked channels:

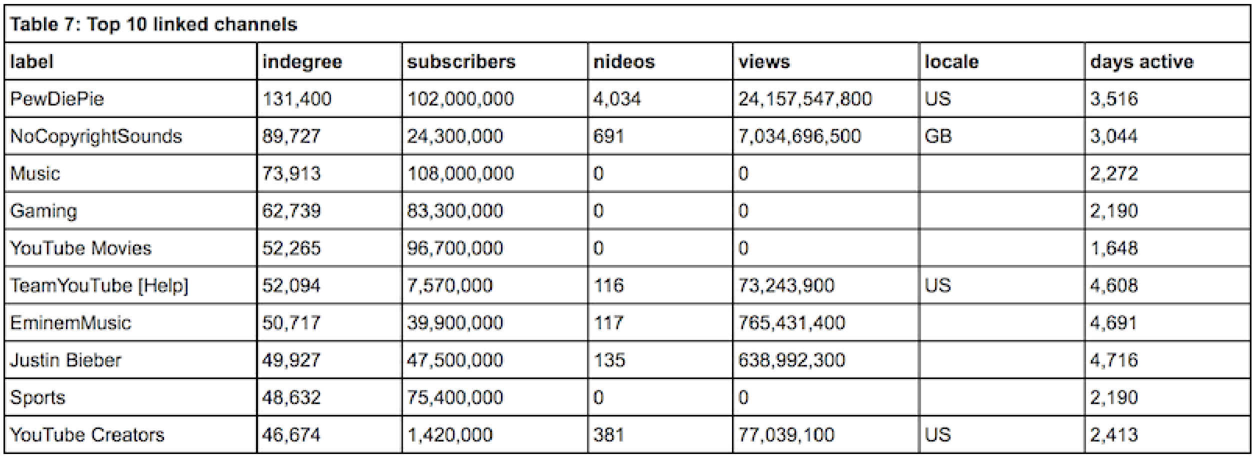

Table 7: Top 10 linked channels

label	indegree	subscribers	nideos	views	locale	days active
PewDiePie	131,400	102,000,000	4,034	24,157,547,800	US	3,516
NoCopyrightSounds	89,727	24,300,000	691	7,034,696,500	GB	3,044
Music	73,913	108,000,000	0	0		2,272
Gaming	62,739	83,300,000	0	0		2,190
YouTube Movies	52,265	96,700,000	0	0		1,648
TeamYouTube [Help]	52,094	7,570,000	116	73,243,900	US	4,608
EminemMusic	50,717	39,900,000	117	765,431,400		4,691
Justin Bieber	49,927	47,500,000	135	638,992,300		4,716
Sports	48,632	75,400,000	0	0		2,190
YouTube Creators	46,674	1,420,000	381	77,039,100	US	2,413

While, unsurprisingly, indegree and subscriber count correlate quite strongly (0.793), absolute numbers for subscribers are vastly higher. We must be conscientious that the sample under scrutiny is only a small part of the entire 'YouTube network', understood as the site's full user base, including those who have not activated their channel feature. Indegree, here, captures the visible part of subscribing within our network, but the actual subscriber numbers reflect the activities of a much wider range of users. Using the powerlaw package for Python (Alstott, *et al.* , 2014), we investigated the distributions of both variables statistically.

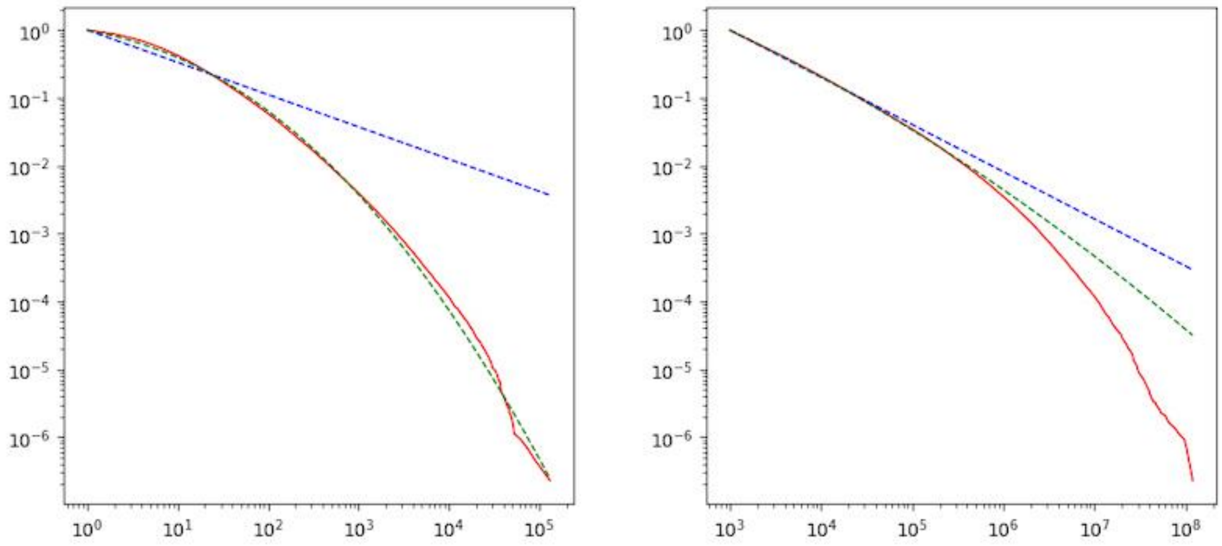

Figure 2: Both sides plot a Complementary Cumulative Distribution Function (CCDF), the left x-axis for indegree, the right for subscribers. The y-axis shows rank. The blue lines represent a power law fit and the green a log-normal fit. In both cases, Xmin is set to the lowest value in the dataset (1 for indegree, 1000 for subscribers).

While Figure 2 raises more questions than it answers, it indicates that neither variable follows a simple power law. As we have seen in the previous section already, the broad logic of the rich getting richer still applies (Borghol, *et al.* , 2012), making the road to visibility and success harder for new creators. But the fact that a log-normal distribution is overall a better fit for both variables, in particular for indegree, indicates that the growth dynamics at play cannot be easily mapped onto a singular process [15]. Many different factors may come into play: social capital transfers across sites (*e.g.* , when a celebrity opens a YouTube channel), ranking and recommendation algorithms affect topic-specific and overall visibility, and so forth. While advertising revenue per view can vary widely between topic domains and countries, there comes a point where a creator may be able to quit or reduce their ‘day job’, allowing them to intensify their publishing schedule and channel growth. Investing actual growth mechanisms and thresholds in more depth is beyond the scope of this paper but would clearly be a worthwhile endeavor.

Channel age distribution

The question how strongly channel age affects success is another element in the platform puzzle. While we can confirm Bärtl’s (2018) finding that there is a statistically significant correlation between channel age and success indicators, it is relatively small at 0.082 for subscriber count and 0.101 for view count — and this is focusing on the elite only. For the full dataset, these values go down to very low levels at 0.009 and 0.003 respectively. To switch perspective, Figure 3 presents the age distribution as percentage histogram for each tier.

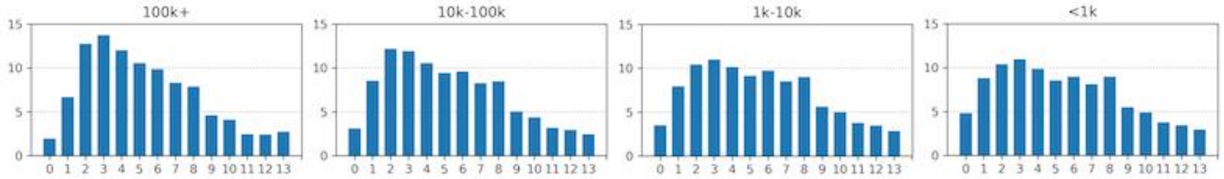

Figure 3: Percentage histogram showing distribution of channel age in years since creation.

As with the days active averages in Table 6 , the differences are not striking: in all four cases, the most common channel age was two or three years. Moving down the tiers, however, we find a ‘flattened’ curve, that is, a higher percentage of both older and younger channels. This cannot be taken as an indicator for greater longevity of less popular creators: despite the bias toward popularity, our dataset includes abandoned spaces in which no content has been published for a long time. Indeed, out of the 153,770 channels in the elite tier, 152,681 (99.29 percent) had at least one video available, but only 137,605 (91.27 percent) featured videos created in 2019. There is an entire YouTube ‘cemetery’ of inactive channels hidden in our data that would merit further investigation.

Video creation tactics

One of the characteristics that have been distinguishing YouTube from other social media sites since 2007 already, is the decision to share advertising revenue — a considerable US$15B in 2019 — with content creators. Interestingly, the controversies around monetization that YouTube has seen almost constantly over the past years have revolved less around the split between platform (45 percent) and creators (55 percent) but rather around algorithmic findability and demonetization. The best tactics to reach more viewers and to increase revenue are heavily discussed topics within the creator community and the opaque and often-changing technical and administrative mechanisms have caused much frustration. This has led many creators to seek other sources of revenue, from crowdfunding to selling merchandise. Unfortunately, there is no automated way to know whether a video or channel has been demonetized, but there are at least three directions to investigate the effects of the 'algorithmic dance' (Kumar, 2019) creators have to engage in to succeed.

First , publishing frequency is considered to be one of the cornerstones of a successful channel strategy and uploading videos often and on a regular schedule is seen as essential to achieving visibility. The pressure this exerts became particularly visible in 2018 when a number of well-known creators had to pause due to burnout (Alexander, 2018). Figure 4 shows the evolution of monthly publication averages per channel, taking into account only channels that published at least one video in a given month. We chose this method of normalization to correct for the successive addition of channels to the elite tier over time and for the sometimes long periods of inactivity for certain channels.

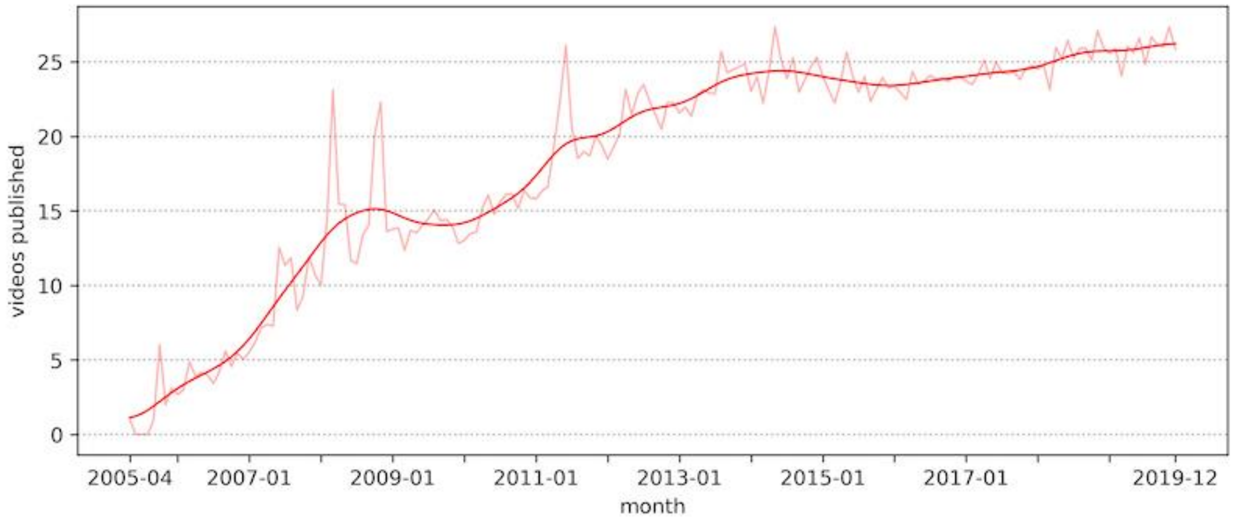

Figure 4: Per-channel publication averages for active elite channels, Gaussian smoothing (sigma=4) added for readability.

Figure 4 shows that publication activity grew steadily until around mid-2014, then taking a slight dip before starting to grow again in 2016. The high variation in our dataset means that these results need to be taken with a grain of salt, but the overall trend toward increased publishing frequency is clear.

Second , the ‘optimal’ length of videos has also received much attention in creator communities. This concerns not only the question of how to handle users’ attention spans, but again relates to estimations of algorithmic preference and advertisement. Although YouTube does not specify a minimum video length for ad eligibility, videos that are longer than 10 minutes can place so-called ‘mid-roll’ ads [16] and, according to some creators, are favored by the all-important algorithms (Peterson, 2018).

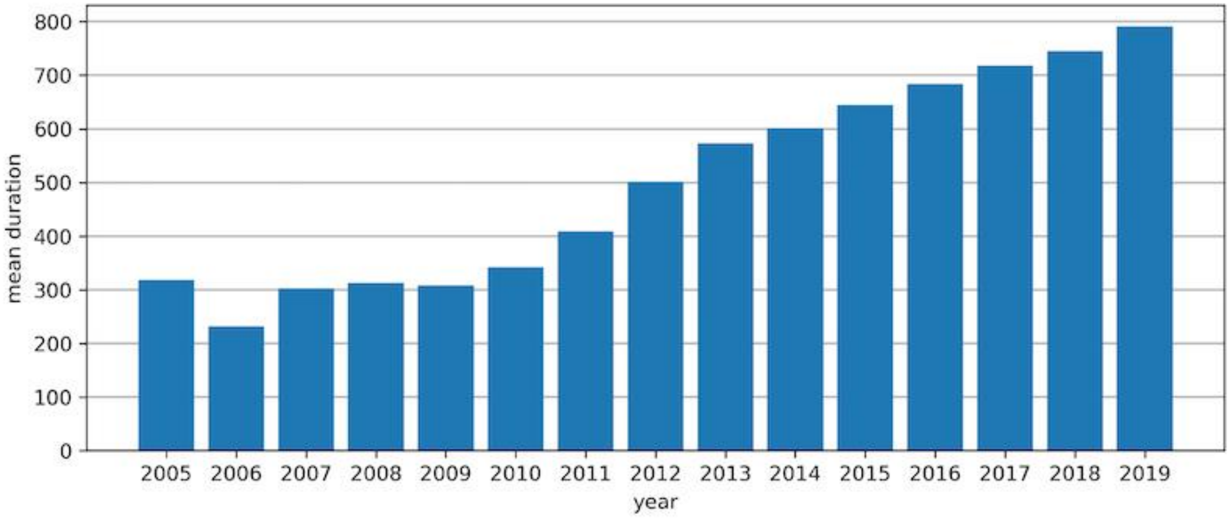

Figure 5: Average video length in seconds per year for the 100k+ channel tier.

As Figure 5 shows, average video length has more than doubled since the beginning for the elite tier, with constant yearly progress from 2011 onward. This shows adaptation to real and imaginary incentives, but also tells a story of a changing platform: while YouTube was, for a long time, considered to be mainly a home for low-effort ‘user generated content’, the trend toward more substantive videos supports the ‘professionalization’ narrative, both on the level of individual channels and for the platform itself. That said, we only detected a very low level of correlation (0.002) between video duration and view count.

Third , in a situation where advertisement rules are unstable and opaque, creators have turned to product placement, sponsorships, affiliate programs, and crowdsourcing as means to generate income. Video descriptions are increasingly important tools to direct viewers to other places on the Web. In the 138M+ videos posted by our elite tier, we found 577,737,068 URLs that represent valuable traces for the specific forms of ‘industrialization’ happening on YouTube. Besides mapping creators’ cross-platform activities, they allow us to trace the appearance and spread of crowdfunding platforms like Patreon, of affiliate links and merchandise stores, and of e-commerce Web sites like Etsy. That is, of ways creators seek to develop their channels into media businesses that are less dependent on advertising income.

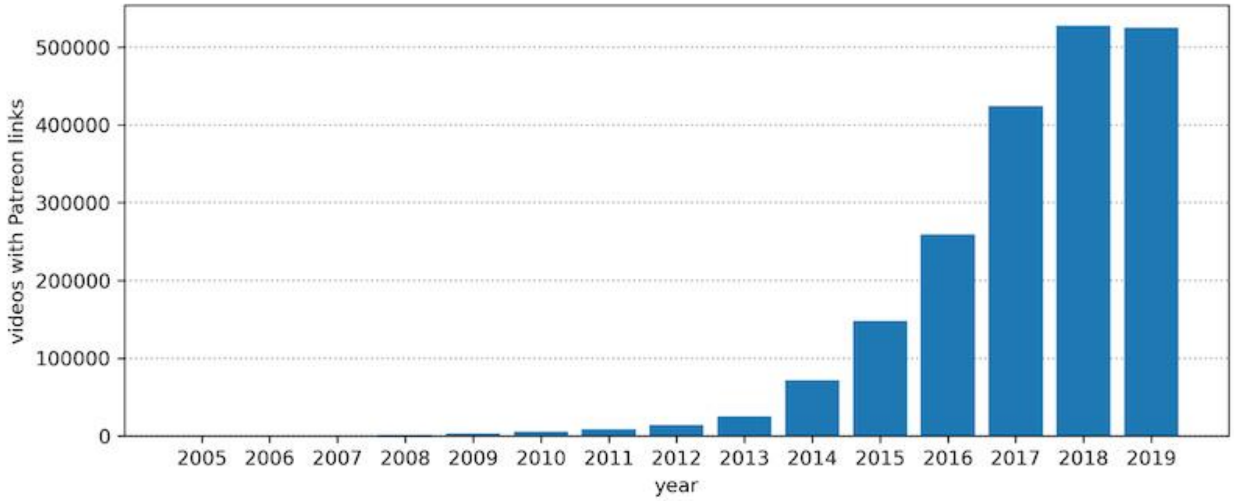

Figure 6: Videos with Patreon links in their description for the 100k+ channel tier.

Figure 6 shows the surge of Patreon links over the years. The slight dip in 2019 is explained by missing data for December, and the fact that we find the first link to the crowdfunding Web site in a video from 2005 confirms the idea that creators do adapt descriptions of older videos - Patreon was founded in May 2013. We will look more deeply into these practices, as well as visibility tactics such as keyword stuffing, in a follow-up publication dedicated to monetization and optimization.

Video overview and user reactions

Yet another way to measure channel stratification and success on YouTube is to investigate user reactions such as likes, dislikes, and comments. Table 8 gives a statistical description of the videos published by the 4.4M monetizable channels with at least 1,000 subscribers. We once more notice the uneven distribution along all variables. Duration, for example, shows that over 25 percent of published videos indeed conform to the ‘short clip’ cliché (less than two minutes), but there are also many much longer ones, reaching up to a whopping 46,043,514 second long video, which is 533 days of live feed from the International Space Station. More than half of the videos published fail to reach 550 views, despite the elite being included in this sample. Like, dislike, and comment counts are low compared to views, which suggests that interaction buttons (other than the play button) have a less central role on YouTube than on other platforms. The fact that comments generally rank higher than dislikes reinforces this observation. Another possible explanation for the centrality of views as a measure of success on YouTube is that the platform counts views from both logged-in and not logged-in users, whereas only logged-in users can interact with videos though liking, disliking, and commenting [17].

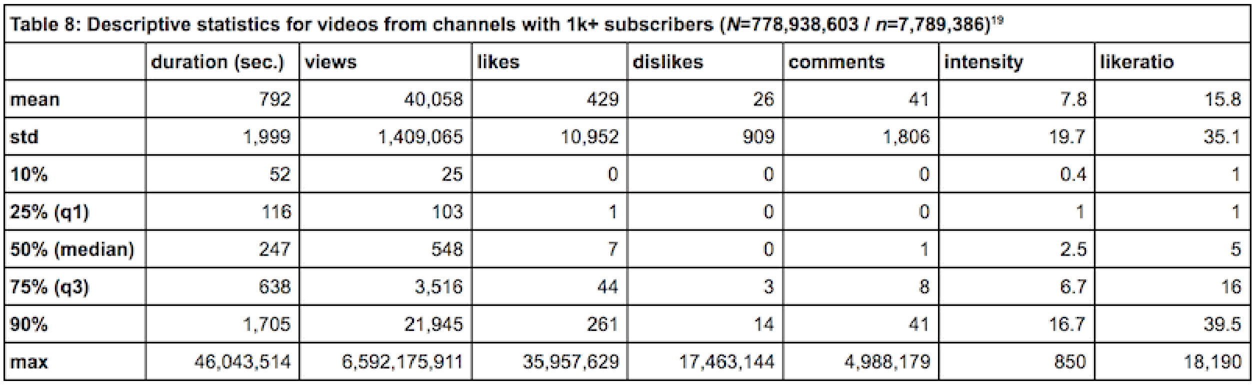

Table 8: Descriptive statistics for videos from channels with 1k+ subscribers (*N*=778,938,603 / *n*=7,789,386)[19]

	duration (sec.)	views	likes	dislikes	comments	intensity	likeratio
mean	792	40,058	429	26	41	7.8	15.8
std	1,999	1,409,065	10,952	909	1,806	19.7	35.1
10%	52	25	0	0	0	0.4	1
25% (q1)	116	103	1	0	0	1	1
50% (median)	247	548	7	0	1	2.5	5
75% (q3)	638	3,516	44	3	8	6.7	16
90%	1,705	21,945	261	14	41	16.7	39.5
max	46,043,514	6,592,175,911	35,957,629	17,463,144	4,988,179	850	18,190

To facilitate aggregate assessment, we generated [18] two additional metrics: *intensity* and *likeratio* . The first one identifies those videos able to generate more engagement per view and corresponds to the sum of likes, dislikes, and comments divided by view count and multiplied by 100 for readability. The second one highlights content that generates more controversy or negative feedback by dividing likes by dislikes. It is rare to receive more dislikes than likes, with an average likeratio of 15.8 in favor of likes for monetizable YouTube. This value goes up to 41 for the elite, adding another element to their success story.

Taken individually these metrics have certain problems: for example, videos made unavailable on copyright grounds retain their like and dislike counts, but their view count is reset to one, resulting in extremely high intensity values. These outliers can fortunately be easily identified, and we removed them from our analysis. Intensity and likeratio become particularly useful further down when intersected with channel topics and countries.

3.2. Channel categories

Categorization exists on (at least) two levels within YouTube's information architecture. When uploading a video, creators have to choose from a list of set labels that are introduced in the following way: 'Content categories organize channels and videos on YouTube and help creators, advertisers, and channel managers identify with content and audiences they wish to associate with.' [19] From these labels and possibly other data, YouTube automatically generates channel categories or topics, which are not identical to the labels used for videos and cannot be changed by creators. As already discussed, categories form navigational hubs, but they also play a role in targeted monetization. This system, as Kumar argues, 'skews the incentives for creating particular genres and types of content' [20], raising concerns regarding its impact on creativity and innovation. This section first introduces categories more broadly and then investigates how various metrics vary across topics.

Categories overview

Channel categories are divided into seven main branches. Entertainment, gaming, lifestyle, music, society, and sports each come with their own subcategories, but a seventh topic, knowledge, is not divided any further. Channels may be categorized into one or several parent topics, and they may or may not be slotted further into subtopics. A certain percentage of channels across all of our tiers were in no category at all (see Figure 8). Figure 7 presents a structural view of category relationships based on co-occurrence for the same channel.

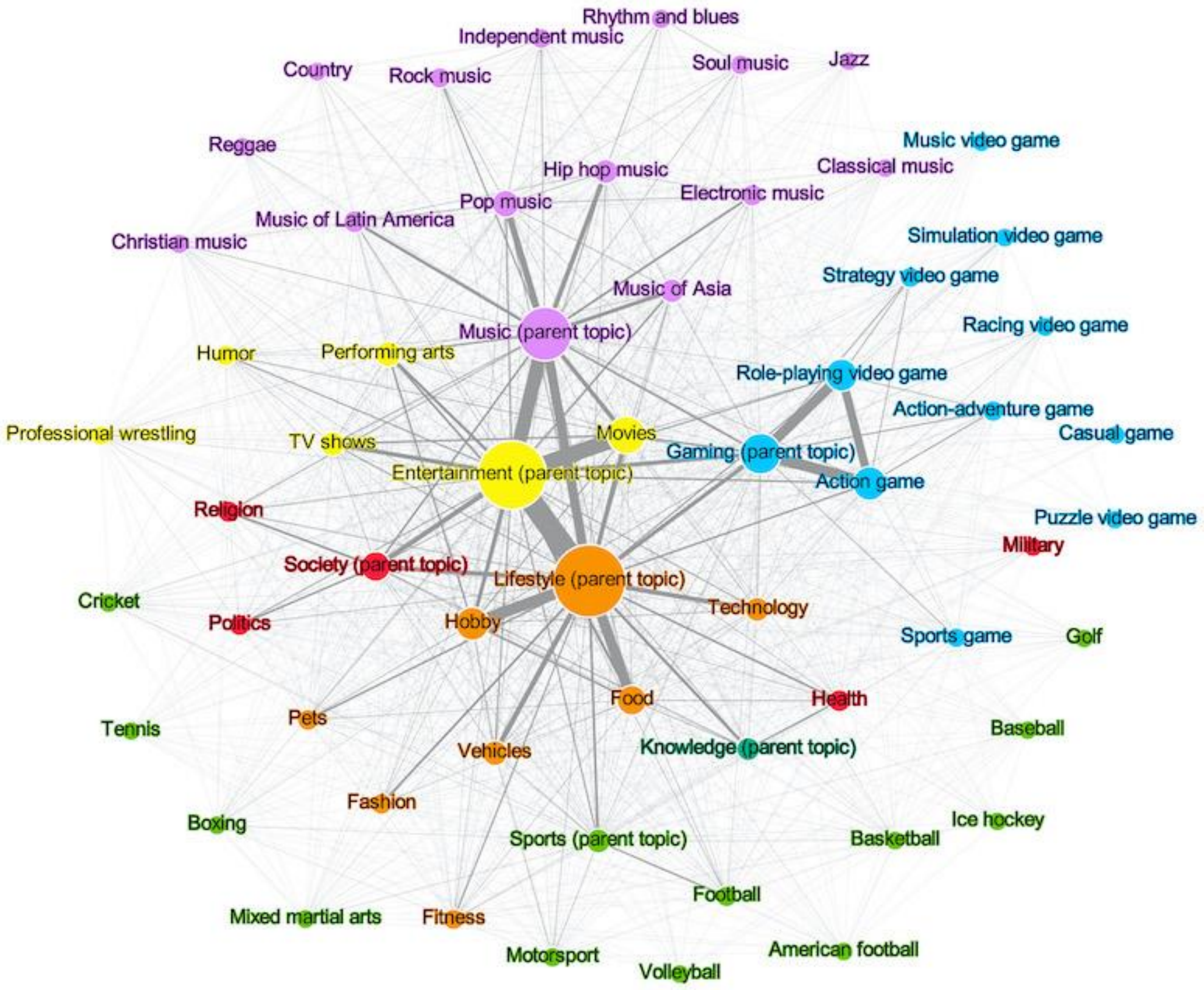

Figure 7: Network of YouTube categories (100k+ tier) based on co-occurrence; only edges with a greater weight than ten are visible.

Subcategories generally cluster close to their parent topics and they are often straightforward specializations: music is organized into genres like jazz, hip hop, or reggae; gaming into role-playing, strategy, or action game, and so forth. Entertainment, however, includes ‘professional wrestling’ as one of only five subcategories and lifestyle covers a very broad range of topics. The most striking exception is the society category, a mixed bag where traditional media outlets such as *BBC News* coexist with channels devoted to military, religion, politics, or health. Consequently, not all society subtopics are positioned close together in Figure 7 .

Connecting with recent debates on radical far-right content on YouTube, we can further problematize YouTube topics. While Politics could be considered the appropriate place for politically charged content, the category is mostly dedicated to news channels and documentary style content. These channels are grouped together into the Politics subcategory almost randomly, which has already been observed in previous literature (Paolillo, *et al.* , 2019). As a short experiment, we took the 31 channels mentioned as the core of Lewis' (2018) far-right 'alternative influence network' and examined the categorization of the 28 channels still on YouTube. Ten were not classified at all, including the the most well-known ones: Steven Crowder, Stefan Molyneux, Jordan Peterson, The Rubin Report, and Ben Shapiro. Is YouTube explicitly reducing the findability of these highly controversial channels or are these signs of demonetization? Out of the remaining 18, only five were tagged as Society and only three as Politics. The most common parent topics, in fact, were Entertainment and Lifestyle (13 each). Making sense of why channels get grouped together in this category would require more forensic analysis, again pointing toward further research. Despite issues in specific areas, categories allow for broad characterizations of what is actually available on YouTube and the following section explores them in greater depth.

Categories in numbers

To provide a basic quantitative overview, Table 9 shows cumulative numbers for channel counts, subscribers, videos, and views per channel category while Figure 8 allows for the interactive exploration of the category system, in both cases for the 153k elite channels with more than 100,000 subscribers.

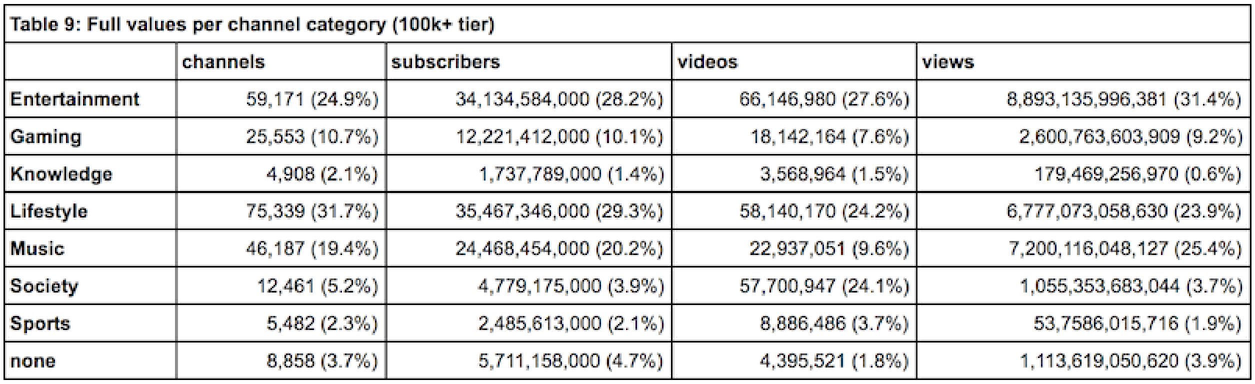

Table 9: Full values per channel category (100k+ tier)

	channels	subscribers	videos	views
Entertainment	59,171 (24.9%)	34,134,584,000 (28.2%)	66,146,980 (27.6%)	8,893,135,996,381 (31.4%)
Gaming	25,553 (10.7%)	12,221,412,000 (10.1%)	18,142,164 (7.6%)	2,600,763,603,909 (9.2%)
Knowledge	4,908 (2.1%)	1,737,789,000 (1.4%)	3,568,964 (1.5%)	179,469,256,970 (0.6%)
Lifestyle	75,339 (31.7%)	35,467,346,000 (29.3%)	58,140,170 (24.2%)	6,777,073,058,630 (23.9%)
Music	46,187 (19.4%)	24,468,454,000 (20.2%)	22,937,051 (9.6%)	7,200,116,048,127 (25.4%)
Society	12,461 (5.2%)	4,779,175,000 (3.9%)	57,700,947 (24.1%)	1,055,353,683,044 (3.7%)
Sports	5,482 (2.3%)	2,485,613,000 (2.1%)	8,886,486 (3.7%)	53,7586,015,716 (1.9%)
none	8,858 (3.7%)	5,711,158,000 (4.7%)	4,395,521 (1.8%)	1,113,619,050,620 (3.9%)

CategoriesLifestyleEntertainmentMusicGamingSocietySportsKnowledgeHobbyFoodVehiclesTechnologyFashionPetsFitnessPhysical attractiveness [Beauty]TourismMoviesTV showsPerforming artsHumorProfessional wrestlingPop musicHip hop musicMusic of AsiaMusic of Latin AmericaElectronic musicRock musicIndependent musicChristian musicSoul musicRhythm and bluesCountryReggaeClassical musicJazzAction gameRole-playing video gameAction-adventure gameStrategy video gameSports gameRacing video gameSimulation video gamePuzzle video gameCasual gameMusic video gameReligionHealthPoliticsMilitaryBusinessFootballMotorsportCricketBasketballBoxingMixed martial artsAmerican footballBaseballGolfVolleyballIce hockeyTennis

Figure 8: An interactive visualization of channel categories; hover over a category to see channel numbers and click on main categories to zoom in.

As we can see, Entertainment and Lifestyle dominate all metrics, with Music not too far behind, which is consistent with findings on popular content on YouTube since its early beginning (Paolillo, 2008; Paolillo, *et al.* , 2019). Society, Sports, and Knowledge together are smaller than (video) Gaming, which covers 10.7 percent of all channels. In terms of sheer output, however, the Society category stands out, which can be explained by the presence of large news channels that publish many videos per day, each receiving few views. Since YouTube changed channel categories in 2017, the results are no longer easily comparable with older data, but Bärtl (2018) already found that the ‘News&Politics’ category accounted for 45 percent of uploads overall. Figure 8 shows that most channels are not classified into child categories, which may indicate that creators avoid focusing on too specific niches. Comparing the four tiers in our sample shows some interesting results.

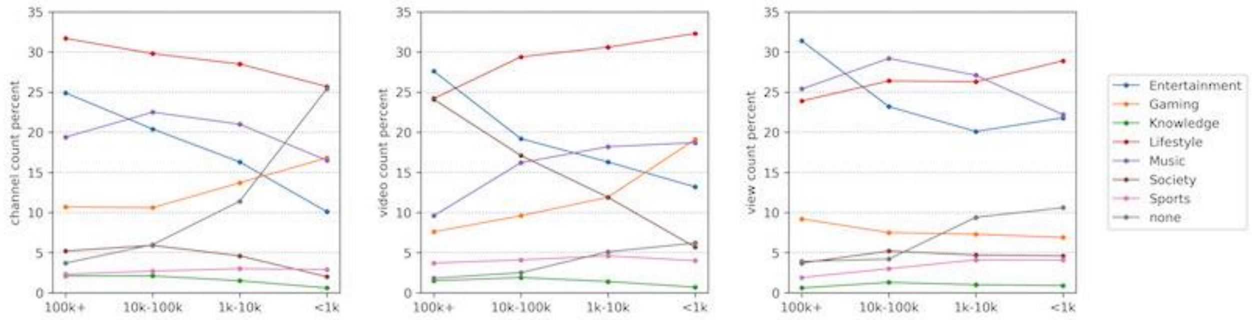

Figure 9: Comparison between channel tiers for channel count, video count, and view count.

The rise of unclassified channels when going down subscriber tiers accounts for much of the drop for all three metrics in Figure 9 , although their effect on videos and views is lower than pure channel count. The decay in metadata quality again resonates with the idea that the elite invests more heavily in optimization tactics. Gaming, however, comes up strong in channel count and videos produced in lower tiers, suggesting that many creators in this category are working hard to ‘make it’ — with rather disappointing results when looking at views. Achieving success is hard in an increasingly crowded market and many abandoned channels in the YouTube ‘cemetery’ are from gaming. A different phenomenon can be observed in the Lifestyle category: channel count remains almost stable over tiers, but both video production and views rise. Here, we find increasingly specialized hobbyist niches that creators pursue more successfully than in Gaming. Society, finally, shows an interesting drop in video production, which confirms that this category is dominated by established television news channels, which are generally part of the upper tiers.

Table 10 again focuses on elite channels and moves the focus to per-video averages. Music stands for the most popular category in terms of views and likes (including likeratio), followed by Gaming and Entertainment. While budding video gamers are struggling, elite creators such as PewDiePie are doing extremely well, including in terms of user reaction intensity. On the other side of the spectrum, Society again stands out with the lowest per-video view count and likeratio. Based on previous research (Bärtl, 2018), most viewed categories change over time, although Music, Gaming, and Entertainment seem to always maintain top positions in the ranking throughout YouTube’s history.

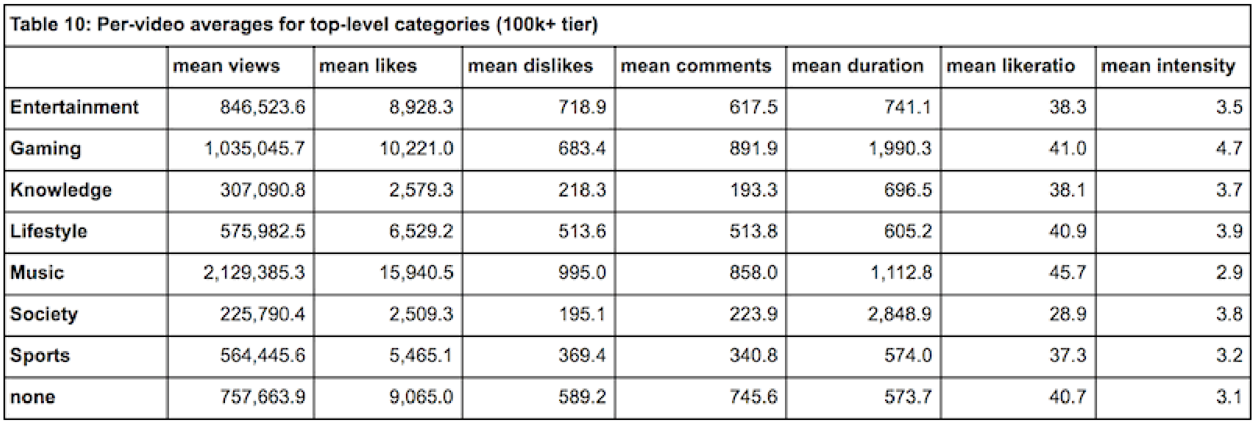

Table 10: Per-video averages for top-level categories (100k+ tier)

	mean views	mean likes	mean dislikes	mean comments	mean duration	mean likeratio	mean intensity
Entertainment	846,523.6	8,928.3	718.9	617.5	741.1	38.3	3.5
Gaming	1,035,045.7	10,221.0	683.4	891.9	1,990.3	41.0	4.7
Knowledge	307,090.8	2,579.3	218.3	193.3	696.5	38.1	3.7
Lifestyle	575,982.5	6,529.2	513.6	513.8	605.2	40.9	3.9
Music	2,129,385.3	15,940.5	995.0	858.0	1,112.8	45.7	2.9
Society	225,790.4	2,509.3	195.1	223.9	2,848.9	28.9	3.8
Sports	564,445.6	5,465.1	369.4	340.8	574.0	37.3	3.2
none	757,663.9	9,065.0	589.2	745.6	573.7	40.7	3.1

Differences in duration across topics are also salient. While most of the categories stray not too far from the ten-minute mark, videos in the Society channel category have an average duration of 45 minutes and those in Gaming an average of 30. The news channels we find in the first category often use YouTube to live stream their full program, leading to video ‘containers’ with long duration in the data. Gamers also stream live on YouTube and normal gameplay videos are often long as well. The somewhat longer duration of music videos can be explained by recordings of live concerts.

While these results are not necessarily surprising, they tell a complex story of differentiation on YouTube, where the quest for viewers can play out quite differently over different topics. Further research could investigate more deeply into subtopics and behind many of the numbers shown here lurks the question of how they connect to optimization and monetization, including the much-discussed issue of ‘advertiser-friendliness’.

3.3. Countries and their relationships

While platforms like YouTube distribute the same interfaces and features to worldwide audiences, the way these are appropriated can differ substantially. The interplay between global accessibility and local embeddedness is therefore one of the most interesting lines of inquiry concerning social media. Within YouTube’s information architecture, countries are managed via ISO 3166–1 alpha–2 country codes. These codes mainly received scholarly attention as means to restrict user access to particular videos in particular areas due to licensing rights or compliance with local laws. (*e.g.* , Lobato and Meese, 2016) And indeed, when looking at the videos posted by the 4.4M channels of ‘monetizable YouTube’, 3.58 percent have some kind of geographic restriction [21]. This number grows to 4.07 percent for the elite, where we found most restricted videos to be in the Music and Entertainment categories. A more detailed analysis of these videos and their respective channels could be the focus of further research.

On a more general level, channel owners have the option to choose a country flag to signal geographic affiliation or to leave the field empty. As with most channel settings, the consequences of this choice remain unclear, although some observers consider that it may help with visibility in the selected locale. From a research perspective, the country field provides insight into the geographic distribution of channels, videos, and reception numbers, and also allows for a deeper analysis of relations between countries. The following two sections address these aspects in turn, showing how our data can contribute to the study of new forms of media globalization that are neither flat, nor simple extensions of traditional hierarchies (Cunningham and Craig, 2016).

Countries overview

A first approach sheds light on the relative presence of countries on YouTube. Since proportions did not vary dramatically over the four tiers, we again focus on the elite in what follows. There is one interesting caveat, however: as we have seen with other metadata elements, channels with higher subscriber numbers take greater care in building their profiles: only 23.4 percent of the elite sample is missing a country flag, going up to 78.1 percent for the full 36M dataset. While the ‘n/a’ group shows proportionally higher view counts for the top tier (Figure 10), it falls behind on lower tiers. This suggests that for elite channels, not selecting a country can be understood as the ambition to target an international audience, while for less popular channels, this is a sign for less attention paid to metadata quality.

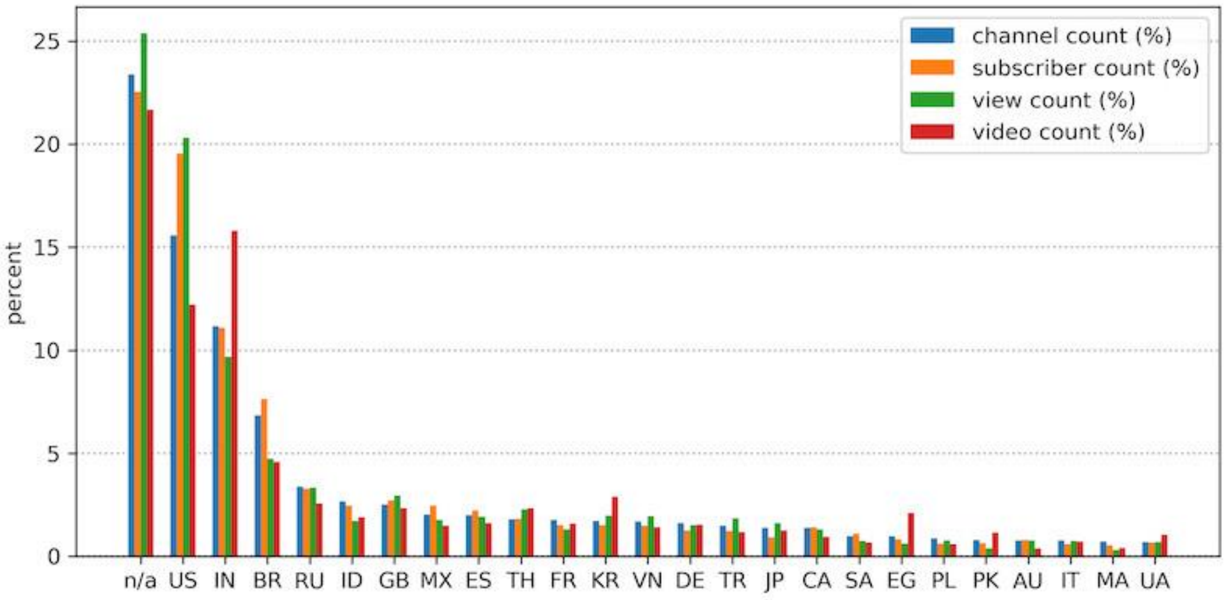

Figure 10: Per-country statistics for the elite tier.

Figure 10 shows a first indication of the dominance of United States (US) channels, which account for more than 15 percent of the elite tier, followed by Indian (IN) and Brazilian (BR) YouTubers. Russia (RU), Indonesia (ID), Great Britain (GB), Mexico (MX), Spain (SP), and Thailand (TH) continue the list of the most prolific countries. What Figure 10 also shows, however, are the relative proportions of subscribers, views, and videos published. Here we see that the US is ‘punching above its weight’ in terms of subscribers and views, achieving greater success with fewer videos published. Only Great Britain achieves a similar feat, pointing toward the crucial advantage the English language confers. Countries like India, South Korea (KR), and Egypt (EG) publish a disproportionate number of videos, although only Korea is able to translate that work into some level of success in terms of views, probably due to the international appeal of K-pop, which has been a popular music genre on YouTube since the platform’s early years (Paolillo, 2008). Countries like Brazil and Indonesia stand out for achieving far fewer views than their overall share of channels would suggest. These patterns remain visible when shifting to per-channel averages:

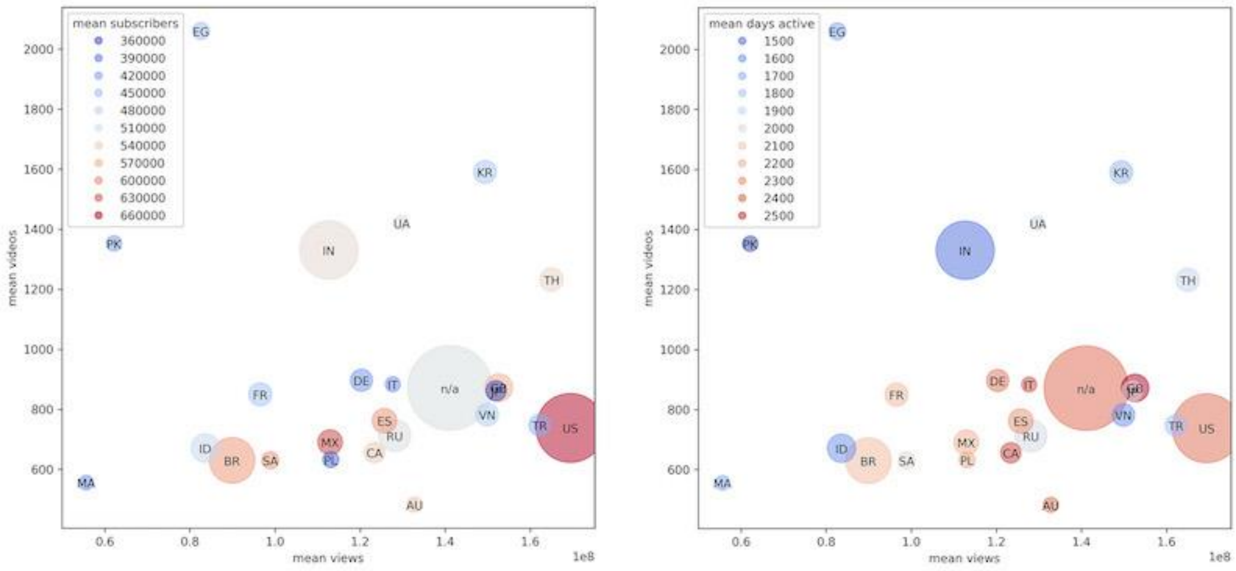

Figure 11: The top 25 countries per number of channels in the elite tier, showing average view count (x-axis), average video count (y-axis), and channel count (size). On the left, color shows average subscriber count and on the right average channel age.

In Figure 11 , we can observe the strength of the US on a per-channel basis, with the highest average subscriber and view counts, as well as the ability to reach these numbers with lower average video counts. These are clear indicators of successful industrialization. We generally observe a dividing line between countries with channels that produce many videos, for example Egypt, South Korea, Ukraine (UA), and India, and everyone else. In addition to their disproportionate success in terms of views (Figure 10), Thailand and Turkey (TR) also stand out with a high level of average views per channel, which signals strong uptake of YouTube in their respective locales. Looking to the right scatter plot in Figure 11, we notice the higher average age of channels from western countries, with mostly Asian countries like India, South Korea, Pakistan (PK), and Vietnam (VN) at the other end of the spectrum, rendering visible the expansion process of YouTube as a truly global platform.

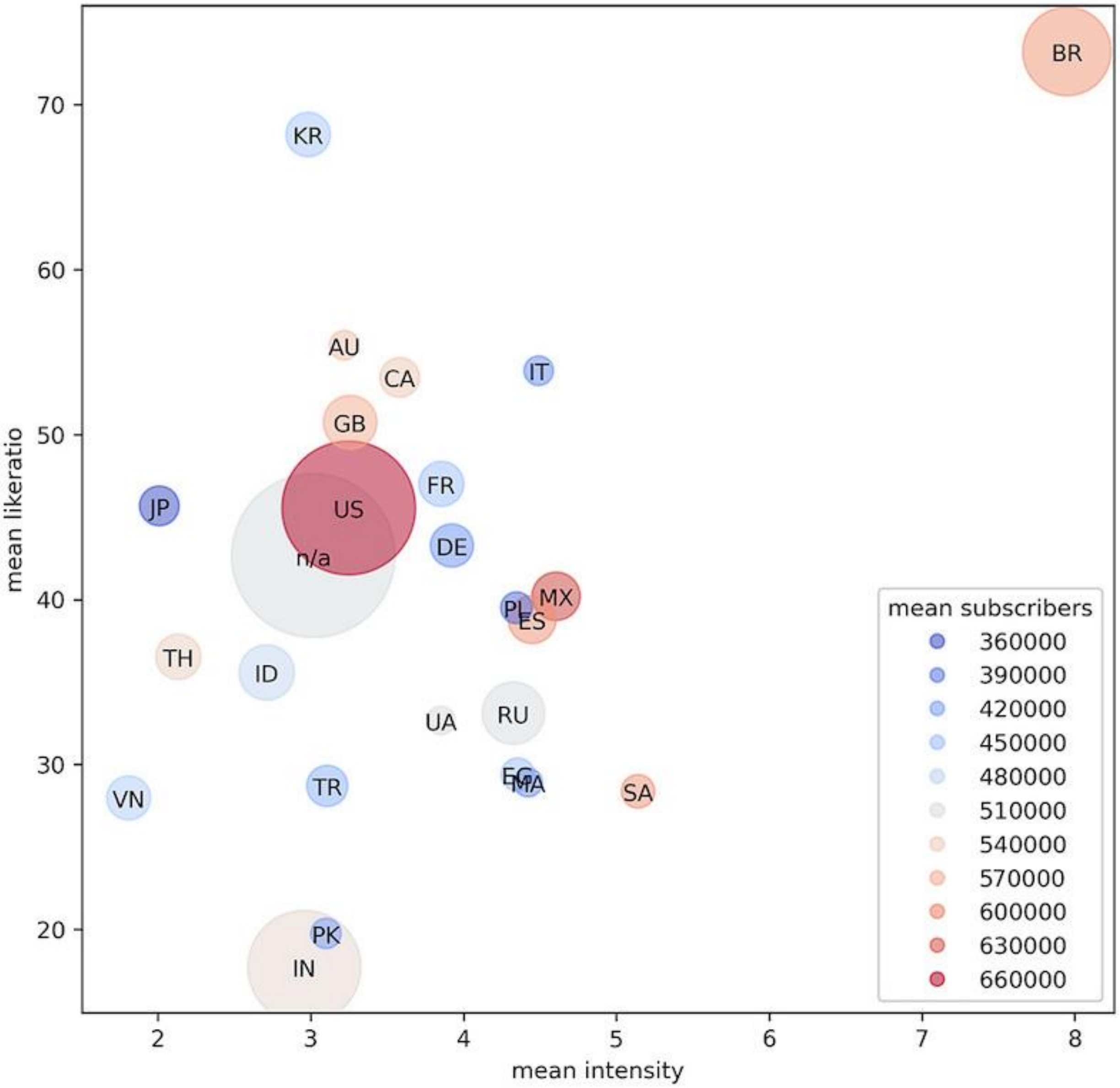

Figure 12: The top 25 countries per number of channels in the elite tier, showing average intensity (x-axis), average likeratio (y-axis), channel count (size), and average subscriber count (color).

Figure 12 shows countries from yet another angle, using intensity and likeratio on the two axes. Pakistan and India stand out as having particularly low likeratios, indicating disapproval or controversy, while South Korea and Brazil lead the other end of the spectrum. Brazil also separates itself with the highest reaction intensity by far, confirming yet again how important social media are for the country's communicative makeup.

Relationships between locales

Our dataset being not just a population of channels but also a network, we can further investigate the relationships between locales. A first approach is a basic visualization of the channel network.

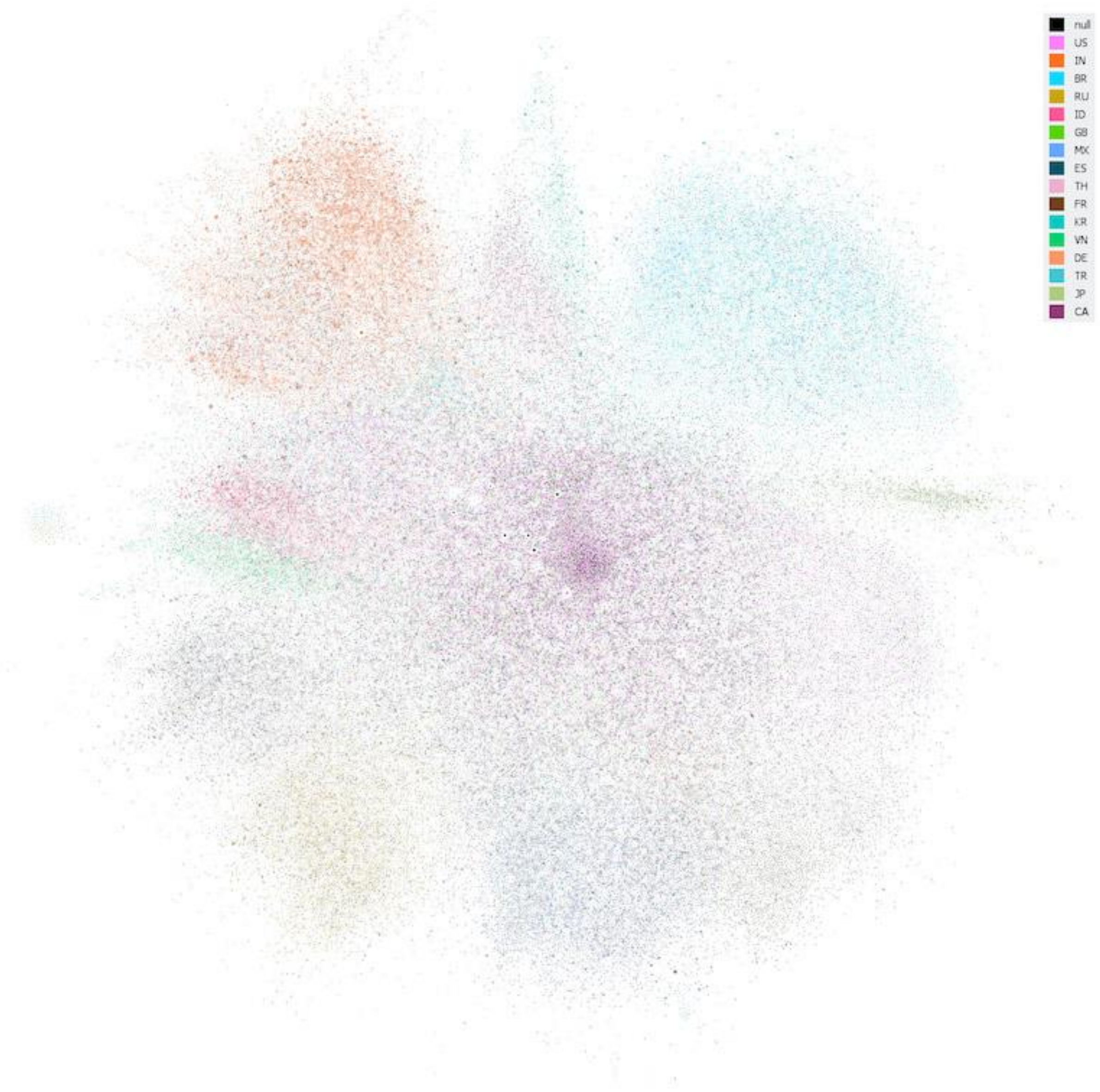

Figure 13: YouTube channel network representation of the giant connected component formed by the 100k+ elite channels (145,117 channels, 2,572,163 edges), size indicating subscriber count.

While visualization cannot replace more involved forms of analysis, the visual interpretation (Figure 13) of the network, made with gephi (*cf.* , Bastian, *et al.* , 2009), yields a number of interesting results. First, looking at the clear clustering around country flags, cultural affinity remains the strongest relational force on YouTube. Second, channels from English-language countries form the topological core, confirming the central place of the United States in particular. Third, language plays an important connective role beyond English, for example for Spanish, where Spain and Latin American countries are closely connected, again confirming previous findings (Paolillo, *et al.* , 2019). Fourth, while several countries/languages form identifiable clusters, Brazil in particular is more detached.

A more structured analysis becomes possible if we transpose the network by fusing locales into single nodes connected by the sum of their respective channel connections. This allows us to bring the full network of monetizable YouTube (4.4M channels with at least 1,000 subscribers) into view. For increased clarity, we limit each country's connections to the top three most linked to (which may include itself) and focus on locales that have at least 100 channels. This leaves us with 150 countries and 450 connections.

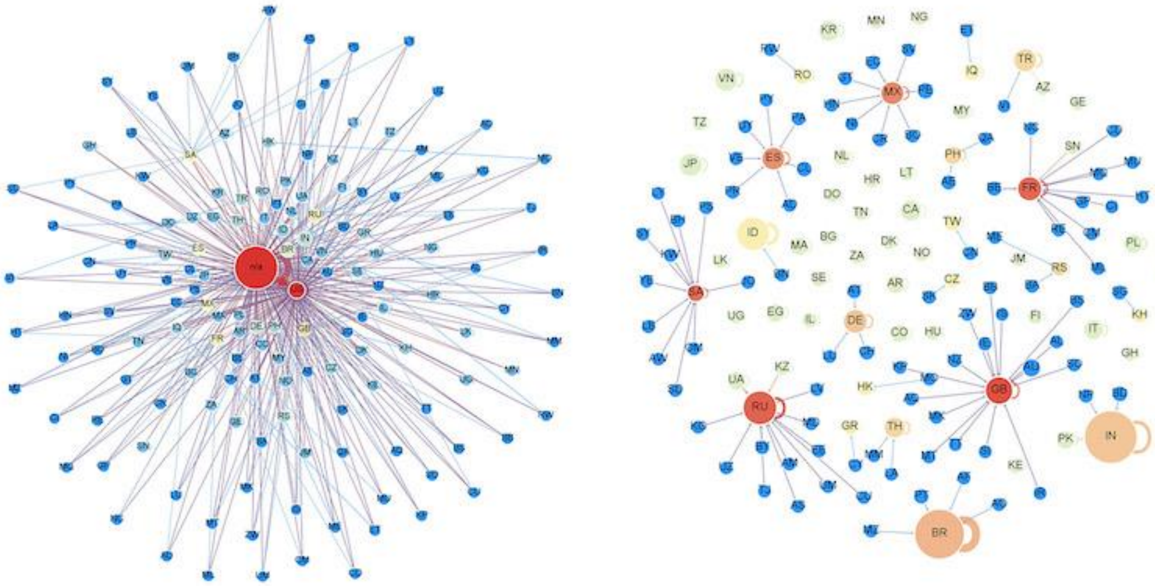

Figure 14: The 150 countries with at least 100 channels in the full channel network. Size indicates the number of channels in the locale and color the indegree.

In Figure 14 on the left, the ‘n/a’ group and the US tower over the rest, with an indegree of 150 and 145 respectively: every country connects to the most dominant group (53.6 percent of all channels) and almost everyone to the next in line (8.4 percent). The more interesting findings appear at a level below, after removing the two top nodes from the network (Figure 14 , right). We now see a relatively large number of countries that link only to themselves as well as the emergence of regional and language based hubs: Great Britain keeps part of its global appeal, while Russia, France (FR), Brazil, India, Germany, and Saudi Arabia (SA) take on the role of language hubs. In the case of Spanish, Spain and Mexico share that role. While the first can be interpreted as a remnant of the colonial relationship, the second aligns with the muscle of the Mexican audiovisual industry and its influence in all Latin America. The same can be said about Brazil in the Portuguese sphere.

In order to better understand linking behavior between countries in numerical ways, we calculate two homophily (*cf.* , Lazarsfeld and Merton, 1954) coefficients for each country, one for incoming and one for outgoing connections. In both cases, we divide the number of connections with the same locale by the number linking to others. A country with 10 incoming links to itself, and five incoming links from other countries thus has an incoming homophily of two. Incoming homophily expresses the interest *of* other countries, while outgoing homophily expresses the interest *in* other countries. We also calculate a metric for relative 'prominence' that is indegree divided by channel count — the higher the number of incoming links per channel, the higher the prominence.

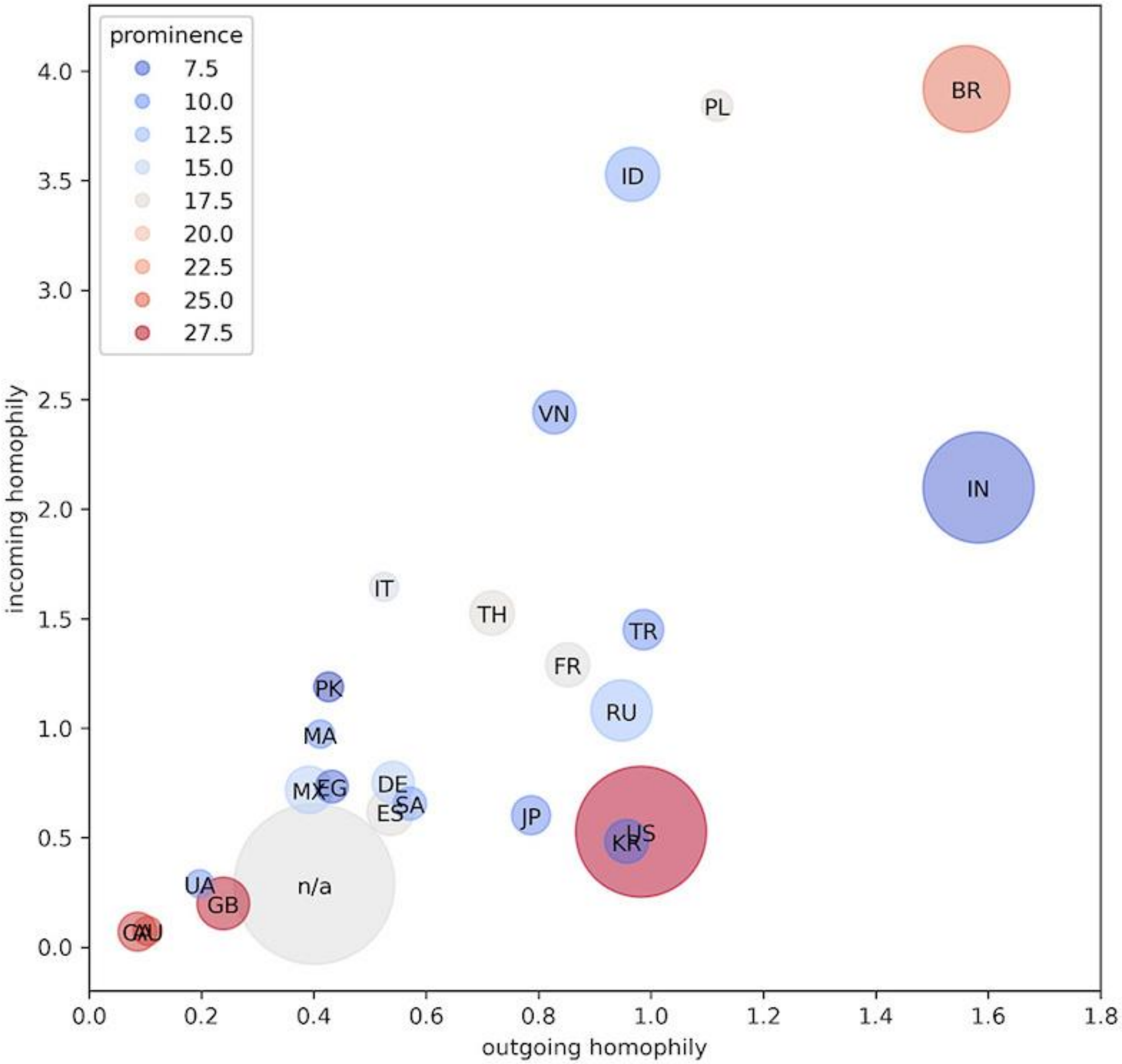

Figure 15: The top 25 countries per number of channels in the elite tier, showing outdegree homophily (x-axis), indegree homophily (y-axis), channel count (size), and prominence (color).

Figure 15 shows a high prominence for English speaking countries like the United States, Great Britain, Canada (CA), and Australia (AU) and low prominence of non-western countries like India, Japan, Vietnam (VN), Morocco (MA), Turkey (TR), or Egypt, which adds another piece of evidence to our previous statements about cultural dominance. We can also see that Brazil has noteworthy prominence in conjunction with the highest indegree homophily, portraying a large and active domestic user base not much interested in other countries' content. India also has a strongly connected local sphere but receives more attention globally, most probably related to Bollywood popularity in Arab countries and elsewhere. Poland, Indonesia, Vietnam, Thailand (TH), Turkey, France (FR), Russia, South Korea, and the United States stand as the strongest domestic markets, but also attract subscribers from other latitudes. Finally, we can see that other English-speaking countries like Canada, Australia, and the United Kingdom emerge as transnational spaces that exchange subscribers with other localities. Ukraine as well presents low outdegree and indegree homophily, which can be attributed to the fact that roughly a third of its citizens are native Russian speakers. These exploratory descriptions show the potential of YouTube data to investigate the complex patterns of media globalization within YouTube's platformed media system and we share the full set of country data in the Appendix .

4. Discussion and conclusion

In this project, we attempted to move beyond the limitations of existing approaches to paint an overall picture of what we termed YouTube's 'platformed media system', a large and complex amalgamation of channels that are hosted and tended within the company's interfaces and data centers. While our crawling strategy cannot claim to be free of shortcomings, our findings can be interpreted with some confidence, at least for the upper tiers of the YouTube hierarchy (channels with 1k–10k, 10k–100k, and 100k+ subscribers). The main roadblock to even more comprehensive sampling is the limited number of daily API units, but the highly connected character of YouTube channels indeed makes crawling a viable technique.

In the spirit of exploratory data analysis, we did not start from a strict research question, but proceeded mainly descriptively to provide a broad overview that can serve as a point of reference for researchers focusing on topic or user samples. Our main goal was to 'map' YouTube, not necessarily in visual terms, but with attention to various stratifying and structuring elements that, in part, are established by YouTube itself, for example when it comes to access to monetization through advertising or through the internal automated channel category system. What emerged is a complex picture of one of the leading social media platforms in mostly quantitative terms. We tackled the over 36M channels in our sample from three distinct but complementary angles, focusing on hierarchical stratification as well as segmentations by content category and country. Four main ‘narratives’ crystalized in the process.

First , in line with longstanding assessments of other networked platforms, we find that YouTube is dominated by a small number of elite channels. Far from the often-cited Pareto ratio of 80/20, the 153k elite channels above 100k subscribers, a mere 0.42 percent of our total sample, account for 69.2 percent of subscribers and 62.4 percent of views. And similar patterns continue *within* the elite tier, where we found that the 15,496 channels (0,04 percent) with more than 1M subscribers account for a little more than a third of all subscribers and views on the platform. While we saw that newer channels have still been able to penetrate the upper echelons, these may well be spin-offs of existing channels, professionally funded start-ups, or creators using their notoriety in other areas to gain visibility on YouTube. Our short analysis of subscriber and connectivity distributions was not able to single out growth mechanisms, but indicated that a number of different principles are at work here. This indeed calls for further research, but our findings suggest that YouTube is no longer a ‘a metaphor for the democratizing power of the Internet and information’ (Levine, 2010), but an increasingly saturated mass-media outlet where a small number of high-visibility channels preside over a large mass of channels struggling to join their ranks and achieve economic viability.

Second , these hierarchies are clearly connected to trends toward professionalization. The emergence of a ‘protoindustry of social media entertainment’ [22] has come full circle and YouTube’s hand in the process is evident. The company has invested in a support environment that expands as subscribers grow, stratifying between creator tiers in different ways. Algorithmic incentives, mediated through ‘algorithmic gossip’ (Bishop, 2019), have had a strong effect on creator tactics, manifesting in demanding publishing schedules, longer videos, channel networking, and attention to details such as metadata quality and completeness. Much remains unexplored here, but the rise of Patreon and other alternative revenue streams signals that creators are looking for means to reduce the grip the platform holds on their income. This is bound to have an effect on available contents. In line with Kumar’s critique that YouTube pushes toward mainstreaming and advertiser-friendly content, to the detriment of socially relevant but ‘inconvenient’ topics, we found that ‘softer’ topics like Lifestyle, Entertainment, and Music indeed make up the vast majority of contents on offer. Is the company ogling Netflix's business model and trying to develop its rooster of megastars into competitors to traditional film and television? While it would seem logical, from this perspective, that YouTube would try hard to separate itself from the dubious ideological content it has become associated with over recent years, we found that the classification of far-right channels into categories such as Entertainment and Lifestyle, which may enhance their findability, raises further questions. This precarious balancing act between different creator groups and audiences remains a defining characteristic of the platform (*cf.* , Gillespie, 2010).

Third , our analysis provides important insights into the structure and dynamics of globalization, as it plays out on YouTube. The US and, to a smaller extent, other English-language channels dominate both in terms of sheer numbers and in the topologies that emerge from channel connections. Countries with large populations, such as India, Brazil, or Indonesia, also have a strong presence on the platform, even if they struggle to attract global audiences. Below the striking presence of US channels, however, we were able to discover a second hierarchical layer that is mainly organized around language: Russia, France, Brazil, India, Germany, and Saudi Arabia take on the role of hubs within their respective linguistic and cultural spheres of influence. In the case of Spanish, Spain and Mexico share that role. The outcome is a complex picture of media globalization that combines elements of traditional US 'cultural imperialism', multipolarity, and stray phenomena such as South Korea's K-pop sensation. (*cf.* , Cunningham and Craig, 2016) Brazil's high level of segregation along almost all variables suggests that YouTube is sufficiently large to host substantial 'media systems within the media system', *i.e.* , large subsections that develop strong internal referentiality and coherence.

Fourth , our exploratory and inductive method raised many new questions for future research. With respect to the first and second narratives, we identified channel growth mechanisms as an important area for more detailed investigation, including algorithmic components, but also creator tactics such as metadata and keyword optimization. Our small experiment with Lewis' 'alternative influence network' highlighted the problematic character of YouTube's automated classification system, suggesting that alternative content classification methods could yield a better lens on what is on offer. Gathering more information on advertising matching and the high variability of revenue per view would allow us to connect topics more clearly with questions of media industrialization. This also connects to the third narrative, because per-video revenue also varies per country, making some audiences less attractive than others, which puts our homophily calculations in an entirely different light. More broadly, one could easily use our data to compare countries much more deeply, showing how content patterns and stratification varies between them. Our intensity and likeratio patterns gave a first glimpse of where this could go, but much remains to be done.

These and the many other questions that one could bring to a large-scale sample of platform data bring us back to the beginning of this paper. With platform data access becoming increasingly more difficult, who will have the ability to actually investigate

YouTube and its colleagues with a similar breadth of scope? How will independent research be able to shed light on the ‘big picture’ that frames the many smaller questions concerning specific users or issues? Despite its limitations, our approach shows a viable way forward and we should demand that it remains open for future analysis.

About the authors

Bernhard Rieder is Associate Professor of New Media and Digital Culture at the University of Amsterdam, the Netherlands.
E-mail: rieder [at] uva [dot] nl

Òscar Coromina is Adjunct Professor in Digital Media in the Faculty of Communication Sciences at Universitat Autònoma de Barcelona, Spain.
E-mail: oscar [dot] coromina [at] uab [dot] cat

Ariadna Matamoros-Fernández is a Lecturer in Digital Media in the School of Communication at Queensland University of Technology, Australia.
E-mail: ariadna [dot] matamorosfernandez [at] qut [dot] edu [dot] au

Acknowledgments

We want to thank Anna Jørgensen, Aikaterni Mniestri, and Janna Joceli Omena for their helpful comments.

Notes

[1] https://www.youtube.com/intl/en-US/about/press/ .

[2] Mohan and Punathambekar, 2019, p. 317.

[3] Kumar, 2019, p. 9.

[4] Social media analytics companies like SocialBlade have been collecting copious amounts of data from YouTube, but they focus on popularity rankings per country and the evolution of individual channels over time.

[5] Cunningham and Craig, 2016, p. 5,412.

[6] Tukey, 1962, p. 13f.

[7] Paolillo, *et al.* , 2019, p. 2,633.

[8] Non-profits like the Internet Archive and Common Crawl are exceptions, but these institutions rely on philanthropy.

[9] YouTube's 'creator benefit levels' overview: https://www.youtube.com/intl/en-US/creators/benefits/ .

[10] We initially started from a wider selection of channels, but since any crawl that does not get stuck in the beginning quickly opens into a giant connected component, seed selection becomes irrelevant in the absence of real random sampling possibilities. We settled on the educational channel Vsauce as a starting point, but since connectivity is so high, any crawl without depth limitation will essentially yield the same results.

[11] Overall, we mainly used three API endpoints: *channels* for channel data and connections, *playlistItems* for retrieving video listings, and *videos* for video metadata.

[12] YouTube calculates the quota cost of a particular query based on the information requested. For example, getting the main metadata for a video costs seven units. See https://developers.google.com/youtube/v3/determine_quota_cost .

[13] https://support.google.com/youtube/answer/9528076?hl=en .

[14] This principle, named after economist Vilfredo Pareto, has often been observed in real-world settings, *e.g.* , when 80 percent of income is generated by 20 percent of shoppers. In Internet research it is often used as a yardstick to compare inequalities in popularity, visibility, and so forth.

[15] If the observed distribution could be attributed to, for example, a preferential attachment mechanism (*cf.* , Barabási and Albert, 1999), we would be able to fit a power law. The fact that we cannot, means that the underlying mechanisms are less easily identifiable. For both variables, power laws can be fit to parts of the curve, but that is true for most empirical phenomena.

[16] https://support.google.com/youtube/answer/6175006?hl=en .

[17] https://support.google.com/youtube/answer/2991785?hl=en .

[18] To avoid calculation problems such as division by zero, we replaced all zeros in the view, like, and dislike count columns in our data with one.

[19] From YouTube's Creators Studio interface, https://studio.youtube.com .

[20] Kumar, 2019, p. 7.

[21] YouTube provides creators with the possibility to exclude specific locales or to limit a video to specific locales.

[22] Cunningham and Craig, 2016, p. 5,412.

References

Crystal Abidin, 2019. "Yes Homo: Gay Influencers, homonormativity, and queerbaiting on YouTube," *Continuum* , volume 33, number 5, pp. 614–629.
doi: https://doi.org/10.1080/10304312.2019.1644806 , accessed 15 July 2020.

Crystal Abidin, 2018. *Internet celebrity: Understanding fame online* . Bingley, U.K.: Emerald.

Swati Agarwal and Ashish Sureka, 2015. "Topic-specific YouTube crawling to detect online radicalization," In: Wanming Chu, Shinji Kikuchi, and Subhash Bhalla (editors). *Databases in networked information systems. Lecture Notes in Computer Science* , volume 8999. Cham, Switzerland: Springer International, pp. 133–151.
doi: https://doi.org/10.1007/978-3-319-16313-0_10 , accessed 15 July 2020.

Massimo Airoldi, Davide Beraldo, and Alessandro Gandini, 2016. "Follow the algorithm: An exploratory investigation of music on YouTube," Poetics, volume 57, pp. 1–13.
doi: https://doi.org/10.1016/j.poetic.2016.05.001 , accessed 15 July 2020.

Réka Albert, Hawoong Jeong, and Albert-László Barabási, 1999. "Diameter of the World-Wide Web," *Nature* , volume 401, number 6749 (9 September), pp. 130–131.
doi: https://doi.org/10.1038/43601 , accessed 15 July 2020.

Julia Alexander, 2018. "YouTube's top creators are burning out and breaking down en masse," *Polygon* (6 June), at https://www.polygon.com/2018/6/1/17413542/burnout-mental-health-awareness-youtube-elle-mills-el-rubius-bobby-burns-pewdiepie , accessed 8 May 2020.

Jeff Alstott, Ed Bullmore, and Dietmar Plenz, 2014. “powerlaw: A Python package for analysis of heavy-tailed distributions,” *PLoS ONE* , volume 9, number 1, p. e85777 (29 January).
doi: https://doi.org/10.1371/journal.pone.0085777 , accessed 15 July 2020.

Jane Arthurs, Sophia Drakopoulou, and Alessandro Gandini, 2018. “Researching YouTube,” *Convergence* , volume 24, number 1, pp. 3–15.
doi: https://doi.org/10.1177/1354856517737222 , accessed 15 July 2020.

Albert-László Barabási and Réka Albert, 1999. “Emergence of scaling in random networks,” *Science* , volume 286, number 5439 (15 October), pp. 509–512.
doi: https://doi.org/10.1126/science.286.5439.509 , accessed 15 July 2020.

Mathias Bärtl, 2018. “YouTube channels, uploads and views: A statistical analysis of the past 10 years,” *Convergence* , volume 24, number 1, pp. 16–32.
doi: https://doi.org/10.1177/1354856517736979 , accessed 15 July 2020.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy, 2009. “Gephi: An open source software for exploring and manipulating networks,” *Proceedings of the Third International AAAI Conference on Weblogs and Social Media* ,
at https://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154/1009 , accessed 15 July 2020.

Sophie Bishop, 2019. “Managing visibility on YouTube through algorithmic gossip,” *New Media & Society* , volume 21, numbers 11-12, pp. 2589–2606.
doi: https://doi.org/10.1177/1461444819854731 , accessed 15 July 2020.

Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti, 2012. “The untold story of the clones: Content-agnostic factors that impact YouTube video popularity,” *KDD ’12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , pp. 1,186–1,194.
doi: https://doi.org/10.1145/2339530.2339717 , accessed 15 July 2020.

Liliana Bounegru, Kari De Pryck, Tommaso Venturini, and Michele Mauri, 2020. “‘We only have 12 years’: YouTube and the IPCC report on global warming of 1.5°C,” *First Monday* , volume 25, number 2,
at https://firstmonday.org/article/view/10112/8332 , accessed 29 April 2020.
doi: https://doi.org/10.5210/fm.v25i2.10112 , accessed 15 July 2020.

James Bridle, 2017. “Something is wrong on the Internet,” *Medium* (6 November), at https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2 , accessed 29 April 2020.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener, 2000. “Graph structure in the Web,“ *Computer Networks* , volume 33, numbers 1–6, pp. 309–320. doi: https://doi.org/10.1016/S1389-1286(00)00083-9 , accessed 15 July 2020.

Axel Bruns, 2019. “After the ‘APIcalypse’: Social media platforms and their fight against critical scholarly research,” *Information, Communication & Society* , volume 22, number 11, pp. 1,544–1,566.
doi: https://doi.org/10.1080/1369118X.2019.1637447 , accessed 15 July 2020.

Taina Bucher, 2017. “The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms,” *Information, Communication & Society* , volume 20, number 1, pp. 30–44.
doi: https://doi.org/10.1080/1369118X.2016.1154086 , accessed 15 July 2020.

Jean Burgess and Joshua Green, 2018. *YouTube: Online video and participatory culture* . Second edition. Cambridge: Polity.

Jean Burgess and Joshua Green, 2009. *YouTube: Online video and participatory culture* . Cambridge: Polity.

Ronald S. Burt, 1992. *Structural holes: The social structure of competition* . Cambridge, Mass.: Harvard University Press.

Andrew Chadwick, 2013. *The hybrid media system: Politics and power* . Oxford: Oxford University Press.
doi: https://doi.org/10.1093/acprof:oso/9780199759477.001.0001 , accessed 15 July 2020.

Stuart Cunningham and David Craig, 2017. “Being ‘really real’ on YouTube: Authenticity, community and brand culture in social media entertainment,” *Media International Australia* , volume 164, number 1, pp. 71–81.
doi: https://doi.org/10.1177/1329878X17709098 , accessed 15 July 2020.

Stuart Cunningham and David Craig, 2016. “Online entertainment: A new wave of media globalization? — Introduction,” *International Journal of Communication* , volume 10, pp. 5,409–5,425, and at https://ijoc.org/index.php/ijoc/article/view/5725 , accessed 15 July 2020.

Matthias Funk, 2020. “How many YouTube channels are there?” *Tubics* (31 January), at https://www.tubics.com/blog/number-of-youtube-channels/ , accessed 29 April 2020.

Carolin Gerlitz and Bernhard Rieder, 2013. “Mining one percent of Twitter: Collections, baselines, sampling,” *M/C Journal* , volume 16, number 2, at http://www.journal.media-culture.org.au/index.php/mcjournal/article/view/620 , accessed 29 April 2020.

Tarleton Gillespie, 2010. “The politics of ‘platforms’,” *New Media & Society* , volume 12, number 3, pp. 347–364.
doi: https://doi.org/10.1177/1329878X17709098 , accessed 15 July 2020.

Barney G. Glaser and Anselm L. Strauss, 1967. *The discovery of grounded theory: Strategies for qualitative research* . Chicago: Aldine.

Matthew Hindman, 2009. *The myth of digital democracy* . Princeton, N.J.: Princeton University Press.

Mathieu Jacomy, Paul Girard, Benjamin Ooghe-Tabanou, and Tommaso Venturini, 2016. “Hyphe, a curation-oriented approach to Web crawling for the social sciences,&tdquo; *Tenth International AAAI Conference on Web and Social Media* , at https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13051 , accessed 15 July 2020.

Frank Kessler and Mirko T. Schäfer, 2009. “Navigating YouTube: Constituting a hybrid information management system,” In: Pelle Snickars and Patrick Vonderau (editors). *The YouTube reader* . Stockholm: National Library of Sweden, pp. 275–291.

Sangeet Kumar, 2019. “The algorithmic dance: YouTube’s Adpocalypse and the gatekeeping of cultural content on digital platforms,” *Internet Policy Review* , volume 8, number 2, at https://policyreview.info/articles/analysis/algorithmic-dance-youtubes-adpocalypse-and-gatekeeping-cultural-content-digital , accessed 8 May 2020.
doi: https://doi.org/10.14763/2019.2.1417 , accessed 15 July 2020.

Patricia G. Lange, 2019. *Thanks for watching: An anthropological study of video sharing on YouTube* . Louisville, Colo.: University Press of Colorado.

Patricia G. Lange, 2007. “Publicly private and privately public: Social networking on YouTube,” *Journal of Computer-Mediated Communication* , volume 13, number 1, pp. 361–380.
doi: https://doi.org/10.1111/j.1083-6101.2007.00400.x , accessed 15 July 2020.

Paul F. Lazarsfeld and Rober K. Merton, 1954. “Friendship as a social process: A substantive and methodological analysis,” In: Morroe Berger,Theodore Abel, and Charles H. Page (editors). *Freedom and control in modern society* . New York: Van Nostrand, pp. 18–66.

Zahavah Levine, 2010. “Broadcast yourself,” *Official YouTube Blog* (18 March), at https://youtube.googleblog.com/2010/03/broadcast-yourself.html , accessed 7 May 2020.

Becca Lewis, 2020. “All of YouTube, not just the algorithm, is a far-right propaganda machine,” *Medium* (8 January), at https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430 , accessed 7 May 2020.

Rebecca Lewis, 2018. “Alternative influence,” *Data & Society* (18 September), at https://datasociety.net/library/alternative-influence/ , accessed 29 April 2020.

Ramon Lobato and James Meese (editors), 2016. *Geoblocking and global video culture* . Amsterdam: Institute of Network Cultures, and at https://networkcultures.org/blog/publication/no-18-geoblocking-and-global-video-culture/ , accessed 28 April 2020.

Geert Lovink and Sabine Niederer (editors), 2008. *Video Vortex reader: Responses to YouTube* . Amsterdam: Institute of Network Cultures, and at https://networkcultures.org/videovortex/vv-reader/ , accessed 15 July 2020.

Sriram Mohan and Aswin Punathambekar, 2019. “Localizing YouTube: Language, cultural regions, and digital platforms,” *International Journal of Cultural Studies* , volume 22, numer 3, pp. 317–333.
doi: https://doi.org/10.1177/1367877918794681 , accessed 15 July 2020.

Fred Morstatter, Jürgen Pfeffer, and Huan Liu, 2014. “When is it biased? Assessing the representativeness of Twitter’s streaming API,” *WWW ’14 Companion: Proceedings of the 23rd International Conference on World Wide Web* , pp. 555–556.
doi: https://doi.org/10.1145/2567948.2576952 , accessed 15 July 2020.

John C. Paolillo, 2008. “Structure and network in the YouTube core,” *HICSS ’08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences* .
doi: https://doi.org/10.1109/HICSS.2008.415 , accessed 15 July 2020.

John Paolillo, Sharad Ghule, and Brian Harper, 2019. “A network view of social media platform history: Social structure, dynamics and content on YouTube” (8 January), at http://hdl.handle.net/10125/59701 , accessed 29 April 2020.

Tim Peterson, 2018. “Creators are making longer videos to cater to the YouTube algorithm,” *Digiday* (3 July), at https://digiday.com/future-of-tv/creators-making-longer-videos-cater-youtube-algorithm/ , accessed 8 May 2020.

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virglio A.F. Almeida, and Wagner Meira, 2020. “Auditing radicalization pathways on YouTube,” *FAT* ’20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* , pp. 131–141.
doi: https://doi.org/10.1145/3351095.3372879 , accessed 15 July 2020.

Bernhard Rieder, 2015. “Introducing the YouTube Data Tools,” *Politics of Systems* (4 May), at http://thepoliticsofsystems.net/2015/05/exploring-youtube/ , accessed 29 April 2020.

Bernhard Rieder, Ariadna Matamoros-Fernández, and Òscar Coromina, 2018. “From ranking algorithms to ‘ranking cultures’: Investigating the modulation of visibility in YouTube search results,” *Convergence* , volume 24, number 1, pp. 50–68.
doi: https://doi.org/10.1177/1354856517736982 , accessed 15 July 2020.

Richard Rogers, 2002. “The Issue Crawler: Makings of live social science on the Web,” *EASST Review* , volume 21, numbers 3–4, pp. 8–11, at http://easst.net/wp-content/uploads/2014/11/review_2002_12.pdf , accessed 29 April 2020.

Sergio Sayago, Paula Forbes, and Josep Blat, 2012. “Older people’s social sharing practices in YouTube through an ethnographical lens,” *BCS-HCI ’12: Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers* , pp. 185–194.

Pelle Snickars and Patrick Vonderau (editors), 2009. *The YouTube reader* . Stockholm: National Library of Sweden, and at http://www.kb.se/dokument/aktuellt/audiovisuellt/youtubereader/youtube_reader_052009_endversion.pdf , accessed 15 July 2020.

Nick Statt, 2020. “YouTube is a $15 billion-a-year business, Google reveals for the first time,” *The Verge* (3 February), at https://www.theverge.com/2020/2/3/21121207/youtube-google-alphabet-earnings-revenue-first-time-reveal-q4-2019 , accessed 29 April 2020.

Zeynep Tufekci, 2018. “YouTube, the great radicalizer,” >*New York Times* (10 March), at https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html , accessed 29 April 2020.

John W. Tukey, 1977. *Exploratory data analysis* . Reading, Mass.: Addison-Wesley.

John W. Tukey, 1962. “The future of data analysis,” *Annals of Mathematical Statistics* , volume 33, number 1, pp. 1–67.
doi: https://doi.org/10.1214/aoms/1177704711 , accessed 15 July 2020.

Peter van Aelst, Jesper Strömbäck, Toril Aalberg, Frank Esser, Claes de Vreese, Jörg Matthes, David Hopmann, Susana Salgado, Nicolas Hub&eracute;, Agnieszka Stpińska, Stylianos Papathanassopoulos, Rosa Berganza, Guido Legnante, Carsten Reinemann, Tamir Sheafer, and James Stanyer, 2017. “Political communication in a high-choice media environment: A challenge for democracy?” *Annals of the International Communication Association* , volume 41, number 1, pp. 3–27.
doi: https://doi.org/10.1080/23808985.2017.1288551 , accessed 15 July 2020.

Duncan J. Watts, 2004. “The ‘new’ science of networks,” *Annual Review of Sociology* , volume 30, pp. 243–270.
doi: https://doi.org/10.1146/annurev.soc.30.020404.104342 , accessed 15 July 2020.

Appendix

Shared data available here:
https://bit.ly/yt-crawl-2019 .
For other data sharing requests please contact rieder [at] uva [dot] nl.

Editorial history

Received 16 May 2020; accepted 4 June 2020.

Mapping YouTube: A quantitative exploration of a platformed media system
by Bernhard Rieder, Òscar Coromina, and Ariadna Matamoros-Fernández.
First Monday , Volume 25, Number 8 - 3 August 2020
https://firstmonday.org/ojs/index.php/fm/article/download/10667/9575
doi: http://dx.doi.org/10.5210/fm.v25i8.10667