Call me Guru: user categories and large-scale behavior in YouTube

Joan-Isaac Biel and Daniel Gatica-Perez

Abstract While existing studies on YouTube's massive user-generated video content have mostly focused on the analysis of *videos*, their characteristics, and network properties, little attention has been paid to the analysis of *users' long-term behavior* as it relates to the roles they self-define and (explicitly or not) play in the site. In this chapter, we present a statistical analysis of aggregated user behavior in YouTube from the perspective of user categories, a feature that allows people to ascribe to popular roles and to potentially reach certain communities. Using a sample of 270,000 users, we found that a high level of interaction and participation is concentrated on a relatively small, yet significant, group of users, following recognizable patterns of personal and social involvement. Based on our analysis, we also show that by using simple behavioral features from user profiles, people can be automatically classified according to their category with accuracy rates of up to 73%.

1 Introduction

Social media sites have become mainstream publishing and communication tools that have globally changed media production and consumption patterns. Among them, YouTube is one of the best examples of this explosion of online user-generated content receiving 24h of new videos every minute (the equivalent of 140,000 Hollywood movies per week) [2], and surpassing 2 billion views per day [1]. While the first key achievement of YouTube was the creation of an easy-to-use integrated

Joan-Isaac Biel
Idiap Research Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
e-mail: jibiel@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
e-mail: gatica@idiap.ch

platform aimed to upload, share, and watch online videos, thus removing barriers to the online video scene, it further promoted participation allowing users to comment, rate, explore, and post related videos. As a result, YouTube has become a site for people and communities to join and interact, from aspiring rockstars to top politicians.

Recently, research on YouTube has analyzed the statistics and social network of uploaded videos, revealing properties of the nature of video sharing systems that may be key for the future of such services [6, 7, 18]. However, few works have focused on characterizing long-term user behavior Existing works have provided a brief statistical analysis on the use of YouTube social-oriented features [11], studied the properties of the social network of friends and subscriptions [18], and a similar characterization based on user video interactions [3]. Understanding how users typically behave and interact, and how they perceive YouTube as both a system and a social outlet is fundamental to improve its performance. Despite this realization, existing work has mainly treated YouTube users as a single homogeneous group, ignoring potential differences between groups of users. We believe that this is an essential point when analyzing users in large social networks, due to the likely wide variety of behavioral patterns.

This chapter presents one of the the first attempts to analyze large-scale aggregated behavior of YouTube users under the lens of *user categories*, i.e., self-assigned roles that people can choose among *Director*, *Comedian*, *Guru*, *Musician*, or *Reporter*. We hypothesize that users' choices and the implicit or explicit ways in which people respond to these roles give rise to different collective behavior, which can be measured quantitatively. Our work has two contributions. First, on a large user dataset, we present a statistical analysis of YouTube users and categories based on easy-to-extract, long-term user behavioral features, which do not require any video or metadata processing. Our analysis reveals clear trends regarding people's category choices, and differences on user behavior (both individual and social) across categories and gender, which suggest that the emerging communities do have differences that could lead the way to automatic modeling of groups of users. Second, we use such behavioral cues in various classification tasks, in order to explore whether, alone or together, they can be used to infer user categories, obtaining promising performance. Overall, our work aims at complementing the emerging (and much needed) work in sociology and ethnography on the understanding of users' motivations to select roles and to create and maintain self and group identities in social media outlets like YouTube, and enquires about some fundamental needs for personalized applications.

This chapter is organized as follows. Section 2 reviews some of the recent literature on YouTube, with focus on user behavior research. Section 3 provides a basic overview on YouTube channels user categories. Section 4 describes the set of behavorial features extracted from the use channels. Section 5 presents a statistical analysis of the feature distributions to study the differences between different user categories. Section 6 focuses on the use of behavioral features for user category classification. Finally, Section 7 summarizes the chapter and discusses future work.

2 Research in YouTube

Today, research in social media has mainly studied YouTube as a video repository system and as a user generated content site, focusing on the analysis of the network of videos and the typical characteristics of video content itself. Using analytical methods from social networks to investigate the macroscopic characteristics of the network of videos, works have studied the impact of user generated content in underlying video-on-demand architectures [6, 7], and in local networks [9, 23]. By concentrating specifically on videos, research has also studied the daily cycles of video reception (based on the number of views), in an attempt to automatically predict their popularity [22]. In relation to video content, several studies have manually coded samples of videos to categorize the types and properties of YouTube content to gain understanding about user-generated media production [13, 12, 5]. Furthermore, other works have attempted to automatically model the topic and ideological perspective of content [16].

Compared to the aforementioned approaches, research on YouTube user behavior, rather scarce specially with respect to automatic analysis, includes a variety of research questions, data sample sizes, and methods. From sociology to communication, some works have focused on studying the practices of small groups of people (typically less than hundred) to understand why and how people participate in online video-sharing sites like YouTube. Some works have done so by observing both the computer-mediated and offline behavior of people to understand the purposes of using YouTube in everyday life [15] and the influence of feedback, criticism and hate behaviors [14]. Others have analyzed videos and their related text content to study how users present themselves from the perspective of creating an online identity [10]. A third set of works have used manual coding and questionares to analyze the processes of creation and reception of online video [19, 21]. Overall, due to their nature, the analysis methods used in all these works are not extensible to study larger samples, which limits the statistical significance of their findings. In addition, findings obtained from the study of small groups and specific communities (e.g., school students or elders) are likely to generalize poorly to the large population of YouTube users. In contrast, very few works in computer science have been devoted to the study of user behavior on large-scale data samples gathered using YouTube's API or web crawlers. Using statistical methodologies, they have focused on analyzing the structure and topology of the social networks that emerge from user interaction, to reveal network characterstics that are relevant for information and communications services [18], and to detect anti-social and spam behaviors [3]. Despite the relevance of their findings, these works do not provide any understanding on the nature of users' behaviors and their motivations to participate and to create and maintain self and group identities.

Few works have focused on studying user behavior as in our work, using the metadata attributes available on YouTube to characterize common user and group patterns that arise from long-term behavior. Halvey and Keane [11] provided a first statistical analysis on the use of some features such as the number of videos watched,

uploads, friends, subscriptions, groups, and comments. Their study evidenced that a big part of users participate in YouTube as consumers rather than as contributors of content, and that few users participate actively, both contributing with content and interacting socially with other users, as it is shown by the power-law behavior of some of these feature distributions. Furthermore, Maia et al. [17] identified different groups of users based on a similar set of social features. Using an unsupervised clustering algorithm, and estimating the number of clusters automatically, they found five different groups of users which were respectively characterized as 1) not active, 2) producers, 3) consumers, 4) producers and consumers, and 5) others. Compared to Halvey and Keane [11], who treated YouTube users as a whole, our analysis emphasizes the potential differences between groups of users. We believe that this is an essential point when analyzing users in large social networks, due to the likely wide variety of behavioral patterns. In addition, compared to Maia et al. [17], our work differs in our focus on predetermined, self-assigned user categories, investigating the significant differences among their behaviors. A preliminary version of our work appeared in [4].

3 YouTube channels and user categories

Most of the people watching videos in YouTube are familiar with the default YouTube *video page* (see Figure 1, left), which allows users not only to watch a video, but to leave comments usually referred to the video, browse related videos suggested by YouTube, or navigate to video answers left by other users. In addition, YouTube provides a *user channel* to registered users (see Figure 1, right), which is equivalent to a user profile in other social networks.

The user channel is a dedicated user page that gathers all the videos uploaded by the user, and thus facilitates interested people to browse videos of a specific user, uploaded at anytime. Furthermore, it includes user information such as the name, location, age, personal description, and interests, together with user participation statistics such as the number of videos uploaded or watched, and links to the user subscriptions, subscribers, and friends. The user is allowed to control the information made public on the channel, as well as, to some extent, customize the look of the channel. In addition, users visiting a channel can leave comments which contrary to the comments in the video page, do no typically relate to any video in particular, but address the user in a more general way.

As with user channels, YouTube is continuously introducing new features to its platform, aiming to enhance the user experience responding to their demands as the community grows and diversifies. Special user categories are another example of features gradually introduced by YouTube and originally serving different purposes.

In April 2006, YouTube introduced the *Director* program in response to a video duration limitation earlier installed to prevent copyright infringement, most likely to be brought about by long videos. A user proving to be a legitimate creator of

Fig. 1 The YouTube video page (left) is designed to watch a video, leave comments, and navigate to related videos, which are not necessarily from the same user. In contrast, the user channel (right) is a user-dedicated page that gathers all the user's videos, in addition to personal information, usage statistics and comments addressed to the user.

Fig. 2 YouTube users can choose to belong to a specific category by changing the default settings of their user channels. Standard, Director, Musician, Comedian, Guru, and Reporter are open to everybody, whereas Politician and Non-profit are granted under special request.

his/her uploaded content could apply for a *Director* account that allowed to upload videos longer than 10 minutes (eventually, with the creation of other user account types, new *Director* accounts were limited to 10 minutes as well). In a span of five months, YouTube created special accounts for other users willing to promote their work. Users with a *Musician* or a *Comedian* accounts had the possibility to customize their user profiles by publishing performer information and a schedule of show dates. By June 2008, four more account categories had been introduced. *Guru* and *Reporter* accounts were respectively addressed to people devoted to create "how to" videos (i.e., videos that teach certain skills or explain how to do something) and to people dedicated to inform others about news and events occurring around them. While these accounts allowed anyone to sign up, two other ones, *Non-profit* and *Politician*, were only to be held by real non-profit organizations and politicians. The first one aimed to support advocacy campaigns and fundrising efforts. The second one was created for candidates of the 2008 United States presidential election and some other elections.

Today, a new user signing up for a YouTube account receives by default a *Standard* or *YouTuber* account, with the basic YouTube features such as uploading, commenting, etc. (we use the term *Standard* to avoid confusing *YouTubers* with all YouTube users). This status remains unchanged unless users *intentionally* modify their channel type to one and only one of the special categories (see Figure 2). In doing so, they are allowed a certain level of customization on their channel, which also exhibits a label with the name of the user category.

Why would users be interested in becoming special users? And why would they choose one category or another? A potential benefit for special users is that only them qualify for the *channels page* in YouTube, where the most subscribed and the most viewed channels are featured. Compared to browsing and searching, video and channel promotion in YouTube are advantageous ways of attracting viewers, and so that results in an obvious incentive for users. However, one may argue that this only benefits a small set of users, which then accumulate a high number of subscriptions and views. Moreover, though the different user categories are described in the YouTube help section, users are free to assign a user category to themselves independently of how well they fit with the respective category description, possibly augmenting the overlap among categories in terms of typical behaviors. For some users, like *Musicians*, the answer may be on the goodness of fit under a specific description. For others, we hypothesize that the process of self-assigning a specific user category conveys a sense of belonging to a particular community of users. If this is true, users might tend to acquire the same user category than the users he follows or interacts with, somehow reassuring his presence in that community. A way to verify this, could be to use networked data to compare the user category labels of users in a same social network, as well as to study the behavioral patterns of different communities of users. Alternatively, in our work, we directly focus on user categories, and hypothesize that users belonging to the same user category, do in fact behave in similar ways.

4 Extracting user behavioral features from YouTube

We gathered a dataset from YouTube user channels. Using YouTube's API, we followed a two-step data collection procedure that consisted on first obtaining the last uploaded videos from the site and extracting the username of the uploader, and then retrieving the channel information for every user. Since video search feeds are constantly updated with a short period of time, we repeated this procedure every 5 minutes, from March 5th to March 9th, 2009. We used video-category specific queries to overcome a 999 entries per feed limitation existing in the API feeds, thus augmenting our capacity to obtain the last uploaded videos.

We collected a dataset of 273,000 distinct users, for whom we obtained a set of descriptive behavioral features. We explicitly limited the set of features to all the attributes that can be easily extracted from the users' channel, being aware that there might be other features that are richer descriptors of behavior at higher computa-

tional cost. We divided the behavioral features in two different groups, capturing the individual participation and the social-oriented behavior of users, respectively. In addition, we extracted the user category information from the user channels.

The **individual participation features** are direct indicators of individual user activity in YouTube, both in terms of production and consumption.

- **Number of Uploads.** The number of videos uploaded by the user is a measure of how much the user contributes to YouTube with content. It is an interesting measure from the point of view of user-generated content production and online video broadcasting.

- **Number of Videos watched.** The number of videos watched is a measure of how much a user participates in the site consuming video content. Watching videos is the simplest, most passive form of participation in YouTube, and it does not require users to be logged-in. Therefore, whereas this feature may be a reliable estimate for some users (those who log-in to interact in the site while watching videos), for others it may vary largely from the actual number of videos watched. Whatever the case, we argue that the feature may tell something about the way users consume video in YouTube.

- **Number of Favorites.** Marking a video as a "favorite" (a.k.a. "fave"-or) is a useful practice for users that helps to maintain a list of videos preferred videos, which thecan be later disseminated in blogs or other websites, or can simply be used to replay and share videos with other people at anytime. It is therefore a descriptor of a specific way of engaging with videos online.

The **social oriented features** are measures of an interaction between users. We differentiate between incoming or outgoing features, depending on the role of the users in the interaction. Incoming features describe how a user is perceived by others, and they are a measure of the social attention achieved by the user.

- **Number of Views.** The accumulated number of views, as it appears in the user channel, is an aggregate of the views over all the videos from the user. The number of views has been typically used as a measure of popularity of videos, as it resembles the way audiences are measures in traditional mainstream media [5]. Here, we use it as a measure of the level of reception of a user's content from other people, which includes registered users as well as not registered. Note however that the number of views does not account for the distribution of views among videos. For example, for some users it could be biased to few highly viewed videos.

- **Number of subscribers.** Subscriptions are a common form of syndication to users and content in social media. In YouTube, users subscribe to other users' channels in order to be notified whenever the users they are subscribed upload new videos to the site. Since it accounts only for registered users, the total number of subscribers of a user is different measure of popularity that complements the number of views.

Alternatively, the outgoing features reveal the level of disposition of a user to proactively interact with other users:

- **Number of subscriptions.** The total number of subscriptions is a measure of the interest of a user to follow other users' content, which clearly denotes social behavior, that goes beyond simply uploading content, being watched, or being followed.
- **Number of friends.** Friendship is an alternative way of connecting with other users in YouTube which does not imply following other users' content. In fact, whereas subscriptions are directed links between users, and therefore do not ensure reciprocity, friendship creates a reciprocal (non-directed) link between two users. As compared to other social media sites, the specific use of friendship connections has not been investigated in YouTube. We hypothesize that for some users it might refer to "true" offline friendships, as opposed to subscription-based connections.

5 Behavioral data analysis

There is a large number of research questions related to the ways in which users participate and interact in YouTube. Here, we focus on issues related to the long-term behavior of users that can be addressed by inspecting the features available from the user channels, that result from the aggregation of the users behavior over their online lifespan. Other questions would likely require more detailed, timestampled data, and cannot be explored here.

5.1 How do features correlate?

We first investigate the interdependence between behavorial features, by computing Pearson's correlation for all the pair-wise combinations, shown in Table 1. Among participative features, the number of uploads and videos watched showed only moderate correlation ($r = .36$). This evidences that uploading and watching videos are two different ways of participation, and that whereas some users contribute to YouTube with content, other users prefer watching videos. Instead, the number of videos watched and times favorited show a larger correlation ($r = .55$), which probably relates to the fact that favoriting a video generally implies having watching it. Overall, we observe larger correlations between social features, as for example, between the number of subscribers and views ($r = .84$), which agrees with the idea that as users have larger pools of subscribers their content is accessed more. In addition, there is a significant correlation across the two groups of features, as for example, between the uploads and the number of views ($r = .67$) and between

Table 1 Correlation between all pairs of features. Feature values were log-scaled. All values are significant with $p < .0001$.

	1	2	3	4	5	6	7
1. uploads		.36	.28	.67	.53	.32	.42
2. watched	.36		.55	.55	.38	.49	.47
3. favorites	.28	.55		.43	.27	.53	.47
4. views	.67	.55	.43		.84	.49	.68
5. subscribers	.53	.38	.27	.84		.44	.68
6. subscriptions	.32	.49	.53	.49	.44		.63
7. friends	.42	.47	.47	.68	.68	.63	

Table 2 Correlation between all pairs of features for the most active users in each feature. The correlation coefficient r_{xy} between feature x (in rows) and y (in columns) is computed after selecting the users contributing to the top ten percentiles of feature x's distribution. $^{*}p < .01$,$^{**}p < .001$,$^{***}p < .0001$.

	1	2	3	4	5	6	7
1. uploads		$.21^{***}$	$.17^{*}$	$.41^{***}$	$.29^{***}$	$.23^{***}$	$.24^{***}$
2. watched	$.12^{***}$		$.36^{***}$	$.14^{***}$	$.15^{***}$	$.32^{***}$	$.26^{***}$
3. favorites	.01	$.21^{*}$		$.02^{*}$	.01	.28	.17
4. views	$.47^{***}$	$.32^{***}$	.25		$.78^{***}$	$.38^{***}$	$.55^{***}$
5. subscribers	$.38^{***}$	$.28^{***}$	$.19^{*}$	$.78^{***}$		$.40^{***}$	$.58^{***}$
6. subscriptions	$.03^{*}$	$.27^{***}$	$.31^{***}$	$.05^{***}$	$.03^{***}$		$.32^{***}$
7. friends	$.17^{***}$	$.29^{***}$	.27	$.32^{***}$	$.33^{***}$	$.49^{***}$	

uploads and subscriptions ($r = .53$). This indicates a certain correlation between the level of attention of users and their content contribution to the site.

Despite what may be suggested by the above results, we hypothesized that the behavioral patterns of users may vary a lot depending on how active users are. For example, we were interested on exploring whether the same correlation levels hold whenever users "specialize" (i.e. they are more active) in one of the features. With this in mind, we recomputed the pair-wise correlations for users contributing to the top ten percentiles of each feature distribution and show them in Table 2. Note that in this case, the correlation matrix is not symmetric and should be read row-wise. Interestingly, we observe a general decrease of the strength of the interdependence, which is significant for some of the pairs. As an example, top uploaders (first row) show lower correlation between uploads and views ($r = .21$) and between uploads and subscribers ($r = .29$) than the ones showed in Table 1 ($r = .36$ and $r = .53$, respectively). This suggests that, compared to the rest of users, top uploaders may have other motifs to contribute to YouTube rather than only social attention. Similarly, the number of favorites and the number of videos watched ($r = .21$) show low correlation for those who favorited the most (third row), compared to the whole sample ($r = .55$). Finally, those who have a lot of subscriptions (fifth row), show

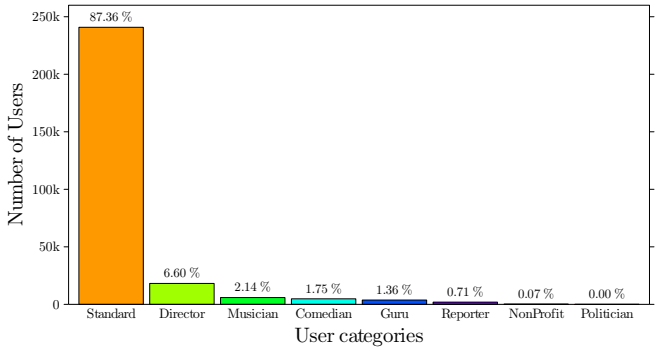

Fig. 3 Distribution of user categories in our dataset. 12.6% of the users self-assigned themselves a special user category (*Director*, *Musician*, *Comedian*, *Guru*, *Reporter*, *Non-profit*, and *Politician*).

low correlation between the number of subscriptions and videos watched ($r = .27$ compared to $r = .49$), and no correlation between subscriptions and number of subscribers ($r = .03$ compared to $r = .44$). The later emphasizes the anti-reciprocity of the subscription links, indicating that active subscribers do not necessarily receive subscriptions from other users, which rather depends on how much they contribute in terms of content. Furthemore, we observe the strength of other relationships, such as the one between views and subscribers ($r = .78$), which continues to hold (and is symmetric) for both the top users receiving views and the top users on the number of subscribers.

This analysis emphasizes that (1) the relationships between pairs of behavioral features are more complex than the linear relation, and (2) these relationships vary between different subsamples of users, depending on their behavior (i.e., how active they are).

5.2 Are you special?

We were interested in exploring the level of popularity that user categories have among the YouTube community. The distribution of user categories is shown in Figure 3. This distribution reveals that 12.6% of the users choose to label themselves with a category different than *Standard*. The relatively moderate level of popularity of special categories could be explained by the poor advertising of such feature in the site. We hypothesize that users are more likely to find about categories by "word of mouth", after seeing other users with labels such as *Director* or *Comedian*. Among special user categories, Directors (52.4%), Musicians (16.8%), Comedians (13.8%), and Gurus (10.8%) are the most popular, followed by Reporters (5.6%), Non-Profit organizations (0.5%) and Politicians (0.0001%). This distribution seems to be biased by the chronological order in which categories were introduced, which could

Table 3 Mean, median and coefficient of variation (cov) values of participatory feature distributions for each user category. First row ("All") groups the totality of users in the dataset, indistinctly of their user category.

Category	uploads			watched			favorites		
	mean	median	cov	mean	median	cov	mean	median	cov
All	31.10	7.00	6.35	2493.42	841.00	2.02	35.99	2.00	2.67
Standard	23.77	6.00	5.13	2054.12	699.00	1.96	28.21	1.00	2.92
Special	81.75	21.00	5.51	5403.49	2803.50	1.62	89.80	22.00	1.68
- Director	95.03	25.00	4.33	5975.35	3301.00	1.61	101.72	28.00	1.58
- Musician	41.84	14.00	3.05	4123.72	2054.50	1.73	68.49	15.00	1.84
- Comedian	40.60	15.00	3.82	5114.65	2580.50	1.56	91.57	23.00	1.68
- Guru	94.85	23.00	8.70	5746.53	3163.00	1.42	85.38	22.00	1.70
- Non-Profit	64.67	26.00	2.47	1447.12	507.00	2.09	15.73	3.00	3.17
- Reporter	127.69	30.00	3.89	4048.75	1415.00	1.77	58.17	8.00	2.10
- Politician	33.57	23.00	1.39	1087.57	591.00	1.42	62.57	6.00	2.43

suggest, in fact, that the process of choosing a category is influenced by a "rich-get-richer" effect, in which new users would tend to choose the most numerous category. Unfortunately, our data does not allow to check this hypothesis. Alternatively, it could also be that *Directors* is perceived as a more broad category than *Musicians* or *Comedians*, and thus users self-associate to it more easily. Whatever the case, this may indicate that the *Director* category is bringing together a larger variety of different users and behaviors, compared to *Comedians*, *Musicians*, or *Gurus*.

5.3 How participative are you?

Based on our sample, users in YouTube uploaded a median of 7 videos, watched 841 videos, and favorited 2 videos. Despite their variance, these figures indicate a relatively low level of individual participation in YouTube compared to other social media sites, as it was exposed in earlier work [11]. However, our category-based analysis reveals that this result is biased by the very low participation of *Standard* users, compared to special users. As shown in Table 3, which gathers some basic statistics of the participatory features for the different user categories, *Standard* users uploaded a median of 6 videos, watched 699 videos, and "favorited" only 1 video, versus the 21 uploads, 2803 videos watched and 22 favorites of special users. In addition, we note that some differences between these statistics for the different user categories.

We propose a methodology to further investigate the differences among user categories at different degrees of activity, going from the most passive to the most active users for each of the features. For this analysis, we first consider a scale of ten different activity levels $l(i)$, $i = 1 \ldots 10$, which are determined for each behavioral feature based on the complete feature distribution deciles, independent from

the user categories. Secondly, we define the discrete distribution $p(u,i)$ as the relative frequency of the user category u on the activity level i, which we computed for each of the user categories in the dataset. In the third place, we build a null model by shuffling the user categories among all the feature values, i.e., we destroy the relational links between features and user categories, keeping the feature values and the proportion of categories. The goal of this null model is to determine the distribution $q(u,i)$ of expected relative frequencies of the user category u given a uniform distribution of the users from this category across the range of feature values. Finally, we can compute the ratio $r(u,i)$ between both distributions:

$$r(u,i) = \frac{p(u,i)}{q(u,i)}, \tag{1}$$

which measures to what extent the feature distribution of a given user category departs from a uniform distribution among all the activity levels. Note that $r(u,i)$ is not a distribution itself, but a measure that indicates how much the category u is underrepresented ($r(u,i) < 1$) or overrepresented ($r(u,i) > 1$) in the i-th bin, as compared to what would result from a uniform distribution. Note also that the accuracy of $r(u,i)$ may vary when used to analyze different user categories, given the differences between sample sizes for the user categories in our dataset. Thus, by repeating the sample procedure several times, which implies a different assignment of user categories to the levels of activity, we are able to build confidence intervals around a mean ratio on $\hat{r}(u,i)$. These confidence intervals help to assess wheather differences between ratios computed for samples of different sizes are significant.

Figure 4 shows the $\hat{r}(u,i)$ values and confidence intervals (equal to one standard deviation) for the three participative features and different user categories. Note that the confidence intervals are larger for categories such as *Non-Profit*, due to the scarcity of data (we do not show *Politicians* for this specific reason). We shall remark two main observations form simple inspection of the figures. First, the feature distributions show clear differences among several user categories. These differences are accentuated when we look at more active levels (see ninth and tenth deciles). Second, we observe a clear difference between *Standard* users and *Special* users behavior. Whereas *Standards* are distributed uniformly along the ranking, *Special* users' behavior is inclined towards higher participation, or in other words, active users (who clearly find valuable their participation in YouTube) also choose to belong to a specific user category. This trend is emphasized as we get closer to the top of the rankings.

We now discuss specific differences for each behavioral feature.

Directors and Gurus consistently upload more videos

In terms of uploads, we observe that very few special users appear among the low participation users. In particular, for the the first decile we obtain $\hat{r}(Directors,1) = 0.20$, $\hat{r}(Gurus,1) = 0.30$, and $\hat{r}(Musicians,1) = 0.31$, and $\hat{r}(Comedians,1) = 0.30$, compared to $\hat{r}(Standards,1) = 1.00$. Special users show ratios larger than one start-

ing from the seventh decile, which clearly indicates that they tend to upload more videos than *Standards*. Focusing on the top percentile we find $\hat{r}(Reporters, 10) = 3.59$, $\hat{r}(Directors, 10) = 2.97$, $\hat{r}(Gurus, 10) = 2.96$, and $\hat{r}(Musicians, 10) = 1.54$, and $\hat{r}(Comedians, 10) = 1.47$. In contrast, $\hat{r}(Standards, 10) = 0.77$. We observe that taken in pairs, *Director* and *Guru*, and *Comedian* and *Musician* are similarly distributed not only on the tenth decile, but across all active levels. We conducted two-sample Kolmogrov-Smirnov (KS) tests[1] for the pair-wise combinations of the original user categories' distributions to assess wheather these similarities are indeed significant. Tests reported p-values larger than 0.1 for these two pairs of user categories, indicating no significant difference on the distributions of uploads, and reported values smaller than 0.001 for the rest of pairs. In addition, right-sided KS tests reported the number of uploads for *Directors* and *Gurus* to be significantly larger than the rest. We argue that the patterns of uploading of different categories are influenced by the time required to create their videos, which in turn may unveil that their videos serve different purposes.

Comedians, Musicians, and Reporters significantly watch less videos

In terms of videos watched, the ratios of special categories for lower active levels do not differ much from those of uploads. In particular, for the the first decile, the ratio $\hat{r}(Directors, 1) = 0.20$, $\hat{r}(Gurus, 1) = 0.22$, and $\hat{r}(Musicians, 1) = 0.31$, $\hat{r}(Comedians, 1) = 0.22$, compared to the ratio of Standards $\hat{r}(Standards, 1) = 1.11$. Focusing on the most active levels, Directors and Gurus again show larger ratios than other categories with $\hat{r}(Directors, 10) = 3.02$ and $\hat{r}(Gurus, 10) = 2.94$, which denote a higher interest of these users for consuming content on YouTube. As with uploads, the no significant difference was found between these two categories across all levels of activity, as reported by KS tests p-values being larger than 0.1. *Comedians* ($\hat{r}(Comedians, 10) = 2.5$), *Musicians* ($\hat{r}(Musicians, 10) = 1.9$) and *Reporters* ($\hat{r}(Comedians, 10) = 1.7$), in this order, are the next on the ranking of videos watched, whereas *Standards* ($\hat{r}(Standards, 10) = 0.75$) remain the last. We argue that compared to *Directors* and *Gurus*, the other special users may be interested in releasing their work or spreading their messages, rather than on exploring other YouTube content.

Directors, Comedians, and Gurus favorite the most

The distribution of the number of videos favorited per user, with up to 45% of the users with zero favorites, imposes a null $p(u,i)$ value for the first four deciles

[1] The two-sample KS test is a non-parametric method which is sensitive to differences in both location and shape of the empirical cumulative distribution functions (CDFs) of two samples, and makes no assumption about the distribution of data. The null hypothesis of this statistic is that the samples are drawn from the same distribution. Thus, a KS test that yields a p-value less than a specified α, leads to the rejection of the null hypothesis, and favors the hypothesis that distributions are different [8].

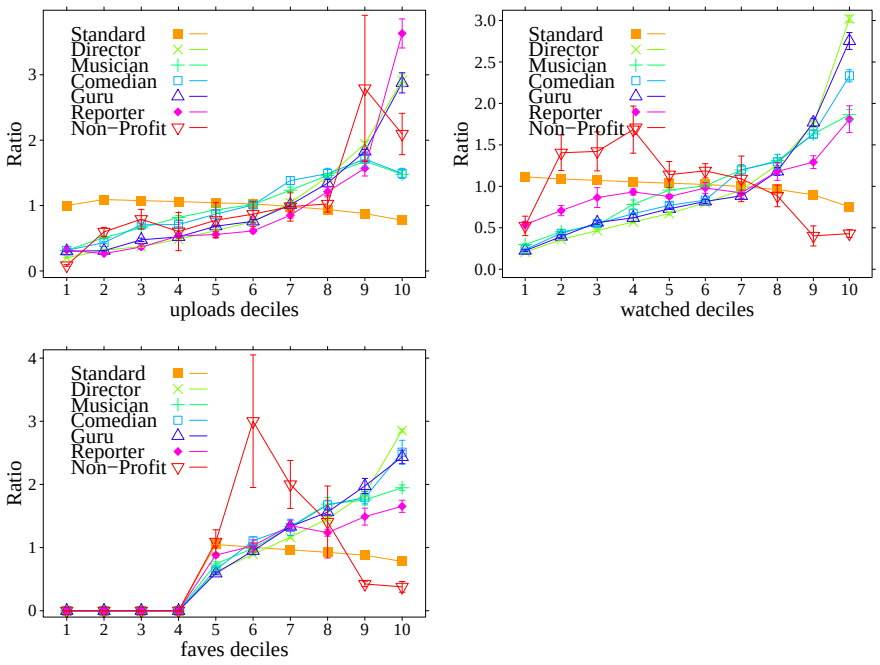

Fig. 4 Mean $\hat{r}(u,i)$ values and confidence intervals of $r(u,i)$ ratio for participative features (confidence values equal to one standard deviation). For each feature, the ten activity levels $l(i)$, $i = 1 \ldots 10$ are fixed by the distribution deciles (i.e. the 1-st and 10-th deciles correspond to the bottom-ten and top-ten feature distribution values respectively).

which results in empty bins for the computation of $\hat{r}(u,i), i = 1 \ldots 4$ (see Figure 4), and all these users are concentrated in the fifth decile. For the fifth decile, we find $\hat{r}(Directors,5) = 0.61$, $\hat{r}(Gurus,5) = 0.58$, $\hat{r}(Musicians,5) = 0.74$, and $\hat{r}(Comedians,5) = 0.65$, which indicate a closer trend to the uniform distribution, compared to uploads and videos watched. Among the most active users in terms of videos favorited, we find *Directors* ($\hat{r}(Directors,10) = 2.80$) , *Comedians* ($\hat{r}(Comedian,10) = 2.49$), and *Gurus* ($\hat{r}(Guru,10) = 2.41$). Contrary to what may be suggested from their $\hat{r}(u,10)$ values, no significant differences were found between Comedian and Gurus across all levels of activity, with two-sided KS tests p-value being larger than 0.1. In addition, one-sided KS tests suggests that the number of favorites is larger for *Directors* than for the rest of categories.

Suming up, we find that *Directors* and *Gurus* are among the most participative in all the aspects, followed by *Comedians* and *Musicians*. Instead, *Non-profit*,and *Reporters* follow a different pattern of participation. Whereas they are also very active uploaders, they typically display lower numbers of videos watched and "favorited", which may indicate that they are more interested in releasing their work or spreading their messages, rather than exploring the site's content.

Table 4 Mean, median and coefficient of variation (cov) values of social feature distributions for each user category. First row ("All") groups the totality of users in the dataset, indistinctly of their user category.

Category	views			subscribers			subscriptions			friends		
	mean	median	cov	mean	median	cov	mean	median	cov	mean	median	cov
All	3114.52	121.00	24.30	89.84	5.00	21.40	14.30	0.00	7.62	21.55	0.00	14.74
Standard	1390.23	89.00	21.59	31.69	4.00	9.49	8.30	0.00	7.18	9.27	0.00	17.71
Special	14158.82	958.00	13.50	300.27	19.00	13.61	55.83	11.00	4.65	106.46	13.00	7.30
- Director	17929.58	1212.00	14.00	320.36	22.00	15.70	59.43	13.00	4.79	104.52	16.00	7.20
- Musician	4179.42	489.00	7.01	114.39	9.00	8.40	33.12	5.00	3.93	88.71	8.00	10.67
- Comedian	9568.23	506.50	11.50	260.06	10.00	14.20	52.32	12.00	5.16	109.52	11.00	8.23
- Guru	17085.11	1250.50	6.86	459.91	33.00	6.58	72.78	18.00	3.66	128.78	18.00	4.40
- Non-Profit	10614.01	1639.50	3.75	397.34	30.00	4.67	29.36	1.00	3.91	81.37	4.00	4.02
- Reporter	11102.75	1016.50	5.43	284.35	22.00	6.13	58.50	6.00	4.82	120.08	8.00	5.39
- Politician	3737.14	2241.00	1.30	52.14	16.00	1.63	44.86	12.00	1.62	16.43	6.00	1.55

5.4 How social are you?

Social-oriented features follow a similar pattern than participative features regarding the differences between *Standard* users and special users. As shown in Table 4 (right), *Standards* accumulated, in median values, 89 views and 4 subscribers, versus the 958 views and 19 subscribers of special users. *Standards* have also no subscriptions or friends in median value. This likely has to do with the motivation behind YouTube users. *Standard* users, in general, might be more interested in sharing few casual videos with their relatives or friends, which would potentially generate a small number of views. Instead, special users seem to be more interested in interacting with the YouTube community at large through videos that are of wider interest among other users, thus receiving more views and subscriptions.

Incoming social features

Figure 5 (top) shows the ratio values for different activity levels of incoming social features. Compared to participatory features, the ratios of special users appearing among the lower activity levels in terms of views are lower, which emphasizes the differences between special users and *Standard* users. We find $\hat{r}(Directors,1) = 0.09$, $\hat{r}(Gurus,1) = 0.13$, and $\hat{r}(Musicians,1) = 0.19$, and $\hat{r}(Comedians,1) = 0.15$, compared to the $\hat{r}(Standard,1) = 1.13$. Clearly, *Gurus*, *Directors*, and *Reporters*, are the ones who get more attention from the community in terms of accumulated views, with $\hat{r}(Gurus,10) = 3.79$, $\hat{r}(Directors,10) = 3.63$, and $\hat{r}(Reporter,10) = 3.46$ respectively. Compared to them, the ratio for the most active *Comedians* and *Musicians* is roughly cut in half, with $\hat{r}(Comedians,10) = 2.06$ and $\hat{r}(Musicians,10) = 1.97$. We performed KS tests for pair-wise combinations of these distributions. Two-sided KS tests reported no significant differences betwen Comedians and Musicians across all the levels of activity. In addition, one-sided KS tests indicate that the dif-

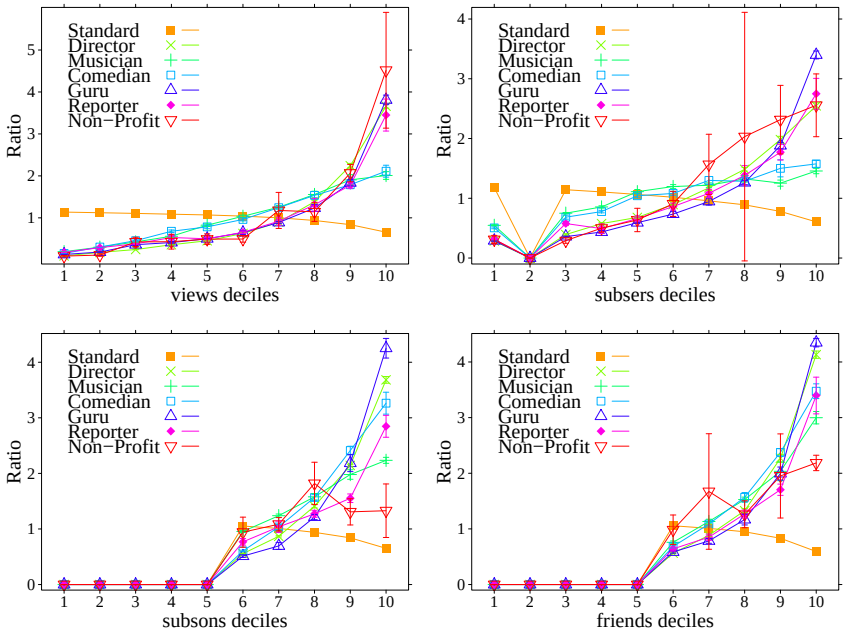

Fig. 5 Mean $\hat{r}(u,i)$ values and confidence intervals of $r(u,i)$ ratio for social features (confidence values equal to one standard deviation). For each feature, the ten activity levels $l(i)$, $i = 1 \ldots 10$ are fixed by the distribution deciles (i.e. the 1-st and 10-th deciles correspond to the bottom-ten and top-ten feature distribution values respectively).

ferences between the Comedians and Musicians and the rest of the special users are consistent, and that they are significantly receiving less views.

In terms of subscriptions, we find similar patterns on the differences between user categories. Gurus, Reporters, and Directors are the ones concentrating more subscribers, with $\hat{r}(Gurus, 10) = 3.42$, $\hat{r}(Reporters, 10) = 2.71$, and $\hat{r}(Directors, 10) = 2.55$, respectively; followed by Comedians and Musicians, with $\hat{r}(Comedians, 10) = 1.65$ and $\hat{r}(Musicians, 10) = 1.41$. This time, two-sided KS tests reported no significant differences between Directors and Reporters. In addition, one-sided KS tests suggest that differences between Gurus, Reporters, and Directors are significant.

Outgoing social features

Figure 5 (bottom) shows the ratio values for different activity levels of outgoing social features. The distributions for subscriptions and friends show that a large percentage of users, both Standard and Special, do not actually use these features, as it is shown by the number of empty deciles. However, among those that use them, there are clear differences between *Special* and *Standard* uses. For the most active level, for example, we find *Gurus*, *Directors*, and *Comedians* as the top cate-

Table 5 Mean, median and coefficient of variation (cov) values of participatory feature distributions for each gender.

Category	uploads			watched			favorites		
	mean	median	cov	mean	median	cov	mean	median	cov
Males	29.56	7.00	6.30	2582.60	884.00	2.03	32.78	1.00	2.78
Females	24.98	7.00	5.43	2115.06	684.00	2.00	43.16	2.00	2.45

Table 6 Mean, median and coefficient of variation (cov) values of social feature distributions for each gender.

Category	views			subscribers			subscriptions			friends		
	mean	median	cov	mean	median	cov	mean	median	cov	mean	median	cov
Males	2394.78	117.00	22.37	70.60	5.00	19.64	13.76	0.00	8.54	18.57	0.00	14.97
Females	2413.76	105.00	20.41	64.74	4.00	11.82	14.43	1.00	5.05	22.49	1.00	6.09

gories with $\hat{r}(Gurus, 10) = 4.25$, $\hat{r}(Directors, 10) = 3.64$, $\hat{r}(Comedians, 10) = 3.19$. Compared to them, Reporters and Musicians seem to be less prominent in terms of subscriptions, with $\hat{r}(Reporters, 10) = 2.86$, $\hat{r}(Musicians, 10) = 2.26$. This is accentuated for Standards $\hat{r}(Standards, 10) = 0.65$. Interestingly, the same ordering holds among the users having more friends, which suggests that YouTube-specific contacts and "real life" contacts might have a similar presence in their online interaction.

5.5 Male or female?

Gender analysis, and in particular, gender distribution, uncovers interesting aspects of large-scale behavior in YouTube. Based on our data, YouTube concentrates a higher participation number of men (73% of the users) than women (27%). Moreover, we find that male users are more likely to enroll in special user categories, with a proportion of 13% compared to the 9% of females. A two-proportion z-test[2] indicate that this difference is significant with $p < 10^{-3}$.

Gender differences in the distribution of special categories are very small. However, the distribution of both participative and social-oriented features show very different patterns of behavior between men and women. As shown by their median values in Table 5, whereas men and women upload the same number of videos (a median of 7 videos), men tend to watch more videos than women (884 and 684 videos watched respectively). Instead, special female users "favorited" a median of

[2] The two-proportion z-test is used to compare proportions of two independent binomial samples. The null hypothesis of this statistic is that the two proportions are equal. Thus, a two-proportion z-test giving a p-value less than a specific α (typically .05), leads to the rejection of the null hypothesis, and indicates that the proportions are different [20].

2 videos, compared to the 1 favorite of men, which suggests a different pattern on the way women watch and engage with video content. One-sided KS-tests among the corresponding features distributions indicate that all these differences are significant with $p < 10^{-3}$.

Regarding social features, we find men to receive more attention than women. Men accumulated 117 views and 5 subscribers compared to the 105 views and 4 subscribers of women. These differences are significant as reported by one-sided KS-tests ($p < 10^{-3}$). However, women accumulate more subscriptions, and more friends (a median of 1 subscriptions compared to 0 of men), which suggests that women, overall, they have a more social-driven behavior in YouTube than men. These differences were also significant, as reported by one-sided KS-tests ($p < 10^{-3}$).

Some of these findings are not completely new but are backed up by substantially more data. In a manual analysis of a small random sample of 100 YouTube vlogs, Molyneaux et al. [19] found a higher presence of male users. They also found that women were most likely to interact with the YouTube community through their videos, and that they receive a higher number of views than men. These results we present here use three orders of magnitude more data.

6 Classifying YouTube users

The analysis of the previous section suggests that the basic features the YouTube users' channel could be used to characterize user categories. In order to explore the goodness of this characterization, we defined a series of classification tasks where a YouTube user is classified between two given user categories using different combinations of features. We use the results not only to evaluate the discriminative power of the features, but to measure the similarity between the behavior of different types of users.

For each task we performed a 10-fold cross-validation using a Support Vector Machine (SVM) classifier with a Gaussian Kernel. In each case, we optimized the Kernel parameters (σ and C) using 5-fold cross-validation on the training data.

Our first binary classification task was between *Standard* and non-*Standard* users (as one unique category) on a balanced subset of 10,000 users (5,000 per category) randomly selected from our dataset. For special users, we only considered the most popular special categories: *Directors*, *Musicians*, *Comedians* and *Gurus*, which were also balanced among the 5,000 corresponding samples. Results in Figure 6 show an average classification accuracy rate (CAR) of $68.9\% \pm 1.2$ for the fusion of features, which performs better than the 50% CAR corresponding to a random decision. However, we found that using the number of friends as a single feature one could achieve a classification rate of $68.3\% \pm 1.4$. This indicates that differences in user behavior between *Standard* users and special users are to be found mainly in the level and pattern of interaction with other users, that is, in how social they are. This observation would also explain why subscriptions and subscribers are the next best single features.

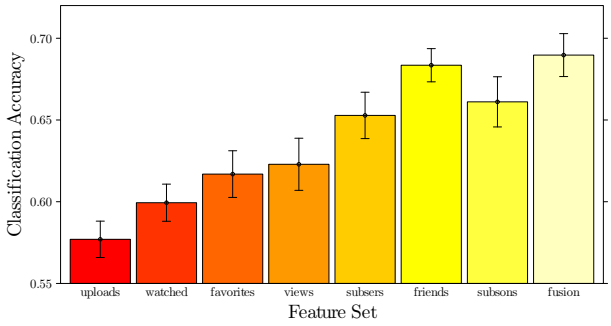

Fig. 6 *Standard* vs. non-*Standard* CAR using behavioral features alone and in combination. Best CARs are obtained for the fusion of features, as well as for "friends" alone.

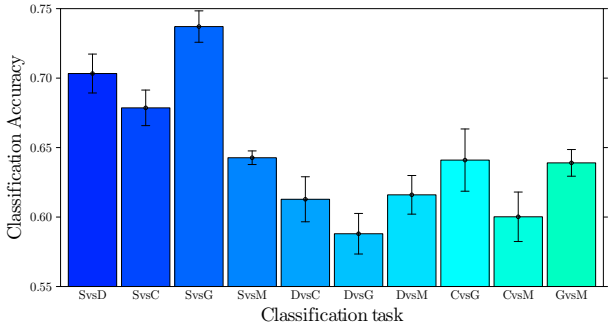

Fig. 7 Best CARs on binary tasks. Capital letters in horizontal axis correspond to the initials of the categories' names (i.e $S = Standard$, $D = Director$, $M = Musician$, $C = Comedian$, and $G = Guru$). Except for those tasks involving Standard users, CARs were obtained with a fusion of at least three social features.

The rest of the binary tasks were defined between pairs of single user categories (see Figure 7), using up to 5,000 users per category. Data used on tasks involving categories with less than 5,000 users (e.g. *Gurus*) were balanced to the minimum among the number of users of such categories. Except for those tasks involving *Standards*, the best CARs were obtained with several combinations (of at least three) social-oriented features, which indicates that higher similarity of special categories requires more complex descriptors of behavior. The use of single features results in a drop of the best CAR to 53%. For tasks involving *Directors* and *Gurus*, the number of views appeared in almost all winning combinations, whereas for *Comedians* and *Musicians* different combinations led to similar results. In some cases, replacing only one of the social features for a participative feature would not cause a drop in performance, suggesting that the latter features still contain information useful to discriminate between special categories.

As suggested in previous sections, Directors and Gurus, and Comedians and Musicians are similar categories on what concerns to several behavioral aspects, which

results in the lowest CARs, with $58.8\% \pm 1.4$ and $60.0\% \pm 1.7$, respectively. In contrast, Comedians and Musicians, and Gurus and Musicians are the pair-wise combinations with best CARs among special users, with $64.1\% \pm 2.2$ and $63.3\% \pm 0.9$, respectively.

7 Conclusions

In this paper, we have presented an analysis of YouTube users' long-term behavior using easy-to-extract features from the users' channels and large-scale data. We have shown that the group of special YouTube users is an undeniably active social community, as opposed to *Standard* users, and we have revealed different patterns of participation and interaction and highlighted some of the possible motivations behind them. Furthermore, we have backed up earlier findings on gender behavioral division using (by three orders of magnitude) more data. A series of binary classification tasks between YouTube users categories has shown social-oriented features to be key when describing differences between user groups' behaviors.

Whereas the features obtained from the user channels are capturing broad statistics on different aspects of long-term user behavior, finer aspects could be obtained by extracting other features in a more computationally expensive manner, which could hopefully lead to a better characterization of users. As an example, new participative features could include the frequency of the uploads, the number of comments posted, or user-specific videos features, including metadata and audiovisual features. Social-oriented features such as friends, subscriptions, or subscribers could be further divided in inter-category and intra-category relations. In addition, considering the antiquity of YouTube users (i.e. the age of their channels) could help to clarify the similarities between users of the same behavioral group. Furthemore, users could be also categorized based on the content they upload in YouTube. Video content is probably one of the most reliable indicators of the interest of users, which drive the way they use and behave in social video-sharing sites. The use of more computationally expensive features to capture finer aspects of the users' behavior will be the subject of our future work.

Acknowledgements We thank the support provided by the Swiss National Science Foundation (SNSF) through the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2.

References

1. Website monitoring blog. YouTube facts & figures (history & statistics). http://www.website-monitoring.com/blog/2010/05/17/youtube-facts-and-figures-history-statistics/
2. Youtube fact sheet. http://www.youtube.com/t/fact_sheet. (accessed November 2010)

3. Benevenuto, F., Duarte, F., Rodrigues, T., Almeida, V.A., Almeida, J.M., Ross, K.W.: Understanding video interactions in youtube. In: MM '08: Proceeding of the 16th ACM international conference on Multimedia, pp. 761–764. ACM, New York, NY, USA (2008)
4. Biel, J.I., Gatica-Perez, D.: Wearing a youtube hat: directors, comedians, gurus, and user aggregated behavior. In: MM '09: Proceedings of the seventeen ACM international conference on Multimedia, pp. 833–836 (2009)
5. Burgess, J., Green, J.: YouTube: online video and participatory culture. Polity, Cambridge, UK (2009)
6. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 1–14. ACM, New York, NY, USA (2007)
7. Cheng, X., Dale, C., Liu, J.: Statistics and social network of youtube videos. In: Quality of Service, 2008. IWQoS 2008. 16th International Workshop on, pp. 229–238 (2008)
8. Conover, W.J.: Practical Nonparametric Statistics. John Wiley & Sons, New York (1971)
9. Gill, P., Arlitt, M., Li, Z., Mahanti, A.: Youtube traffic characterization: a view from the edge. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 15–28 (2007)
10. Griffith, M.: Looking for you: An analysis of video blogs. In: Annual meeting of the Association for Education in Journalism and Mass Communication (2007)
11. Halvey, M., Keane, M.: Exploring Social Dynamics in Online Media Sharing. In: Proc. of the 16th International Conference on World Wide Web, pp. 1273–1274 (2007)
12. Kruitbosch, G., Nack, F.: Broadcast Yourself on YouTube - Really? In: Proceedings of the 3rd ACM international workshop on Human-centered computing, pp. 7–10 (2008)
13. Landry, B., Guzdial, M.: Art or circus? characterizing user-created video on YouTube. Tech. rep., Georgia Institute of Technology (2008)
14. Lange, P.: Commenting on comments: investigating responses to antagonism on YouTube. In: Conference on Society for Applied Anthropology (2007)
15. Lange, P.: Publicly private and privately public: social networking on youtube. Journal of Computer-Mediated Communication **1**(13) (2007)
16. Lin, W.H., Hauptmann, A.: Identifying ideological perspectives of web videos using folksonomies. In: AAAI fall symposium on Multimedia Information Extraction (2008)
17. Maia, M., Almeida, J., Almeida, V.: Identifying user behavior in online social networks. In: SocialNets '08: Proceedings of the 1st Workshop on Social Network Systems, pp. 1–6. ACM, New York, NY, USA (2008)
18. Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and Analysis of Online Social Networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 29–42 (2007)
19. Molyneaux, H., O'Donnell, S., Gibson, K., Singer, J.: Exploring the gender divide on youtube: An analysis of the creation and reception of vlogs. American Communication Journal **10**(2) (2008)
20. Newcombe, R.G.: Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in medicine **8**(17), 857–872 (1998)
21. O'Donnell, S., Gibson, K., Milliken, M. Singer, J.: Reacting to YouTube Videos: Exploring Differences Among User Groups. In: Proceedings of the International Communication Association Annual Conference, pp. 22–26 (2008)
22. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Commun. ACM **53**(8), 80–88 (2010)
23. Zink, M., Suh, K., Gu, Y., Kurose, J.: Watch global, cache local: YouTube network traffic at a campus network - Measurements and implications. In: MMCN '08: Proceedings of SPIE/ACM conference on Multimedia Computing and Networking (2008)