Broadcast Yourself: Understanding YouTube Uploaders

Yuan Ding†, Yuan Du†, Yingkai Hu†, Zhengye Liu‡, Luqin Wang†,
Keith W. Ross†, Anindya Ghose⋆

†Computer Science and Engineering, Polytechnic Institute of NYU
‡AT&T Labs, San Ramon
⋆Stern School of Business, New York University

{dingyuan1987, duyuanvvv, yingkaihu, zhengyeliu, lukelwang}@gmail.com
ross@poly.edu, aghose@stern.nyu.edu

ABSTRACT

YouTube uploaders are the central agents in the YouTube phenomenon. We conduct extensive measurement and analysis of YouTube uploaders. We estimate YouTube scale and examine the uploading behavior of YouTube users. We demonstrate the positive reinforcement between on-line social behavior and uploading behavior. Furthermore, we examine whether YouTube users are truly broadcasting themselves, via characterizing and classifying videos as either user generated or user copied.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: General—*measurement techniques*; C.2.3 [**Computer-Communication Networks**]: Network Operations—*network monitoring, public networks*

General Terms

Measurement, Performance

Keywords

YouTube, system scale, social network, content classification

1. INTRODUCTION

YouTube is a major Internet phenomenon. Driving YouTube's success are its *uploaders* – the millions of users who upload content to the YouTube site. YouTube uses the slogan "Broadcast Yourself" to highlight its unique feature of allowing ordinary Internet users to freely distribute videos. By encouraging users to upload content and broadcast themselves, YouTube has transformed Internet users from video consumers to video producers.

Due to its remarkable success, there have been several recent studies about YouTube [13, 4, 6, 10, 12, 1, 2, 3, 7, 19, 17, 14, 16, 18], providing important insights into YouTube videos, viewers, social behavior, video traffic, and recommendation system. However, the YouTube uploaders – who are at the core of the YouTube phenomenon – have yet to be studied carefully. In this paper, we provide a large-scale, comprehensive measurement study of YouTube uploaders. Such a study is important, as a big-picture understanding of YouTube must take into account its uploaders. Our contributions are as follows:

- **Analysis of YouTube uploaders:** All the videos in YouTube are uploaded by users. The most fundamental question is *who are the uploaders and what are their characteristics*. To this end, we carefully design crawling, sampling and analysis methodologies. Estimating the scale of YouTube, we find that there are approximately 47.3 million uploaders, having uploaded 448 million of videos (with an aggregated length of 2,649 years), and having attracted more than 1.5 trillion of views. To our knowledge, this represents the only modern (and systematic) estimate of YouTube scale currently available. We find that the top 20% of the uploaders (in terms of number of uploaded videos) contribute 72.5% of videos, while the top 20% of most popular uploaders attract approximately 97% of views. We find that uploaders belonging to the social network are more active than those not belonging, and upload on average more than twice as many videos. We study the uploading behavior of users with different gender, age, and geographic locations.
- **UGC vs. UCC:** YouTube is designed for sharing user generated content. However, many videos in YouTube are simply copied from other places – such as from movies, TV shows, and other professional video websites – rather than being originally generated by ordinary Internet users. We classify videos as UGC and UCC and investigate the properties of UGC and UCC videos from three different groups of representative uploaders. We make the critical observation that most uploaders consistently upload either UGC or UCC videos.

2. YOUTUBE UPLOADERS

In the jargon of YouTube, each subscribed user has its own *channel*. Throughout this paper we use the term *user* and *channel* interchangeably. Each channel has a channel profile that includes the channel name and (optionally) personal information about the user, such as the user's name, gender, age, interests and homepage (outside of YouTube). A user is said to be an *uploader* if it has uploaded at least one video. Each channel also includes links to all of the user's uploaded videos, links to users favorite videos, links to the users friends, and links to its subscribers and subscriptions.

IMC'11, November 2–4, 2011, Berlin, Germany.

2.1 Methodologies and Datasets

In this paper we use the related video graph to crawl the uploaders. Specifically, for a given uploader, we first obtain the list of uploaded videos. For each of its uploaded videos, we crawl the related videos and determine all the corresponding uploaders, which we refer to as the *related uploaders*. We then repeat the process. Unlike most other social network measurement studies, where typically a small part of the social graph is crawled, and the results are highly dependent on the sampling methodologies, we attempt to crawl the entire related video graph to avoid any assumptions underlying the sampling methodologies that may not be held for the related video graph. After crawling the entire related video graph, we then randomly select a subset of uploaders and conduct deeper investigation.

Using a depth-first-search (DFS) of the related uploader graph, we crawled approximately 44.8 million unique uploader IDs over a 90-day period (from 9/22/2010 to 12/22/2010). In order to crawl such a large number IDs in a 90-day period, we did not download the channel page for each ID; instead we simply collected the ID for each uploader. In the last weeks, our crawler, based on related videos, continued to see a large number of uploaders every hour; but, the number of new IDs per hour dropped off dramatically. To verify that the 44.8 million uploader IDs nearly fully cover the YouTube uploaders, we conducted another independent crawl via the YouTube friend social network. The related video network and the social network are two separate networks. The social network is created by the YouTube users, and the related video network is created by YouTube. Using DFS of the friend social network, we crawled 100,000 YouTube social uploaders. We found that 94,737 of the social uploaders (94.7%) were found in our set of 44.8 million uploaders. Given that crawling rate dramatically dropped in the last weeks of our crawl, and that the resulting set contained 94.7% of the uploaders from an independent crawl, we believe that the crawl is representative and nearly complete over all YouTube uploaders. From the 44.8 million IDs, we randomly selected 100,000 uploaders and then fully crawled all their channel information. We call this dataset `UploadSam`.

2.2 Scale of YouTube

In this subsection, by measuring and analyzing the YouTube uploaders, we estimate the scale of YouTube, including the total number of uploaders, the total number of videos, the aggregate length of YouTube videos, the total number of views. YouTube currently does not make this information public. To our knowledge, there has not been any systematic attempt to estimate the scale of YouTube in the past several years.

First, we estimate the total number of uploaders in YouTube. Since we claim that most uploaders in YouTube have been crawled, the number of uploaders can be simply estimated as 44.8/0.947=47.3 million. We then define a *scale factor* for 473 by comparing the total number of uploaders and the number of uploaders in `UploadSam`. We now estimate the total number of videos and aggregate length in YouTube. We find that there are 947,960 videos uploaded by the uploaders in `UploadSam` and the aggregate length of these videos is 5.6 years. Using the same scaling factor of 473, we therefore estimate that, up to December 2010, the total number of videos published to be 448 million and the aggregate video length to be 2,649 years.

We can also estimate the total number of views and the total viewing time in YouTube. The total number of views of the videos in `UploadSam` is approximately 3.1 billion. Because we also know the length of individual videos and their view times for the videos in `UploadSam`, we can calculate the total viewing time on these videos, which is 20,891 years. Once again using the scale factor of 473, we can estimate the total number of views to be 1.5 trillion and total viewing time to be 9.9 million years. Table 1 summarizes the scale of YouTube.

Table 1: Estimated YouTube Scale

Number of uploaders	47.3 million
Number of videos	448 million
Aggregate video length	2,649 years
Number of views	1.5 trillion
Aggregate view time	9.9 million years

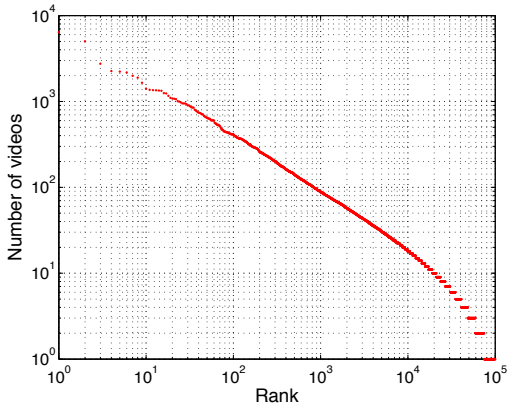


Figure 1: YouTube uploaders rank ordered by number of uploaded videos.

2.3 Uploading Behavior

Having investigated the scale of YouTube, we now investigate the behavior and characteristics of YouTube uploaders. To our knowledge, this is the first study of YouTube uploaders. We first investigate the distribution of the number of videos uploaded by the uploaders. As shown in Figure 1, the number of uploaded videos clearly follows a Zipf distribution. The maximum, mean, and median of uploaded videos are reported in Table 2. *Among these uploaders, the most active* 20% *of the uploaders contribute* 72.5% *of the videos, which largely follows the famous 80-20 rule [11].*

We further investigate the uploading pattern of YouTube uploaders. We observe that many uploaders upload very infrequently, but when they upload, they often upload several videos at the same time. An uploader who uploads multiple videos may only upload once during its life time. We investigate the "active time" of uploaders in different time scales. For example, if the time scale is one day, and if an uploader uploads at least one video, we consider this uploader active in that day, no matter how many videos it uploads. Figure 4 plots the PDF of active times of uploaders, with time scales of one day, one calendar month, and one calendar year. We eliminate the uploaders that newly joined and only consider the 83,769 uploaders that joined YouTube before 2010. We observe that 32.5%, 40.0%, and 56.1% of uploaders have been active for only one day, one month, and one year, respectively. Thus, less than half of the uploaders have been active for a period extending over one year.

We now examine whether a YouTube uploader tends to upload videos into a small number of categories. To this end, for each

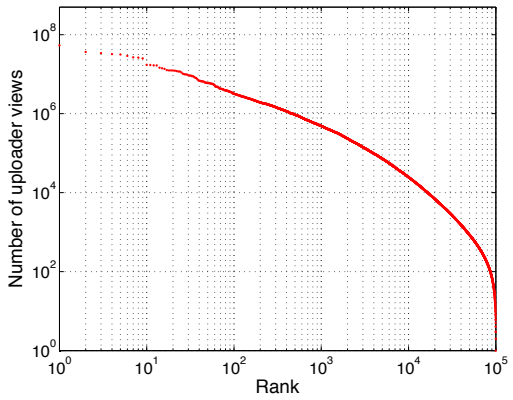


Figure 2: YouTube uploaders rank ordered by popularity.

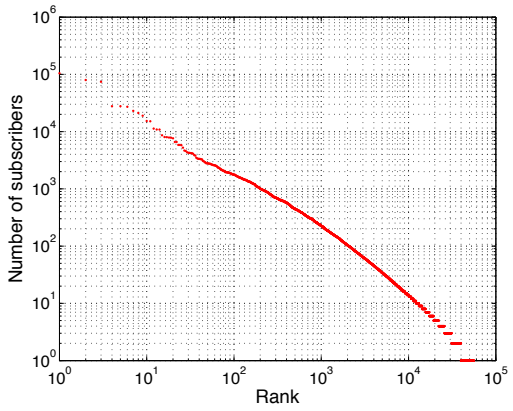


Figure 3: Distribution of number of subscribers.

uploader, we rank the categories based on the number of videos uploaded to each category, and calculate the fraction of its videos in the top category. We refer to this measure as the *top-category ratio*. If all the videos of an uploader belong to a single category, then its top-category ratio is 1. Similarly, we define the top-three-category ratio, which is the fraction of videos belonging to the top three categories of the uploader. To have a yet more comprehensive measure of upload category diversity, we introduce the category entropy for uploader i, which is expressed mathematically as:

$$e_i = \frac{-\sum_{k=1}^{K} \frac{u_{ik}}{u_i} \ln \frac{u_{ik}}{u_i}}{\ln K}, \quad (1)$$

where K is the number of categories, u_{ik} is the number of uploaded videos from user i to category k, and $u_i = \sum_{k=1}^{n} u_{ik}$ is the total number of uploaded videos of user i. If an uploader uploads all videos to a single category, then its category entropy will be 0; while if an uploader uploads uniformly to all 15 categories, then its category entropy will be 1. Note that category entropy takes into account the videos uploaded to all categories for individual uploaders, unlike the ratio of the top category (and the top three categories) which only takes into account the top categories.

Figure 5 illustrates the category uploading behavior. We consider three groups of uploaders, namely, the uploaders that upload more than one, 10, and 100 videos respectively. Figure 5 (a) shows the

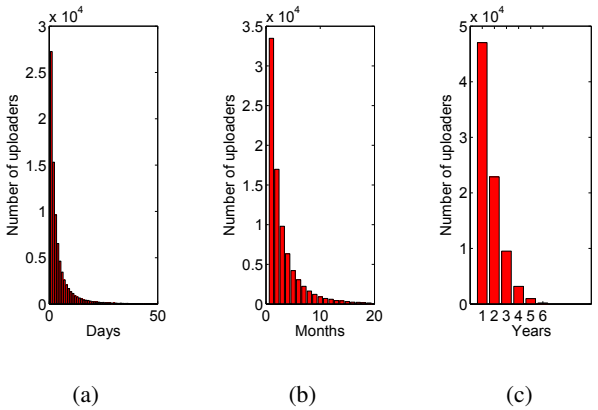


Figure 4: Distribution of active times with different time scales. (a) one day; (b) one month; (c) one year.

Table 2: Basic statistics of uploaders

	Maximum	Mean	Median
Number of videos	11,323	9.5	3
Number of views	61,130,983	31,143	847
Number of subscribers	138,583	19.6	1

CDF of the top-category ratio. Focusing on the uploaders who upload more than 100 videos, we see that about 20% of the uploaders only upload to a single category, and more than 85% of uploaders upload more than 50% of their videos to their top category. By considering the relatively large number of videos uploaded by this group of uploaders, *we see there is a strong tendency to concentrate videos in a small number of categories.* These can be further demonstrated in Figure 5 (b) and (c). As shown in Figure 5 (b), more than 70% of uploaders upload more than 80% of their videos to the top three categories. As shown in Figure 5 (c), more than 80% of uploaders have a category entropy less than 0.5. Interestingly, the category concentration trend appears to be independent of how active the uploader is. All three groups present similar behavior no matter how many videos the uploaders upload.

2.4 Uploader Views and Subscribers

In YouTube, some uploaders are more popular than others, with more users viewing their uploaded videos. YouTube incentivizes the very popular uploaders by sharing its ad revenue with them. To some extent, the number of views of an uploader influences its motivation to upload. We investigate the number of views on individual uploaders in the dataset `UploadSam`.

Figure 2 plots the uploader rank ordered by the number of uploader views in the dataset `UploadSam`. Surprisingly, *the number of views does not follow a Zipf distribution*. Our conjecture is that uploader popularity has been tuned by YouTube intentionally by YouTube's recommendation system (that is, the related videos). On one hand, there are fewer unpopular uploaders than there would be in a Zipf distribution. Specifically, more than 47% of uploaders attract more than 1,000 uploader views, implying that YouTube uploaders enjoy good visibility, even for those who upload a very small number of videos. Thus, YouTube's recommendation system (that is, related videos) seems to be somewhat biased towards less popular uploaders. On the other hand, among all uploaders, the most popular 20% of the uploaders attract 97.0% of views, which does not follow the 80-20 rule. Thus YouTube's recommendation

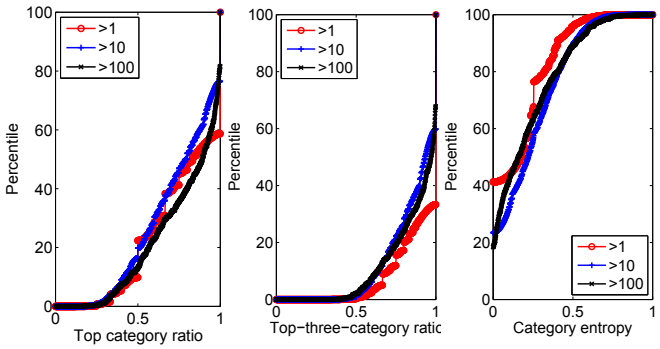


Figure 5: Category-concentrated uploading. (a) uploaded video ratio of the top category; (b) uploaded video ratio of the top-three categories; (c) uniformed category entropy.

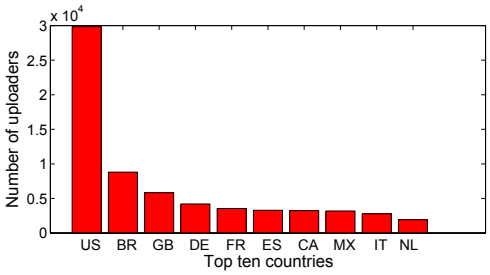


Figure 6: Distribution of geographic location .

system seems to be also biased to the very popular uploaders. The maximum, mean, and median of uploaded videos are reported in Table 2. We briefly remark that there are also many very unpopular uploaders in our random sample: 13.5% of the 100,000 uploaders have less than 100 views (aggregated across all of the uploader's videos).

Figure 3 plots the uploader rank ordered by the number of subscribers in dataset `UploadSam`. We observe that the number of subscribers follows a Zipf distribution. *Therefore, although a recommendation system may be able to shape uploader views, it is harder to change Internet users preference when it comes to subscriptions, which is a more active action than having a "quick view" upon a recommendation.* The maximum, mean, and median of number of subscribers are reported in Table 2.

2.5 Gender, Age and Location

YouTube uploaders may provide gender information in their profiles. In `UploadSam`, 97.9% of uploaders specify their gender information, 91.2% specify their ages, and 90.0% specify both. *We find that there are more than three times the number of male uploaders than female uploaders*. Consistent with the number of uploaders, male uploaders contribute more than three times the number of videos and attract more than three times the number of views than female uploaders. We also observe that the uploaders with ages from 20 to 30 are the most active uploaders, contributing approximately 40.0% of YouTube videos.

YouTube is a US-based Internet application, but it is widely used all over the world. We investigate the geographic location of the YouTube uploaders. The YouTube API provides the uploaders' country information. We carefully examine the geographic location information for the uploaders in dataset `UploadSam` and eliminate 4,137 uploaders whose location information is noisy.

Figure 6 shows the number of uploaders from the top ten countries. As expected, the US is the top country in terms of number of uploaders, accounting for 31.1% of the total YouTube uploaders. The other top nine countries account for more than 43.7%, while the remaining countries account for 25.2%. In addition to the US, YouTube is very popular in Europe and South America. However, it is less popular in Japan and Korea, which is surprising since these two countries have high Internet usage. YouTube is not accessible from Mainland China.

2.6 Social Uploaders

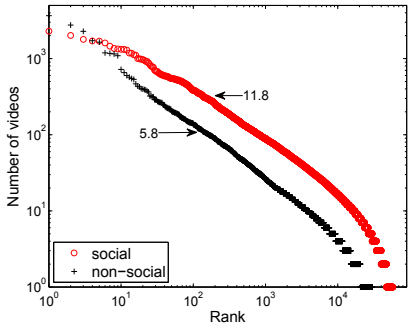
Some YouTube users join this social network (who we call *social users*), while others are not (who we call *non-social users*). We now study how the YouTube social network influences uploader behavior. We first ask, what fraction of the social users are uploaders, and what fraction of YouTube users on the whole are uploaders? To this end, using breath-first search, we crawl a large number of YouTube users (10,000,000) in the YouTube friend social network. We then randomly select a subset of 100,000 users (named `SocUsr`) to investigate whether these social users are uploaders. We find that 43% of the 100,000 social users in `SocUsr` have uploaded at least one video. To estimate the fraction of YouTube users on the whole who are uploaders, we assume that there are totally 135 million U.S. YouTube users. Note that this is a very conservative estimate, since it has been reported that 135 million unique U.S. viewers viewed YouTube videos in December 2009 [8]. Recall that approximately 31.1% of the uploaders are from the U.S. and that there are around 47.3 million uploaders in YouTube. By considering both social and non-social YouTube users, we estimate the fraction of YouTube users who are uploaders to be at most $\frac{47.3}{135} * 0.311 = 0.109$. In summary, whereas at most 10.9% of the YouTube users are uploaders, 43% of the social users are uploaders. *Thus, social users are much more inclined (at least four times more!) to be uploaders than the average YouTube users.*

To compare the behavior of the social uploaders and non-social uploaders, we re-visit the dataset `UploadSam` and group the uploaders into social uploaders and non-social uploaders, based on whether an uploader belongs to the social network. Figure 7 (a) plots the rank of uploaders ordered by the number of uploaded videos. Figure 7 (b) plots the rank of uploaders ordered by the number of uploader views. *These figures clearly show that social uploaders upload more videos and receive more views – on average twice as many – than their non-social counterparts.*

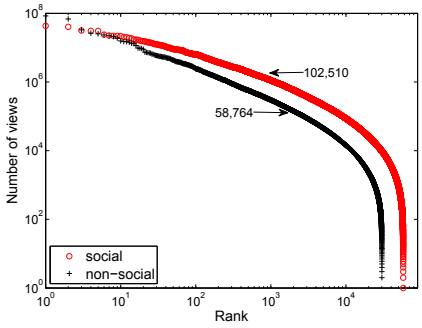
There are two possible reasons why social uploaders upload significantly more than non-social uploaders. On one hand, there is a feedback loop: when a social uploader publishes a new video, it will likely be viewed by many of the uploader's friends; this positive feedback incentivizes the social user to upload even more. On the other hand, users who upload a lot are more likely to join the social network, since they are already very active in YouTube. Without actually interviewing the uploaders, it is difficult to say which of these two reasons weighs more heavily in user behavior. However, it is reasonable to believe that these two reasons together explain why social uploaders are more driven to contribute.

3. USER GENERATED OR USER COPIED?

With its slogan "Broadcast Yourself", YouTube has been designed for sharing user generated content. However, a significant fraction of videos are not originally from YouTube, but are instead simply copied from other places, such as movies, TV shows, and other professional video websites. Thus, content on YouTube can



(a)



(b)

Figure 7: Social uploaders vs. non-social uploaders. (a)Number of uploaded videos; (b) number of uploader views.

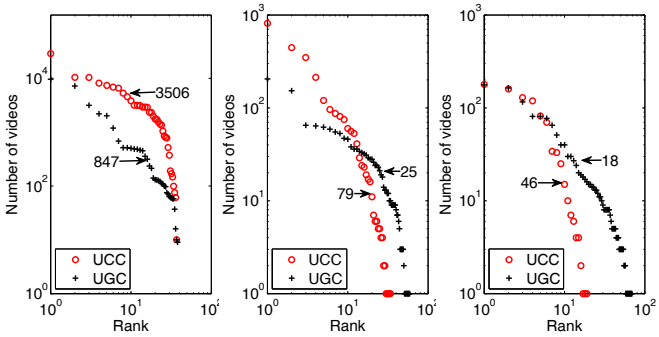
be classified as User Generated Content (UGC) or User Copied Content (UCC). In this section, we seek to classify each uploader as a UGC uploader, a UCC uploader, or a combined UGC/UCC uploader. As a consequence of the classification, we will be able to estimate the importance of UCC content in YouTube.

The classification of a video as either UGC or UCC is not always black and white. We begin this analysis by defining UGC and UCC videos:

- **UCC:** A contiguous snippet of a video that was originally distributed outside of YouTube. For example, a snippet of a movie, of a TV show, or of anything originally broadcasted over TV. This also includes videos created by professional video producers, such as NBC. These professional uploaders publish these videos on their own websites, and then on YouTube to broaden their visibility.
- **UGC:** A video that is originally generated for YouTube-like video sharing websites. This not only includes videos captured with user's digital cameras and webcams, but also mashups of copied video snippets.

We now manually classify uploaders and videos as either UGC or UCC. We first create a dataset for three groups of uploaders:

- **Popular uploaders:** This group contains the 100 most popular uploaders, as explicitly presented in the YouTube website. Although it is unclear to us exactly how YouTube decides which uploaders to include in this list of 100 uploaders, in general, each of these uploaders attracts a large number of uploader views.
- **Social uploaders:** This group contains 100 randomly chosen uploaders from the YouTube social network. As shown in Section 2.6, the uploaders in the YouTube social network are in general more active than those not in the social network.
- **Random uploaders:** This group contains 100 randomly selected uploaders from `UploadSam`. This group represents the typical behavior of YouTube uploaders.



(a) (b) (c)

Figure 8: UCC uploaders vs. UGC uploaders in terms of number of uploaded videos. (a) Popular uploaders; (b) Social uploaders; (c) Random uploaders.

Table 3: Manual classification on uploaders

	UCC	UGC	non-dominate
`Popular:`	63	36	1
`Social:`	34	61	5
`Random:`	20	77	3

3.0.1 *UGC or UCC Uploader?*

We first examine whether YouTube uploaders upload mostly UGC content or mostly UCC content, but not both types of content. We say an uploader is a UCC (respectively, UGC) uploader if more than 90% of its videos are UCC (respectively, UGC).

As shown in Table 3, for each of the three groups, *more than* 95% *of the uploaders are either UGC uploaders or UCC uploaders*. This is a critical observation. One immediate application of this finding is, for automatically classifying videos as UGC or UCC, in addition to the features of individual videos (e.g., number of video views, video title, video category, etc.), one can indirectly classify the video by classifying its uploader. Thus, the uploader information provides useful features for automatic video classification.

Although the fraction of non-dominant uploaders is small for the three groups, the relative fractions of UCC/UGC uploaders is strikingly different among the three groups. Observe that 77% of the random uploaders are UGC uploaders. Thus most of the users in YouTube are indeed publishing user generated content. However, in the popular group, 63% of uploaders are UCC uploaders, implying that *the majority of the most popular uploaders are not uploading user generated content!* We mention that many (but not all) of the uploaders in the popular group are professional video producers, such as NBC.

3.0.2 *Number of uploaded videos*

We now investigate the detailed properties of UCC/UGC uploaders. For both UGC and UCC uploaders, and for each of the three groups, Figure 8 plots the uploader rank ordered by number of up-

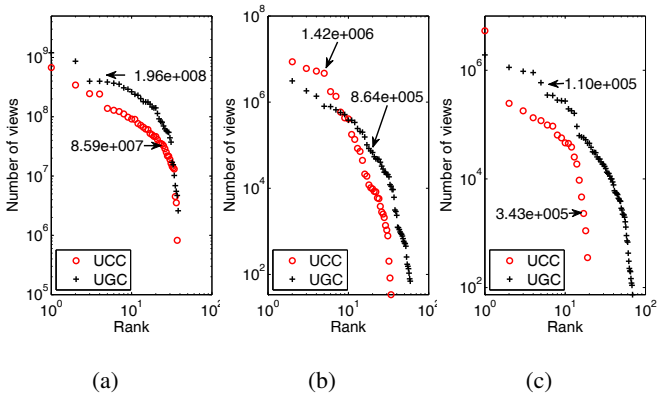


Figure 9: UCC uploaders vs. UGC uploaders in terms of number uploader views. (a) Popular uploaders; (b) Social uploaders; (c) Random uploaders.

loaded videos. The figure also indicates the average number of uploaded videos for each curve. We observe that, on average, *UCC uploaders upload many more videos than UGC uploaders.* This is probably because it is much easier for a UCC uploader to generate video clips from movies and TV shows than it is for a UGC uploader to generate original content. This figure clearly shows that video content influences an uploader's uploading behavior.

3.0.3 Views and subscribers

We have shown that UCC uploaders generally upload more videos than UGC uploaders. This does not necessarily mean, however, that UCC uploaders are more popular than UGC uploaders in terms of number of views. Figure 9 plots the uploader rank ordered by number of views for UCC/UGC uploaders in different groups. Similarly, we indicate the average number of views for each curve. Surprisingly, *in the popular group, the UGC uploaders on average attract more than twice as many views as the UCC uploaders, even though many of the popular UCC uploaders publish professionally generated videos and upload more videos.* This demonstrates that ordinary Internet users are capable of creating more interesting and popular channels than the professional producers! A similar observation can be made for the uploaders in the social network group and the random group: except for the top uploaders, UGC uploaders are in general more popular. However, UCC uploaders are on average more popular than the UGC uploaders in the social network group and the random group, because the very top UCC uploaders attract significantly more views than the very top UGC uploaders.

We have seen that several video features (video category and video title) and several video uploader features (number of uploaded videos, number of active times, maximum number of uploaded videos in a day, gender, and uploader dominant category) provide hints about whether a particular video is UGC or UCC. In future work, we will explore using machine learning with these features for automatic classification of videos as UGC and UCC.

4. RELATED WORK

As mentioned in the Introduction, there are several studies on measuring and analyzing YouTube. Mislove *et al.* investigated several on-line social networks, including YouTube [13]. Their study mainly focuses on the properties of social graphs, e.g., the power-law, small-world, and scale-free properties. Cha *et al.* comprehensively studied YouTube from the perspective of YouTube videos and viewers [4]. The authors studied the popularity life-cycle of videos, the statistical properties of requests and their relationship with video age, and the level of content aliasing and illegal content. From the system design perspective, they also discuss the potential of using P2P in UGC VoD systems. Similar to [4], Cheng *et al.* investigated the video properties of YouTube [6]. They found that the related video network forms a small world network. Leveraging the small world property, Cheng *et al.* optimized P2P system design to effectively distribute videos for YouTube [7]. Zhou *et al.* study the view sources of YouTube videos, and find that related video recommendation is one of the most important sources [18].

Gill *et al.* [10] investigated YouTube from the perspective of YouTube traffic. They examined YouTube usage patterns, file properties, and transfer characteristics. They also investigated the YouTube video properties, such as video popularity and referencing behaviors. Zink *et al.* also investigated YouTube from the perspective of YouTube traffic [19], with a focus on measurement methodology and modeling.

Benevenuto *et al.* investigated the user behavior in a social network created by video interactions, and characterized a social network created by the video interactions among users in YouTube [2]. They identified typical user behavior patterns and showed evidence of anti-social behavior such as self promotion and content pollution. Benevenuto *et al.* considered the problem of detecting video spammers, and applied machine learning to provide a heuristic for classifying an arbitrary video as either legitimate or spam [3]. Qiu *et al.* [15] and Cha *et al.* [5] studied the user behavior in large scale IPTV systems. Cuevas *et al.* [9] studied the user behavior of content publishers in BitTorrent.

With its slogan "Broadcast Yourself," YouTube has been designed for sharing user generated content, where uploaders play the critical role. Surprisingly, to our knowledge, none of the prior studies carefully examines YouTube uploader properties in a systematic and comprehensive way. Our study fills this gap and paints a precise picture of YouTube uploaders.

5. CONCLUSION

As Internet researchers, we are interested in the characteristics and behavior of the content creators, arguably the most important entities in the Internet phenomenon. By identifying which uploaders in the near future will likely attract many views, advertisers can make more informed decisions. Google currently makes little information publicly available about the scale of YouTube. To understand YouTube's relative importance in the Internet landscape, we need to have estimates of it scale, which are provided in this paper.

In this paper, we provide a comprehensive study on YouTube uploaders, the central agents in the YouTube phenomenon. We conduct extensive measurement and analysis and obtain an in-depth understanding of YouTube uploaders. We estimate YouTube scale and examine the uploading behavior of YouTube users. Furthermore, we examine whether YouTube users are really broadcasting themselves, via characterizing and classifying user generated videos and user copied videos. Moreover, we demonstrate the positive reinforcement between on-line social behavior and uploading behavior.

Perhaps the most surprising result in the paper is the discovery that much of the content in YouTube is not user generated. We found that 63% of the most popular uploaders are primarily uploading UCC content, and that UCC uploaders on average upload many more videos than UGC uploaders. The results and observations in Section 3 can be used as a first step towards an automatic algorithm for classifying UGC and UCC content.

6. REFERENCES

[1] BALUJA, S., SETH, R., SIVAKUMAR, D., JING, Y., YAGNIK, J., KUMAR, S., RAVICHANDRAN, D., AND ALY, M. Video suggestion and discovery for youtube: Taking random walks through the view graph. In *WWW* (2008).

[2] BENEVENUTO, F., DUARTE, F., RODRIGUES, T., ALMEIDA, V., ALMEIDA, J., AND ROSS, K. Understanding video interactions in youtube. In *ACM Multimedia* (2008).

[3] BENEVENUTO, F., RODRIGUES, T., ALMEIDA, V., ALMEIDA, J., ZHANG, C., AND ROSS, K. W. Identifying video spammers in online social networks. In *Fourth International Workshop on Adversarial Information Retrieval on the Web* (2008).

[4] CHA, M., KWAK, H., RODRIGUEZ, P., AHN, Y., AND MOON, S. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *IMC* (2007).

[5] CHA, M., RODRIGUEZ, P., CROWCROFT, J., MOON, S., AND AMATRIAIN, X. Watching television over an ip network. In *IMC* (2008).

[6] CHENG, X., DALE, C., AND LIU, J. Statistics and social networking of youtube videos. In *IEEE IWQoS* (2008).

[7] CHENG, X., AND LIU, J. Nettube: Exploring social networks for peer-to-peer short video sharing. In *INFOCOM* (2009).

[8] COMSCORE. http://comscore.com/.

[9] CUEVAS, R., M.KRYCZKA, CUEVAS, A., KAUNE, S., GUERRERO, C., AND REJAIE, R. Is content publishing in BitTorrent altruistic or profit-driven? In *ACM CoNext* (2010).

[10] GILL, P., ARLITTZ, M., LI, Z., AND MAHANTIX, A. Youtube traffic characterization: a view from the edge. In *IMC* (2007).

[11] KOCH, R. *The 80/20 Principle: The Secret to Success by Achieving More with Less*. Broadway Business, 1999.

[12] LANGE, P. G. Publicly private and privately public: Social networking on youtube. *Journal of Computer-Mediated Communication 13*, 1-2 (2007).

[13] MISLOVE, A., MARCON, M., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. Measurement and analysis of online social networks. In *IMC* (2007).

[14] PAOLILLO, J. C. Structure and network in the youtube core. In *HICSS* (2008).

[15] QIU, T., GE, Z., LEE, S., WANG, J., XU, J., AND ZHAO, Q. Modeling user activities in a large IPTV system. In *IMC* (2009).

[16] SAXENA, M., SHARAN, U., AND FAHMY, S. Analyzing video services in web 2.0: A global perspective. In *NOSSDAV* (2008).

[17] SIERSDORFER, S., PEDRO, J. S., CHELARU, S., AND NEJDL, W. How useful are your comments?- analyzing and predicting youtube comments and comment ratings. In *WWW* (2009).

[18] ZHOU, R., KHEMMARAT, S., AND GAO, L. The impact of youtube recommendation system on video views. In *ACM IMC* (2010).

[19] ZINK, M., SUH, K., GU, Y., AND KUROSE, J. Characteristics of youtube network traffic at a campus network - measurements, models, and implications. *Journal of Computer-Mediated Communication 53*, 4 (2009).

Summary Review Documentation for

"Broadcast Yourself: Understanding YouTube Uploaders"

Authors: Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, A. Ghose

Reviewer #1

Strengths: Amazing result on the ability of user generated content to compete directly with professionally produced content.

Very nice results on the (estimated) size of the YouTube universe. I particularly liked the methodology for doing that (high level scalable crawl of IDs, detailed profiling of a subset, and then final extrapolation from the combination of the two).

Excellent insight on the impact of recommendation on number of views and its relatively less strong impact on number of subscribers.

The fact that uploaders tend to be highly concentrated (specializing) in very few categories is also new and gives interesting insights into the behavior of uploaders of YouTube.

Weaknesses: The paper fails to cite a highly related reference: R.Cuevas, M.Kryczka, A. Cuevas, S. Kaune, C. Guerrero and R. Rejaie, "Is Content Publishing in BitTorrent Altruistic or Profit-Driven?", Proc. ACM CoNext 2010.

Comments to Authors: **N**ot much to add. I greatly enjoyed reading this paper. My sincere compliments to the authors.

The only glitch that I see (more of a wish probably) is that the authors do not put Youtube uploaders in context by comparing them with the uploaders of other systems eg BitTorrent or 1-click hosting sites (rapidshare). For BitTorrent there exists a recently published paper (Is Content Publishing in BitTorrent Altruistic or Profit-Driven?) and I believe that there have also been some studies of 1-click hosting.

In Section 2.1: during the last weeks of the crawler when you saw a drop on new ID's returned per hour did you also see a drop in the overall number of ID's that you saw per hour. If yes then this might mean that YouTube was rate limiting your crawler, leaving open the possibility that you have not seen all the uploaders.

Reviewer #3

Strengths: While significant studies have been done on YouTube, the paper finds a new angle.

Weaknesses: The paper focuses on the data analysis and falls short on the implication of the analysis results. Most of the findings are expected.

Comments to Authors: The paper is well written and has solid analysis of the interesting youtube data. Most of the findings are expected, such as socially active users tend to upload more, males tend to upload more than female, and significant user generated content being uploaded.

The paper studies the content based on the videos uploaded by 100 users selected in three different ways. How representative are the data given it's still very small percentage of the users? Perhaps increasing the number of users to be studied or having some sensitivity analysis by using a different number of users and see if the same conclusions hold.

It'd be great to see some more applications about the findings.

Reviewer #3

Strengths: Reasonably well executed measurement study, rather useful results, well written paper.

Weaknesses: Measurement methodology has some problems and do not seem to capture proper datasets which affects accuracy of presented results.

Comments to Authors: This paper presents a measurement study for characterizing youtube uploaders. The presented analysis is sound and the results are somewhat interesting. However, authors do not in general describe implications of their findings (with a couple of exceptions).

The main concern with this paper is the soundness of its measurement methodology that affects the accuracy of its datasets (samples) and thus the correctness of the presented results. Here are more details on this problem with the methodology:

Authors select a seed user, collect her uploaded video, crawl all the related videos and then identify their uploaders. They repeat this process by selecting a new seed user. The key question is how these seed users are selected? If these seeds are randomly selected, then the authors are likely to have a reasonably complete set of uploaders. Otherwise, the captured uploaders are likely to be a (possibly small) fraction of uploaders. Authors do not describe their strategy for seed selection and it does not seem trivial to select a random user since user IDs do not have numerical forms. It appears that authors selected known users as seed. Therefore, this approach is likely to capture only the uploaders in the largest connected component, and is going to miss uploaders in other components and singleton ones. The conducted test for completeness via comparison with DFS crawl of social network is not reliable because the seeds were selected in the same region of the graph and do not reveal other components of the graph. Youtube does not seem to have a strong social network (like facebook or other social networks). Therefore it is very likely that this approach misses a large number of

uploders. This could significantly affect the presented results that are the main contributions of this paper.

It would be useful that authors provide some information about the speed of crawling and key factors that may affect this. The crawling rate seems to be around 21K users/hour, which is modest.

Using BFS approach to crawl YouTube friendship network also reveals only high degree users in SocUser dataset. It is not clear why authors use just 100K samples from SocUsers. Since there is not need for further data collection for SocUSers, it seems useful to consider all of them for related analysis.

It is odd to see that there are no uploaders with a small (<100) number of views in Fig 2, this suggests that data mostly has captured bigger (and more visible) uploaders, which relates to the problem with data collection mentioned earlier.

In section 3, clearly determining whether a video is UCC or UGC is difficult and requires a manual approach. But it is not clear that this can be done reliably even in a manual fashion, For example how do you know whether a mash-up video is created for Youtube or it already used in a tv program. Authors do not elaborate on difficulties related to classification of videos into these two classes including the reasons for their inability to classify a few videos in each dataset.

The provided reasons for various findings in the paper are mostly intuitive guesses without any evidence to back them up.

Reviewer #4

Strengths: This is a new take on a well-mined problem -- looking at uploaders instead of videos etc.

Weaknesses: I find the motivation for this work rather weak -- I dont see a good reason why understanding uploaders is interesting other than a simple curiosity-driven study (and the authors don't really point out any either). Also, many parts of the paper seem to be rehashing well known or well-expected results.

Comments to Authors: This is a reasonably nice and in-depth study of the uploader behaviors as a self-contained topic. My biggest problem with the paper as is I am not sure why we should care about this uploader behavior beyond a "curiosity" result! I would really like to see more implications or motivation for these results - how would your findings change the behavior dynamics of Youtube, viewing patterns, content caching, etc?

Section 2.1: How do you bootstrap the initial uploader list before you do the crawl?

Maybe I don't understand what is happening here, but I found the "sampling" towards the end rather puzzling. Arguably, you have down the hard work of collecting the entire graph, why do you need to down sample it; 43 million users should be pretty easy to analyze using a beefy machine? Put another way, if you were going to do this sampling anyway, why go through the trouble of collecting the full user list?

Section 2.2: For the numbers you report in Sec 2.2 - I remember seeing a bunch of nice Youtube stats at this URL earlier: http://www.youtube.com/t/fact_sheet; but Google seems to have removed this content now.

Also, given this "heavy-tail" behavior is the "linear scaling" justified?

Section 2.3: I was also mildly annoyed by your references to the 80-20 rule; that to me is just a "heuristic" observation not a golden rule for every human-driven system. So the references to "X follows this but Y does not" seem to be out of place. All of them are skewed distributions and the data shows that; whether its 80-20 or 90-20 or 70-20 to me seems somewhat secondary.

How do you identify the "Categories" - I am particularly worried that the "concentration" in category you observe is merely an artifact of incorrect categorization or Youtube tagging it automatically into some default category.

Section 2.4: I find your claim that "Youtube's related video recommendation seems to be biased toward unpopular" pretty darn speculative. I can imagine a whole slew of other causes for this; e.g., most of these uploaders are video spammers, many of the views are from click bots, Youtube also posts a bunch of recent most videos on the home page (maybe many views come from that?)

The italicized conclusion at the end of this section too seems a bit doubtful in this light.

Section 2.5: What is SamRanUpl -- a new/different dataset? Why does age/gender matter?

Section 2.6: Maybe this reflects my ignorance, but what exactly do you mean by "the social network" or "join" here? Is it someone explicitly invoking a friend relation to another user?

Also, this could just be a wording issue, but some of your conclusions at the end of Sections 2.6 and 3.0.2 seem to suggest an implicit "causal" relationship, when there could be hidden variables or the causality could in fact be reverse, e.g., social users are more inclined - it could be that "active" users tend to be social and upload more. That's the hidden causality!

Section 3: Two high-level comments about this section. First, I dont really understand how you classify UGC vs UCC; the text is vague about this. Second, I am not sure I agree with what you call UCC; NBC uploading trailers or snippets are not "copying" anything - I got the feeling you wanted to distinguish content that the user "authored" herself vs. content that the user "borrowed" from others and reposted? Your definition of copying (which is vague, btw) seems the muddle these two.

Section 4: do you apply the techniques for detecting spammers/video spam to filter your dataset. I wonder if the results will change if you do that.

Reviewer #5

Strengths: - Interesting analysis of the uploaders' profile (to the best of my knowledge, it is the one of the first studies at this level on Youtube uploaders)
- The paper reads pretty well.

Weaknesses: - The analysis is somewhat superficial, should be more careful before drawing conclusions. Assessing that the number of views/subscribers does not/does follow a Zipf distribution needs some more rigorous justification than a graphical plot.
- There may be a bias in the sampling process to construct the dataset, and the paper does not present compelling sanity checks on this point.

Comments to Authors: The paper presents an interesting set of observations, but suffers from two main problems in my view:

1. The claims about distributions are made by visual inspection of cdf plots, which is not rigorous (e.g, that Figure 2 shows that the number of views does not follow a Zipf distribution but that Figure 3 shows that the number of subscribers does follow a Zipf distribution). QQ plots, and even better, hypothesis testing, should be conducted to have a more quantitative validation. Given the emphasis given to these observations in the paper, this more rigorous validation is needed.

2. The sampling process may be biased and corrupt the analysis, for example in the way you "randomly (that is, uniformly over the set of all users?) select a subset of uploaders and conduct deeper investigation". In addition DFS adds a bias favoring high degree nodes of the graph (but which fades away if the coverage is large, see Kurant et al's paper on unbiased sampling, IEEE JSAC, 2011). The paper does not convince the reader that the proper sanity checks have been made in order to be sure that these measurement biases do not invalidate the results.

Response from the Authors

We would like to thank the reviewers for their insightful comments. We included the suggested reference but otherwise did not make any significant changes to the paper.

As Internet researchers, we are interested in the characteristics and behavior of the content creators, arguably the most important entities in the Internet phenomenon. By identifying which uploaders in the near future will likely attract many views, advertisers can make more informed decisions. As mentioned by one of the reviewers, Google currently makes little information publicly available about the scale of YouTube. To understand YouTube's relative importance in the Internet landscape, we need to have estimates of it scale, which are provided in this paper.

We are confident that we crawled the majority of the Youtube uploaders. In the last weeks, our crawler, based on related videos, continued to see a large number of uploaders every hour; but, as indicated in the paper, the number of previously un-observed uploaders dropped off dramatically in the final weeks. Most importantly, we validated our claim by examining the overlap with YouTube's social network. The related video network and the social network are two separate networks. The social network is created by the YouTube users, and the related video network is created by YouTube. We acknowledge that there may be some singleton uploaders who are not included. We also did some small studies (not reported in paper) with some low-view uploaders (the authors and their friends) and found that they were all included in our crawl. Also, 13.5% of the 100,000 uploaders have less than 100 views in our data set.

Perhaps the most surprising result in the paper is the discovery that much of the content in YouTube is not user generated. We defined UCC content in the submitted paper as: A contiguous snippet of a video that was originally distributed outside of YouTube. We believe this definition is quite clear. We manually examined all of the videos from 300 uploaders (100 from each of the three categories), a very laborious and painstaking process. Of course, we would prefer to use a larger sample size. We found that 63% of the most popular uploaders are primarily uploading UCC content, and that UCC uploaders on average upload many more videos than UGC uploaders. The results and observations in Section 3 can be used as a first step towards an automatic algorithm for classifying UGC and UCC content.