Content Creation on YouTube: A Longitudinal Study

Abstract

AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions. AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions. AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions. AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions.

Introduction

YouTube, the world's second biggest online social network, plays a sizeable role in our society. In a 2018 Pew Research study (Pew Research 2018), 54% of adult U.S. users said the platform was somewhat or very important for helping them understand things that are happening in the world, 52% claimed it was somewhat or very important to decide whether to buy a particular product or not, and 34% report that their children regularly watch videos on the platform.

Previous research on the YouTube platform give insight into geographical patterns of video virality (Brodersen, Scellato, and Wattenhofer 2012), patterns related to video popularity and traffic (Gill et al. 2007; Cha et al. 2007), as well as troublesome phenomena happening on the platform (Horta Ribeiro et al. 2019; Papadamou et al. 2019). Yet, little is know about *what content* has prospered in the platform through the years, as well as how *the process of creating content* for YouTube has evolved.

The proccess of content creation are particularly interesting on YouTube as: *(i)* the platform allows content creators to directly monetize their content through a comprehensive partner program where ad revenue is shared (Google 2020); *(ii)* during much of the time users spend on YouTube, they watch videos algorithmically recommended by the platform (Solsman 2018), often for content creators that they have not explicitly subscribed.

Since YouTube plays an ever-large role in entertaining and informing society, we argue that studying the evolution of content and content creation in the platform through the years is key to understand our current information landscape and to provide insights on how to improve it.

Present Work. In this paper, we conduct a large scale longitudinal study examining the metadata of 69 million videos of English speaking YouTubers. More specifically, we ask two research questions:

RQ1 What are the kinds of content and the formats that prospered on YouTube through the years?

RQ2 How have content creation practices changed through years on YouTube?

To answer these questions, we explore several different aspects of YouTube videos and channels, such as: the distribution of views, videos and channel activity through different categories; how the duration of videos changed across the years; the regularity and the dates when videos are posted. Our results indicate three main trends:

1. There was a *topical shift* in the platform, which contains increasingly more news-related content, less music, and (in recent years) less gaming content.
2. There is an increasing professionalization of successful YouTubers, which whose content creation strategy seems to differs from the average content creator.
3. We find that the YouTube's monetization criteria significantly affects the production of content. For instance, video duration

Besides these findings, we also: *(i)* develop a sampling algorithm to obtain a more representative sample a biased sample of channels and their overall ranking, and, *(ii)* make available a YouTube dataset of unprecedented size, with more than 69 million videos from 89 thousand channels across 11 years.

Related Work

Methods

Dataset Collection

We illustrate our data collection methodology in Fig.**??**, and explain each of the steps below:

1. **Channel Pool:** We begin by gathering a pool of $164,649$ channels by crawling all english channels with more than 10k subscribers from `channelcrawler.com`. Language classification is made by the website using an automatic classifier. The website exists atleast since mid 2013[1], and uses a snowball sampling approach to collect YouTube channels. Their repository of channels is particularly helpful, as scrapping data from YouTube was easier in previous times.
2. **Video Metadata**
3. **Rankings**
4. **Filtering Language**

Obtaining a (More) Representative Sample

The rankings from Social Blade show us that our data is not entirely representative, for example, we have around 8% of Comedy channels with between 10^4 to 10^5 subscribers, but around 19% of those with between 10^4 to 10^5 subscribers. Yet, it also allows us to partially correct these biases.

Let $R = \{1 \ldots, k\}$ be an ordered set of all ranked channels. We know that our sample correspond to a subset O, which we call *observed channels*. The complement of this subset are the *unobserved channels* U. In this scenario, we want to compensate for our biases by weighting each channel inversely to its sampling probability.

While the sampling probability over R is not homogeneous globally, we hypothesize that it is likely to be homogeneous *locally*. We confirm our intuition by sliding windows of varied sizes through the ranks and performing the Mann-Whitney rank test comparing the ranks of observed and unobserved channels. We find that smaller window sizes yield higher p-values. For example, for technology channels, when sliding a window of size 5000, we have that the p-values are always smaller than 0.1, where for a window of size 500, they are consistently around 0.5. Recall that the null hypothesis of the test is that, if we draw values X and Y from the two populations at random (here, the rankings of observed and unobserved channels), that $P(X > Y) \neq P(Y > X)$. This suggests that locally, our sampling bias is much smaller.

In that context, assuming that, locally, the sampling probability is homogeneous, we do as follows:

Step 1 Estimate the local sampling probability with a moving average. This moving average calculates the chance of a channel being observed at rank i as the average of channels observed between ranks $[i - k, i + k]$, where k is the width of the moving average.

Step 2 Normalize the estimated probability of being sampled for each observed sample, and assign weights inversely proportional to the normalized value.

[1] `https://web.archive.org/web/*/https://www.channelcrawler.com/`

Collapsing Categories

Topic Model for Keywords

Categories are the only categorical variable w.r.t. the content published on YouTube.

High Performing Videos and Channels

An important dychotomy to study on YouTube is performance

.

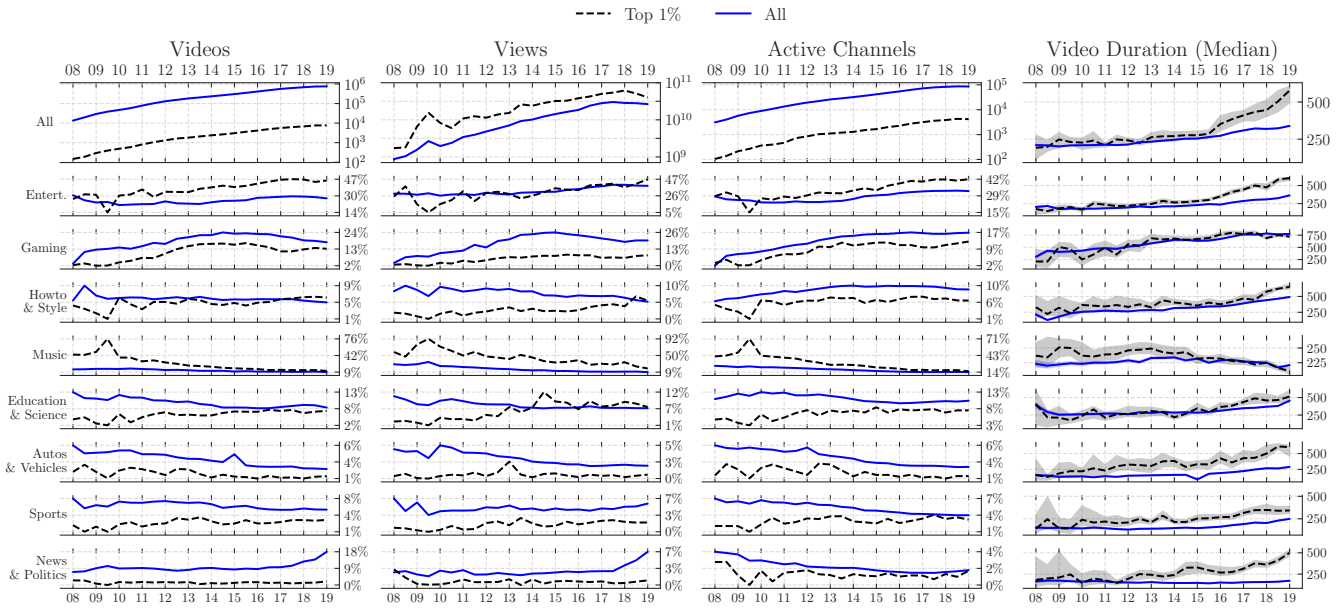

Figure 1: Overview of the number of videos (first column), views (second column), active channels (third column) and the median video duration (fourth column) throughout the last 11 years. In the first row we show the total number of videos/views/active channels, and the median video duration in seconds for all channels in blue. In black, we portray the proportion of views contributed by the top 1% videos with most views, the proportion of active channels responsible for the top 1% of videos, and the median duration for these most popular videos. Below, we show the composition of videos/views/active channels in terms of categories for the top 1% of videos and for the bottom 99%. We also stratify the median duration per category for these two groups.

General Characterization

We begin by providing an overview of the video production during the last 11 years according to the collected dataset, as portrayed in Fig. 1.

Videos. In the first column we examine the amount of videos published on YouTube from the beginning of 2008 to mid 2019. In the first row, we portray the number of videos uploaded each semester (in blue). Below, we show the composition of the videos uploaded each semester in terms of their categories for the bottom 99% of videos (in the second row), and for the top 1% of videos (in the last row).

We find interesting trends regarding the composition of the categories of the videos published. The "Other" category, which agglutinates "Sports", "Education", "Science & Technology", etc, decreases from roughly 40% to roughly 24% in the bottom 99% of videos through the 11 years of data, becoming less popular than the "Entertainment" category. Also, we can observe the explosion of "Gaming" content, which goes from around 4% of the bottom 99% of videos in 2008 to 24% in mid 2014, and then decreases again. Another noteworthy trend is the rise in videos related to "News & Politics" from 2015 to 2019, where the percentage of videos roughly doubles (going roughly from 8% to 16%).

The top 1% of videos paints a different picture. First, there is a peak of "Music" videos in the top 1% in mid 2009, and a decrease ever since. The peak coincides with the creation of *Vevo*, a joint venture among record companies which distributed music by famous artists on YouTube **??**. Second, the dominance of "Entertainment" videos is more pronounced in the top 1%. Lastly, "News & Politics" is under-represented in the top 1%. In 2019, for instance, while in the bottom 99% the categeory represented more than 16% of videos, here it represents less than 2%.

Views. In the second column we show the number of views collected at crawl time. Importantly, videos have different behavior when it comes to views, *e.g.*, music videos may accumulate views indefinitely, while news often grab little attention after a few days or weeks. In the first row we show the total amount of views per semester (in blue), and also the percentage of the total number of views that correspond to the top 1% of videos (in black). Below, we show the percentage of views that belong to videos from each of the categories for the bottom 99% and top 1% of videos.

The peak in total number of views in late 2018 (in the first row) may be explained by the cumulative nature of views: videos that are older had more time to gather them. Moreover, we also see the peak of musical videos in mid 2009 due to *Vevo*, which here corresponds to more than 80% of views from the top 1% of videos. Lastly, the percentage of videos that correspond to the top 1% of views decreases, from roughly 70% in 2008 to around 60% in 2019. This suggests that YouTube views are becoming less concentrated, although it could be that some videos distort this statistic as they accumulate views through the years.

Comparing the percentage of videos in each category and their share of total views, we can deduce which categories output less, but more popular videos. For example, the percentage of "Music" videos in both scenarios is much lower than the percentage of views associated with these music videos. For "Entertainment" videos, on the other hand, we

have that the percentage of views associated with the bottom 99% of entertainment videos is proportionally higher than the proportion of videos belonging to this category. So while around 30% of videos produced in 2018 were Entertainment videos, they corresponded to roughly 40% of the views.

Active Channels In the third column we display statistics related number of active channels per month, that is channels channels that published atleast one video in the 6 months period being analyzed. In the first row, we portray the total number of such channels (in blue). Below, we show the composition of the videos uploaded each semester in terms of their categories for the bottom 99% of videos (in the second row), and for the top 1% of videos (in the last row).

Duration In the fourth and last column we show how the median video duration evolved during the last 11 years. In the first row we portray the median duration for all videos (in blue), and for the top 1% of videos (in dashed black). Below, we stratify and show the median duration per category for the bottom 99% of videos (in the second row) and for the top 1% of videos (in the last row).

We find that the median duration of the top 1% and for all videos increased throughout the years. For all videos, it reached around 950 seconds in 2019 while for the top 1% it was roughly 400 seconds in 2008. For the top 1%, it went from around 250 seconds in 2008 to roughly 500 in 2019.

Analyzing the duration of different categories for the top 1% of videos and the bottom 99%, we can see that different categories have very different median video lengths, for example, videos in the "Gaming" and the "Howto & Style" category tend to be much longer. Moreover, we see that in some categories there are the duration profiles between the median bottom 99% video and median top 1% are different. For example, the median duration of top 1% "News & Politics" is much higher in recent years than their bottom 99% counterpart ($\sim$ 500s vs. $\sim$ 180 median duration in 2019). The same is true for "Entertainment" videos ($\sim$ 520 vs. $\sim$ 360 median duration in 2019).

Video Duration

We explore further how the video duration on YouTube evolved over the last 11 years.

.

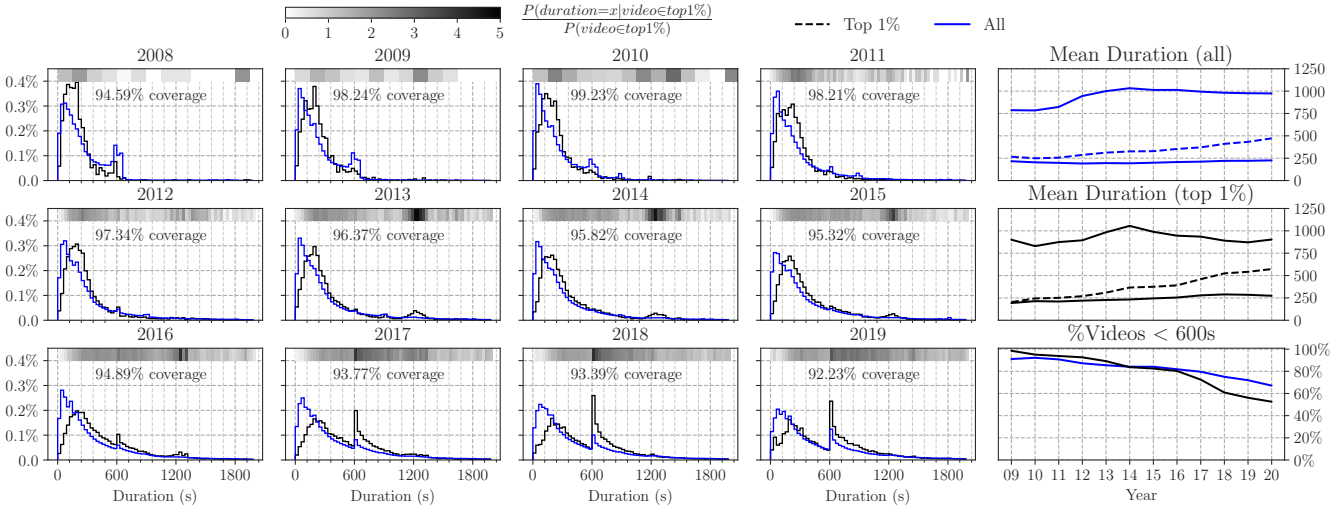

Figure 2: this is a caption.

.

Activity and Regularity

.

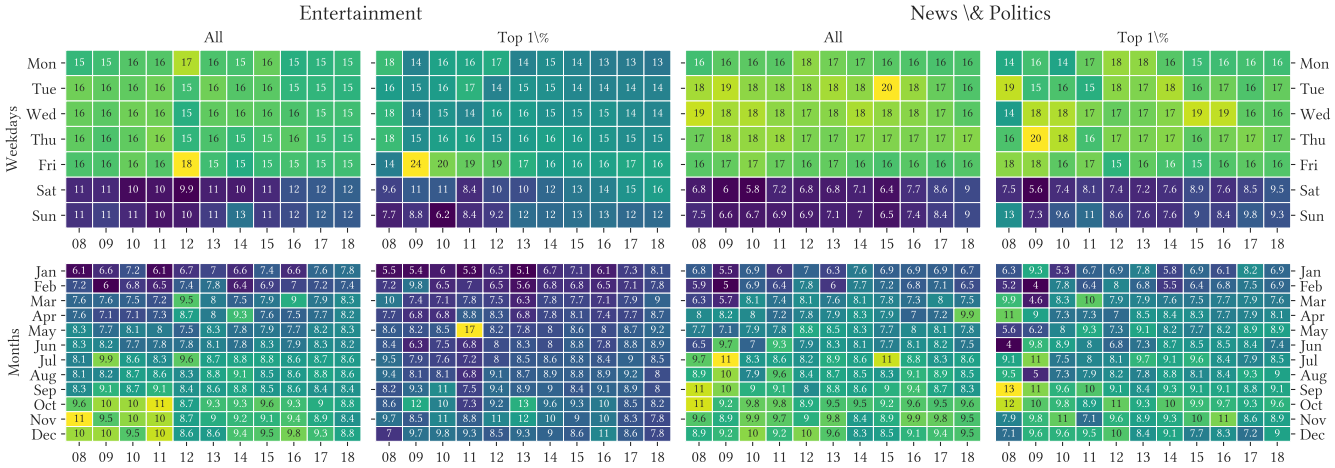

Figure 3: this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption.

.

.

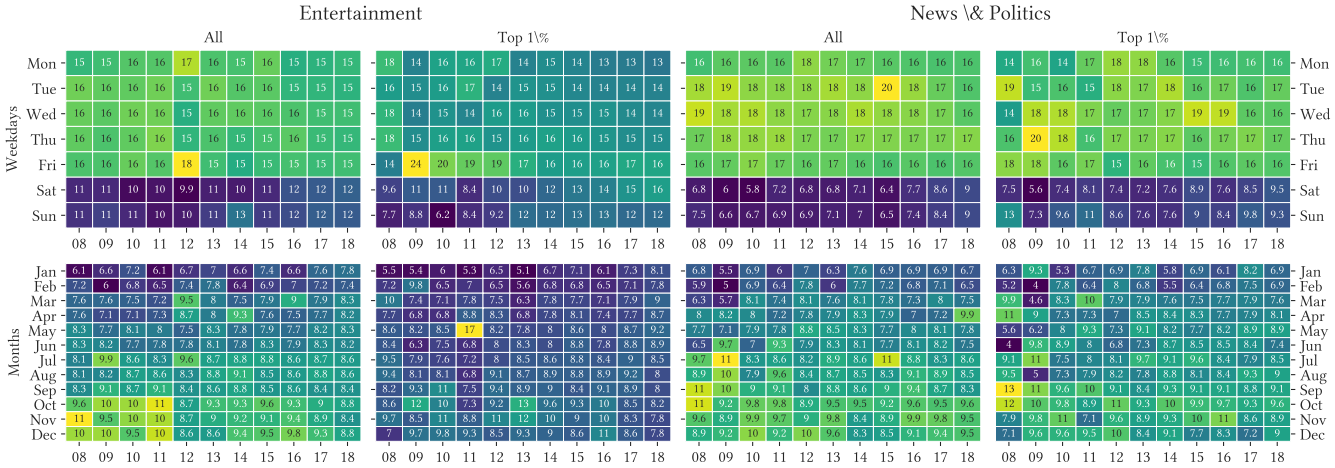

Figure 4: this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption. this is a caption.

What's on YouTube?

.

.

Discussion and Conclusion

- Exciting perspective: obtain representative samples in other websites using the same methodology.
- Begets further inspection: impact of monetization strategies in content we consume.
- Another interesting line: how YouTubers compete for user's attention.
- Last but not least, cool dataset to many other applications.

.

References

Brodersen, A.; Scellato, S.; and Wattenhofer, M. 2012. YouTube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 241–250. Lyon, France: Association for Computing Machinery.

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, 1–14. San Diego, California, USA: Association for Computing Machinery.

Gill, P.; Arlitt, M.; Li, Z.; and Mahanti, A. 2007. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, 15–28. San Diego, California, USA: Association for Computing Machinery.

Google. 2020. YouTube Partner Program. https://web.archive.org/web/20200316084418/https://support.google.com/youtube/answer/72851?hl=en.

Horta Ribeiro, M.; Ottoni, R.; West, R.; Almeida, V. A. F.; and Meira, W. 2019. Auditing Radicalization Pathways on YouTube. In *FAT* 2020*.

Papadamou, K.; Papasavva, A.; Zannettou, S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; and Sirivianos, M. 2019. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *arXiv:1901.07046 [cs]*. arXiv: 1901.07046.

Pew Research. 2018. Many Turn to YouTube for Children's Content, News, How-To Lessons.

Solsman, J. 2018. CES 2018: YouTube's AI recommendations drive 70 percent of viewing. https://web.archive.org/web/20191107090451/https://www.cnet.com/news/youtube-ces-2018-neal-mohan/.

Stratified by Country							
Kind	Bracket	#Channels	% Channels	Kind	Bracket	#Channels	% Channels
All	10^4 to 10^5	91278	0.27	United States	10^4 to 10^5	33879	21.51
	10^5 to 10^6	34673	4.07		10^5 to 10^6	14233	12.27
	10^6 to 10^7	5636	4.17		10^6 to 10^7	2626	9.44
	10^7 to 10^{10}	268	45.97		10^7 to 10^{10}	150	61.22
United Kingdom	10^4 to 10^5	6822	6.16	Australia	10^4 to 10^5	1895	25.01
	10^5 to 10^6	2516	21.53		10^5 to 10^6	699	2.14
	10^6 to 10^7	411	66.94		10^6 to 10^7	105	61.05
	10^7 to 10^{10}	26	78.79		10^7 to 10^{10}	3	100.00
Canada	10^4 to 10^5	3398	25.45	New Zealand	10^4 to 10^5	325	29.65
	10^5 to 10^6	1262	29.51		10^5 to 10^6	90	51.43
	10^6 to 10^7	180	61.02		10^6 to 10^7	8	100.00
	10^7 to 10^{10}	11	100.00		–	–	–
Stratified by Category							
Kind	Bracket	#Channels	% Channels	Kind	Bracket	#Channels	% Channels
film	10^4 to 10^5	4923	10.30	entertain.	10^4 to 10^5	13866	9.40
	10^5 to 10^6	1724	21.07		10^5 to 10^6	6696	5.89
	10^6 to 10^7	306	32.08		10^6 to 10^7	1505	29.43
	10^7 to 10^{10}	8	72.73		10^7 to 10^{10}	79	43.17
music	10^4 to 10^5	17410	10.28	comedy	10^4 to 10^5	1931	8.13
	10^5 to 10^6	6061	5.76		10^5 to 10^6	1187	18.84
	10^6 to 10^7	1064	7.11		10^6 to 10^7	279	32.44
	10^7 to 10^{10}	96	59.63		10^7 to 10^{10}	17	65.38
games	10^4 to 10^5	13385	14.78	tech	10^4 to 10^5	3182	13.55
	10^5 to 10^6	4902	6.02		10^5 to 10^6	1202	27.41
	10^6 to 10^7	669	31.25		10^6 to 10^7	144	43.90
	10^7 to 10^{10}	23	58.97		10^7 to 10^{10}	3	75.00
sports	10^4 to 10^5	3414	13.60	education	10^4 to 10^5	4945	6.78
	10^5 to 10^6	1143	26.48		10^5 to 10^6	2142	18.54
	10^6 to 10^7	155	42.01		10^6 to 10^7	464	24.42
	10^7 to 10^{10}	3	100.00		10^7 to 10^{10}	7	87.50
people	10^4 to 10^5	13093	5.44	howto	10^4 to 10^5	6875	13.35
	10^5 to 10^6	3849	8.57		10^5 to 10^6	3337	7.44
	10^6 to 10^7	464	24.42		10^6 to 10^7	465	37.11
	10^7 to 10^{10}	15	55.56		10^7 to 10^{10}	13	56.52
nonprofit	10^4 to 10^5	804	11.59	autos	10^4 to 10^5	2649	15.88
	10^5 to 10^6	175	22.41		10^5 to 10^6	745	28.50
	10^6 to 10^7	3	37.50		10^6 to 10^7	44	35.20
animals	10^4 to 10^5	848	11.44	travel	10^4 to 10^5	1571	0.42
	10^5 to 10^6	332	24.65		10^5 to 10^6	367	26.31
	10^6 to 10^7	41	46.59		10^6 to 10^7	24	32.43
news	10^4 to 10^5	1455	7.28	shows	10^4 to 10^5	19	35.85
	10^5 to 10^6	722	13.16		10^5 to 10^6	1	100.00
	10^6 to 10^7	147	39.20		–	–	–
	10^7 to 10^{10}	1	100.00		–	–	–

Table 1: Assessment of the representativeness of the channels collected according to their social blade ranking.