J Braz Comput Soc (2010) 16: 201–214
DOI 10.1007/s13173-010-0019-x

ORIGINAL PAPER

Equal but different: a contextual analysis of duplicated videos on YouTube

Tiago Rodrigues · Fabrício Benevenuto · Virgílio Almeida · Jussara Almeida · Marcos Gonçalves

Received: 8 March 2010 / Accepted: 23 July 2010 / Published online: 19 August 2010

Abstract Videos have become a predominant part of users' daily lives on the Web, especially with the emergence of online video sharing systems such as YouTube. Since users can independently share videos in these systems, some videos can be duplicates (i.e., identical or very similar videos). Despite having the same content, there are some potential context differences in duplicates, for example, in their associated metadata (i.e., tags, title) and their popularity scores (i.e., number of views, comments). Quantifying these differences is important to understand how users associate metadata to videos and to understand possible reasons that influence the popularity of videos, which is crucial for video information retrieval mechanisms, association of advertisements to videos, and performance issues related to the use of caches and content distribution networks (CDNs). This work presents a wide quantitative characterization of the context differences among identical contents. Using a large video sample collected from YouTube, we construct a dataset of duplicates. Our measurement analysis provides several interesting findings that can have implications for how videos should be retrieved in video sharing websites as well as for advertising systems that need to understand the role that users play when they create content in services such as YouTube.

Keywords Video duplicates · Metadata association · Social network · YouTube

T. Rodrigues · F. Benevenuto · V. Almeida (✉) · J. Almeida · M. Gonçalves
Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
e-mail: virgilio@dcc.ufmg.br

T. Rodrigues
e-mail: tiagorm@dcc.ufmg.br

F. Benevenuto
e-mail: fabricio@dcc.ufmg.br

J. Almeida
e-mail: jussara@dcc.ufmg.br

M. Gonçalves
e-mail: mgoncalv@dcc.ufmg.br

1 Introduction

Content is rapidly moving towards more video. The signs are evident. According to Comscore [9], a measurement company, 161 million US Internet users watched online video during the month of August 2009 and viewed more than 25 billion videos. Another Comscore report indicates that the total number of videos viewed online in the UK in April 2009 was 4.7 billion videos [10]. Video search also became a popular service on the Web, and YouTube accounts for a large fraction of all Google search queries in the US, generating 3.5 billion searches in August 2009 [10]. Video is now mainstream and users can find an online video on virtually any topic.

However, the exponential explosion of video data combined with the user participatory model of the Web 2.0 created a large fraction of duplicated or near-duplicated videos on the Web. Duplicated videos bring problems to both users and content producers. Several aspects of the Web are affected by the growing presence of duplicated content, such as copyright enforcement, video clustering, recommendation, annotation propagation, multimedia authoring, and video search. Most video search engines rely exclusively on text keywords or user-supplied tags to select video content. As a result, a typical search on a popular topic often returns many duplicated and near-duplicated videos in the top results. Figure 1 shows three identical results in terms of the

Fig. 1 Illustrative example of a search on YouTube for the words "Susan Boyle"

video content for a search on YouTube using the query "Susan Boyle". We can see that not only do these videos have differences in terms of the metadata associated to them, but they also present different statistics that indicate popularity. In fact, Cha et al. [7] pointed out that duplicated videos can negatively impact the performance of caching mechanisms and content distribution networks (CDNs).

Due to the practical importance of detecting duplicates, the literature has focused on techniques to identify and remove duplicates in video sharing services [15, 29]. Although these efforts are essential to improve the quality of service, they do not explore important aspects of duplicates. It is important to understand what is behind the creation of duplicate content.

Despite having similar content, duplicated videos may exhibit different metadata (e.g., tags and categories) and may have different popularity indicators (e.g., number of views and ratings). The same content can be viewed in completely different ways by different users. Quantifying the different perceptions of the social community is important for two reasons. The first one refers to the need for understanding how users associate metadata to videos on video sharing services, such as YouTube. Such understanding is crucial for video information retrieval mechanisms since the current systems rely mostly on tags and other metadata associated by users. The second reason is associated with understanding possible reasons that influence the popularity of videos, which is important to the association of advertisements to videos and to performance issues related to the use of caches and CDNs. Moreover, it is also important to increase the understanding of factors that contribute to the creation of duplicated videos.

This paper presents a thorough characterization of the differences that exist in duplicated videos. The differences are analyzed from different viewpoints, i.e., from the content creator's viewpoint and from the content consumer's viewpoint. We are particularly interested in answering the following questions: (1) *Do users associate similar metadata to identical videos*? (2) *Do users agree on the topic assigned to identical videos*? (3) *Do users agree on the ratings given to identical videos*? (4) *What is the impact of the owners' characteristics on the popularity of their videos*? (5) *Is there any evidence of opportunistic behavior in the creation of duplicates on YouTube*?

To answer these questions, we crawled and created a large collection of videos considered as duplicates by YouTube. Our measurement analysis provides several interesting findings that can have implications for how duplicated videos should be retrieved in video sharing websites as well as for advertising systems that need to understand the role of duplicated videos in services such as YouTube.

The rest of the paper is organized as follows. Section 2 discusses our main findings. Section 3 discusses related work. Section 4 describes how we collected and created a dataset of duplicates. Section 5 presents a characterization of differences among duplicates. Section 6 investigates how users create duplicates. Lastly, Sect. 7 concludes the paper and discusses some directions towards which this work can evolve.

2 Main findings

In this work we provide insights towards understanding why some videos became more popular than others, comparing the popularity of identical videos. This is the first work we know of that considers this strategy to analyze video popularity in video sharing systems. Another novel contribution of this work is the understanding of how different users perceive similar content, which is important to many applications such as video recommendation and video search. We also analyze factors related to the creation of duplicates, which may help in the development of methods to detect and avoid identical videos in the system. Our main findings are summarized below:

- About 39% of the duplicates were created less than a month after the creation of the oldest content, and 1%

were created more than one year after the first content appeared in the system.
- Most duplicates receive similar user evaluation (e.g., star rating feature). Only a small fraction of the videos are evaluated in a discrepant manner.
- Duplicates have few metadata (tags, description and title) in common, which can be an indicator of the different perceptions that users have of a video content.
- A significant fraction of the duplicates do not share the same user-defined category, falling into different predefined topics on YouTube.
- Duplicate videos exhibit different degrees of popularity. Duplicates created by owners of popular videos tend to be more popular, suggesting that characteristics of the owners may influence the video popularity. For example, users with many friends tend to have duplicates that are more popular.
- In 8% of the groups of duplicates there are videos in which the owners create more than one duplicate. By analyzing these videos, we noted the existence of video response promotion (i.e., a large number of video responses posted to a unique target video in an attempt to promote the target video to lists of most responded videos) as well as videos containing tag spamming.

3 Related work

The existence of duplicates is a problem in other systems such as blogs [1] and photo sharing systems [31]. In [30], the authors propose a mechanism to filter near-duplicates results from the video search. They create a collection of near-duplicates based on 24 search queries to YouTube, Google Video, and Yahoo! Video. Using a hierarchical clustering algorithm, they are able to find 27% of near-duplicates of the most popular video resulting from a search. In [15], the same authors propose to combine contextual information regarding time duration, number of views, and thumbnail images with content analysis derived from color and local features to achieve real-time near-duplicate elimination 164 times faster than the effective hierarchical method proposed in [30], with a slight loss of effectiveness. More recently, Tan et al. [29] proposed a new method to detect near-duplicates which focuses on the scalable detection and localization of partial near-duplicate videos by jointly considering visual similarity and temporal consistency. Huang et al. [17] developed a Web-based integrated platform which performs online detection of near-duplicates over continuous video streams, as well as retrieval of near-duplicate clips from segmented video collections.

A recent study from Cherubini et al. [8] highlights the importance of studying duplicates from a human-centric perspective. They study the different definitions of near-duplicates and show that some near-duplicates that add visual content to the original video are not perceived as near-duplicates by users. In another work [25], the same authors conducted a study with 217 users of video sharing websites and reported that participants had a preference for one video when compared to its duplicates. Additionally, their study revealed that users were more tolerant to changes in the audio than in the video channel. Our work is complementary to these efforts as it has a different focus, which is to characterize the contextual differences of videos with identical contents.

There are several recent efforts that characterize different aspects of video sharing systems, especially YouTube. In particular, [13] presents a characterization of YouTube traffic from the point of view of a university campus, comparing their results with those previously reported for the Web and traditional video servers. Another characterization of YouTube, based on a traffic collected in a university, is presented in [32]. The authors also perform simulations showing that client and proxy caching and P2P distribution can reduce network traffic and improve response time in video sharing systems. Another important effort that characterizes video sharing systems is presented in [7]. The authors analyze the popularity distribution, evolution, and characteristics of YouTube videos. Additionally, they present evidence of the existence of duplicates through a manually built duplicate database. They analyze the popularity of duplicates and discuss the potential problems that duplicates can cause to the system. The properties of a social network created by interactions through video responses on YouTube are analyzed in [3, 6], revealing the existence of malicious users who post video responses to unrelated discussion topics aiming at promoting some content such as advertisers and pornography. Recently, we have approached the problem of identifying these users using a machine learning approach [4, 5]. The present work studies the existence of possible malicious behavior associated with the creation of duplicates.

Lastly, Suchanek et al. [28] quantified two common assumptions about social tagging in Web pages: that tags are "meaningful" and that the tagging process is influenced by tag suggestions. Their analysis was based on a corpus of search keywords, contents, titles, and tags applied to several thousand popular Web pages. Their results showed that the most popular tags of a page tend to be the most "meaningful". They also developed a model to measure the influence of tag suggestions. Another relevant related work, presented in [23], proposes an approach to discover social interests based on user-generated tags. Based on a large analysis of real data extracted from del.icio.us [11], the authors showed that, in general, user-generated tags are consistent with the Web content associated to them. Marshall [24] compared the characteristics of public tags with other forms of descriptive metadata (titles and narrative captions) that users

have assigned to a collection of very similar images gathered from Flickr [12], a popular photo sharing service. Her work showed that narrative metadata may be more effective than tags for capturing certain aspects of images that may influence their subsequent retrieval and use. Our work is complementary to these previous studies since we quantify the similarity of the metadata associated by different users to videos with the same content.

In a previous work [27], we provide a set of analysis about contextual differences of duplicated videos. The present work greatly builds on our first effort not only by providing a much more thorough, richer, and solid investigation of contextual differences of duplicated content, but also by studying the social perception of duplicate content from the viewpoint of the content producer (i.e., creator/owner) and from the consumer's viewpoint (i.e., user).

4 Data collection

In order to study the characteristics of duplicates, we collect data from YouTube, aiming to obtain different groups of identical videos. The strategy used to collect duplicates is based on searching random words on YouTube and collecting videos which appear as search results. When YouTube shows the search results, it currently filters the duplicates out, showing only one video per group of duplicates. However, YouTube used to offer links to these groups. Our crawler followed these links, collecting information on duplicates and their owners. From this point on, we refer to the sets of duplicates grouped by YouTube on the search results as *groups of duplicates*.

Our crawler was built in a distributed fashion (1 server and 10 clients). The server selects random words from an English dictionary obtained from an open source tool called *ispell* [18]. The server sends one word at a time to each client. The clients execute Algorithm 1 in a loop until the server stops sending words. For each video and its duplicates, we collected a number of pieces of information available, including video identifier, video contributor identifier, title, category, description, tags, upload time, video duration, number of ratings, average rating, number of views, number of users who set the video as favorite, number of comments received, number of video responses received, etc. Table 1 shows an example of information collected about a video d uploaded by user u. We also collected characteristics of the other videos of the duplicate owners.

After running for one week, our crawler found more than 100 thousand duplicates, grouped into 9,178 groups. In total, we collected 100,373 videos from 80,297 users. We noted that some groups have videos with different durations (i.e., a full version of a video and another with only some of the scenes are considered duplicates by the YouTube algorithm). Thus, since our goal is to have a dataset of groups of identical content, we filtered videos with duration differing by more than 2% of the mean duration of the group. In total, we filtered 31,709 duplicates, which reduced the number of groups to 7,330 since some groups had all duplicates filtered. Table 2 presents a summary of the data collected and the data filtered. Groups of duplicates are, in general, small. About 51% of the groups have only 2 duplicates and only 13% have more than 10 videos. The largest group, composed of duplicates of a popular video, contains 947 duplicates.

Algorithm 1 Crawler to collect duplicates

```
1: Obtain word W from server
2: Search YouTube using W
3: for each video v that appears on the search results do
4:     if v has a list of duplicated videos DV then
5:         for each video d in DV do
6:             Collect information on d
7:             Collect information on d owner's u
8:             Collect information of all videos uploaded by u
9:         end for
10:    end if
11: end for
```

Table 1 Example of information collected about a video d uploaded by user u

Video d information	
Video identifier	RXPZh4AnWyk
Title	Susan Boyle—Britains Got Talent 2009
Tags	singer, episode, dreamed, talent, 2009
Number of views	37,788,942
Duration	5:50
...	...
User u information	
User name	userchannel
Number of videos	49
Number of subscribers	1,519
Number of friends	823
Country	Scotland
...	...

Table 2 Summary of the duplicate dataset

	Collected	After filtering
Crawling period	05/24/08 – 05/31/08	–
# words searched	319	–
# groups	9,178	7,330
# duplicates	100,373	68,664
# owners	80,297	58,922
# videos collected	1,844,611	1,321,407

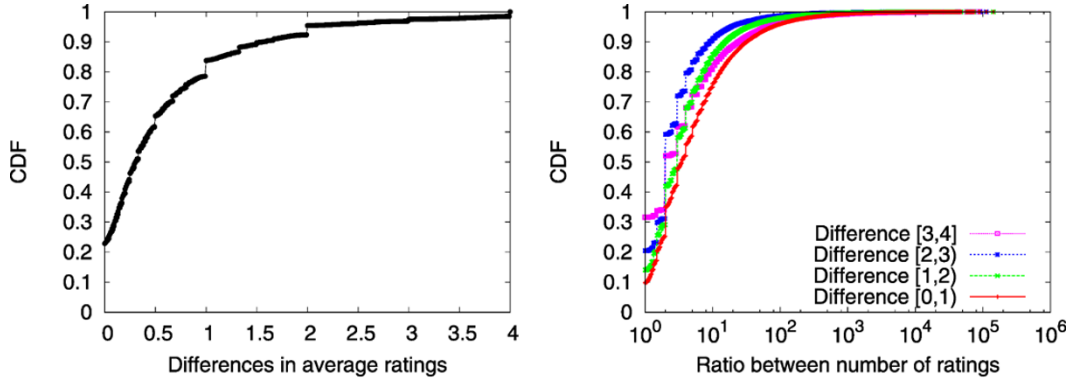


Fig. 2 (Color online) Differences between pairs of video duplicates in terms of ratings

Since our approach to create a dataset of duplicates relies on the YouTube algorithm to filter duplicates, it is possible that there are some videos which are not duplicates in our dataset. In order to verify this, we created a script to randomly selected 154 groups of videos for manual inspection. This sample size was calculated in order to obtain a confidence interval of 7.8% with a 95% confidence level, using the equation:

$$ss = Z^2 * p * (1 - p)/c^2$$

where Z is the Z-value, p is the percentage of picking a choice, and c is the confidence interval. This equation is presented in [19]. In total, we watched 1,059 videos.

The videos were analyzed and flagged as correctly or incorrectly identified as duplicate by YouTube. In order to minimize the impact of human error, three volunteers were used. All groups were analyzed and independently classified by two volunteers. The third volunteer was used as a tie-breaker. We considered the definition of duplicates presented in [30], which says that videos with small differences, such as changes in color, quality, or audio operations and small differences in size, are considered duplicates. The volunteers were instructed to, in case of doubt, flag incorrectly classified videos as duplicates, thus adopting a conservative strategy. With 95% of confidence, only $5 \pm 3.4\%$, were considered incorrectly classified as duplicates by YouTube. Thus, in the following sections, we assume that the percentage of videos classified erroneously as duplicates by YouTube does not have a strong impact on our analysis. All the videos classified erroneously as duplicates that we found were removed from the dataset.

5 Contextual analysis

Duplicates can present differences in several aspects, such as in the indicators of popularity and quality (i.e., number of views, ratings, comments, etc.) and in their metadata information. Additionally, differences in the characteristics of the owners of the duplicates may influence the popularity of these videos. Next we analyze these contextual differences among duplicates, providing insights into their impact on the popularity of videos as well as on how users associate metadata to content. The analysis of contextual differences is based on a quantitative characterization of the collective view of duplicate content. It also gives insight into the social behavior of users that create and consume duplicate content.

5.1 Quality and popularity

In this section, we analyze the differences among duplicates in each group of our dataset with respect to different indicators of content quality and popularity.

In YouTube, registered users may evaluate a video after watching it, giving a rate that varies from 1 (the lowest/worst) to 5 (the highest/best). Thus, we start our analysis by comparing the ratings assigned by users to different duplicates in each group. For each pair of duplicates within a group, we compute the absolute difference between their average ratings. Videos without ratings are not considered in this analysis. Intuitively, we would expect very small discrepancies in these differences, since the videos have similar or identical content. Figure 2(a) shows the cumulative distribution function (CDF) of these differences. We note that 23% of the pairs of ratings do not differ and, in 95% of them, the difference is at most 2. This result shows, as expected, that the majority of the duplicates are evaluated similarly by users. However, there is a small fraction of pairs (about 3%) with differences of ratings larger than 3. One possible explanation for such large differences could be related to the number of people who rated each video. For example, if a duplicate A was evaluated by a single person and another duplicate of the same video, say B, was evaluated by 100 persons, we are comparing the perception or opinion of one unique

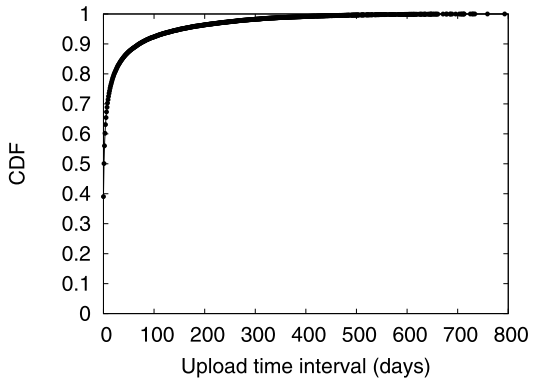


Fig. 3 Post interval between duplicates

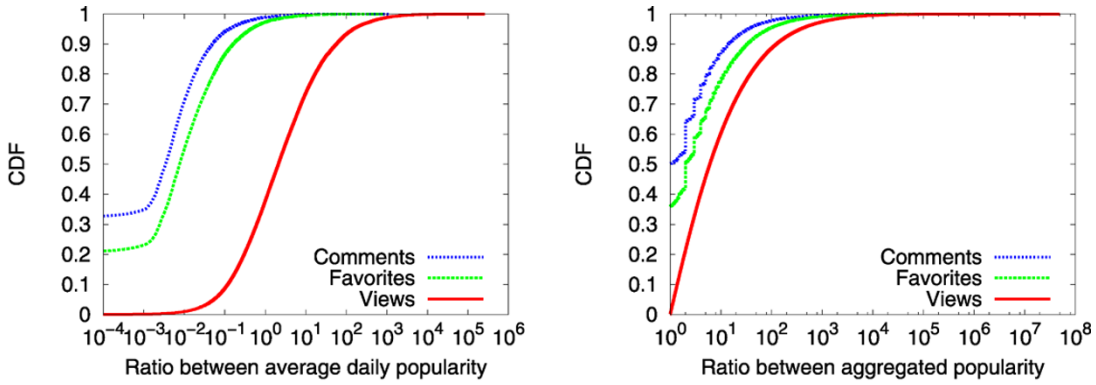
person against the average rating given by 100 people, which may result in a large difference. In order to verify if this is the case, we calculate the ratio between the numbers of users who evaluated each pair of duplicates. Figure 2(b) shows the CDF of these ratios, in which pairs of duplicates are separated in certain fixed ranges based on the corresponding ratings differences. We observe that about 32% of the pairs with differences of ratings larger than 3 have ratio 1, i.e., their videos were evaluated by basically the same number of people. This value is higher than that for pairs whose differences of ratings are smaller than 1, which is about 10%. Thus, we can conclude that such large differences are not caused by large variations in the number of user ratings but rather reflect different user perceptions about the same content. Interestingly, similar results were found in systems of movie recommendation and evaluation [14]. As one might expect, it is quite possible that the same video (at a theater or on YouTube) may receive very different evaluations by different people.

Figure 3 shows the CDF of the time between successive posts of videos with the same content, i.e., the time between successive posts of videos in a group of duplicates. Taking the duplicates within each group ordered by their upload time, we can see that around 39% of the duplicates are posted on the same day of their corresponding predecessor in their group. Moreover, 83% of the duplicates are posted in less than 30 days after the predecessor, while only 1% is uploaded more than 1 year after it. This graph shows that the majority of the duplicates appear in the system within a short period of time after the last post of the content, although, in some cases, a long time passes until the content reappears in the system. Thus, a mechanism that compares videos to detect duplicates could compare an uploaded video with only recent videos and still be able to detect most of the duplicates. We also found that approximately 11% of the videos were uploaded on the same day as the original video or within a week, and about 21% were uploaded within a month.

Now we turn to the analysis of potential differences among duplicates in terms of some indicators of content popularity that may be influenced by the age of a video in the system. These indicators are the number of comments received, the number of times the video was added as a favorite, and the number of views. In order to minimize temporal factors, we consider in Fig. 4(a) the difference between the ratios of each popularity indicator divided by the age (in days) of the duplicate, i.e., the ratio between average daily popularity estimated by each popularity indicator. We calculate the age of a duplicate as the difference between the day it was crawled and the day it was uploaded. As an example, if a video a is on the system for 10 days and received 1200 views in this period, its ratio is 120 views per day. In the same way, if a video b is on the system for 5 days and received 1000 views in this period, its ratio is 200 views per day. The difference of ratios between videos a and b is 80 views per day. Figure 4(a) shows that 71% of the pairs of duplicates differ by at most 0.1 in the number of comments received per day, while 32% of them received the same number of comments per day (i.e., no difference). Moreover, around 1% of the pairs differ by more than 1 comment received per day. The largest observed difference was around 1,094 comments per day. The figure also shows that, in terms of differences in the number of times added as favorite per day, 87% of the pairs differ by less than 0.1 (21% do not differ at all), 3% differ by more than 1, and the largest observed difference was around 798. Finally, more than 62% of the pairs of compared duplicates differ in the number of daily views by more than 1, 50% differ by more than 2, and the largest observed difference was almost 261,667 views per day.

We note, however, that two duplicates with different daily ratios may still have the same aggregated values for a determined indicator, say, the number of views. As an example, if a 5-day old video has a ratio of 2 views per day, it has, in total, 10 views. Similarly, a 10-day old video with a ratio of 1 view per day has also received 10 views since its upload. Among other typical uses, online video sharing systems such as YouTube use aggregated values of indicators to rank videos in lists of top videos and as an option to sort search results, so it is also interesting to analyze the differences of duplicates in terms of their aggregated values for each content popularity indicator. However, comparing aggregated values using the absolute difference between them may not be very significant. For example, a difference of 1 view between duplicates A and B may reflect quite different scenarios depending on whether their corresponding numbers of views are 1 and 2, or 999 and 1000.

In order to verify if there exists a significant difference in the indicators of (aggregate) content popularity among duplicates within a group, we calculated, for each pair of duplicates and each indicator of content popularity, the ratio of



(a) Absolute difference between average daily popularity (b) Ratio between aggregate popularity

Fig. 4 Differences between pairs of duplicates: comments, favorites, and views

the highest aggregated indicator value to the lowest one. If both indicators were 0, we considered the ratio as 1, and if only one of them was equal to 0, we considered the value of the nonzero indicator as the ratio. This ratio shows the number of times a duplicate is more popular, in aggregate terms, than another, considering each of the popularity indicators previously studied: number of comments received, number of times added as a favorite, and number of views. Figure 4(b) shows the CDF of the ratio for each popularity indicator. Starting with the number of comments, we note that 50% of the compared pairs of duplicates do not differ, i.e., have a ratio equal to 1. Interestingly, only 2% of the pairs differ by a factor larger than 100, and the largest registered ratio was 288,855. Analyzing the number of times added as a favorite, we observe that 36% of the compared pairs of duplicates do not differ, and 5% differ by more than 100 times. The largest observed ratio was 163,722. Lastly, we investigate the ratios of the total number of views. Only 2% of the pairs of duplicates did not exhibit any difference, 33% differed by less than 3 times, and 12% differed by more than 100 times. The largest observed ratio was for a pair of duplicates in which one of them had received 48,728,567 times more views than the other one.

In summary, we observed that some video duplicates, in spite of their similar content, may receive quite different evaluations as well as reach quite different popularity levels among users. These contextual differences among videos with similar contents may impact the system in several ways. Multiple copies of the same content dilute popularity, directly impacting the design of recommendation and ranking systems, since it is no longer straightforward to track the popularity of that content based on a single popularity indicator [7]. Another possible impact is related to the usage of caching systems and content distribution networks (CDNs), since the same content may be very popular but this popularity may be diluted among multiple videos that are not, individually, nearly as popular.

One intriguing and important question that arises from these analyses of popularity differences among duplicates is: *what factors influence the popularity of a video*? In the next section we analyze if the characteristics and actions of the duplicate owners can drive the popularity of their videos.

5.2 Duplicate owners

In the previous section we showed that there are differences in popularity indicators between videos with the same content. In this section we analyze if these differences may be (partially) caused by differences in characteristics of the owners of the videos.

Research has shown that many users access content through social networks [22]. Thus, one could think that popular users (e.g., users with many friends or owners of popular videos) tend to attract more visibility and popularity to their own duplicates, in comparison with other duplicates in the same group. In order to verify if such a trend exists, we consider the ranking of duplicates of each group in terms of the number of views, and we compute the Pearson correlation coefficient of this ranking with the ranking of their owners according to six different user characteristics: (1) number of views of all other videos (i.e., excluding the duplicate analyzed) owned by the user; (2) number of comments received by all other videos of the user; (3) sum of the ratings received by all other videos of the user; (4) number of times all other videos of the user were added as a favorite; (5) number of friends of the user, and (6) total number of videos uploaded by the user. For example, for a group of 2 videos, $C = 1$ if the owner of the most viewed video is the most popular user of the group (in terms of one of the aforementioned characteristics), and $C = -1$ otherwise. When a user had more than one duplicate in a group, we considered the average views of her videos in this group to calculate the video ranking.

Figure 5 shows the CDF of the measured correlations for each of the six aforementioned user characteristics.

We clearly note weaker correlations when considering the number of uploads done by the duplicate owners. In fact, 26% of the groups have significant negative correlation (lower than −0.5), and 26% have significant positive correlation (higher than 0.5). The remaining groups (48%) have weak (positive/negative) correlations. We also note that, considering the number of friends of the duplicate owners, the correlations between the two rankings are clearly more biased towards positive values. Illustrating, only 17% of the groups have strong negative correlations (<−0.5), while 36% of them have positive correlations higher than 0.5. Lastly, we observe that the distributions of correlations considering the remaining four characteristics are quite similar. Taking the total number of views as an example, only 15% of the groups present strong negative correlations, while 44% of them have correlations higher than 0.5. These results indicate that, in general, users with other popular videos tend to have the most popular duplicates. Intuitively, if a user has attracted a larger audience to her videos in the past, her audience would tend to have a certain degree of loyalty to new videos that she posts. Moreover, the number of friends of a user also has some impact (though weaker) on the popularity of her videos, indicating that part of their accesses may come from interactions established among users in the social network, similarly to observations performed in Flickr [22].

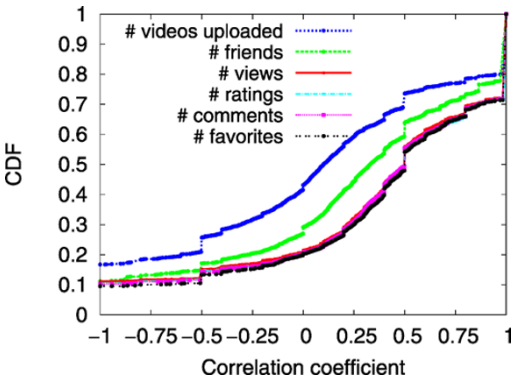
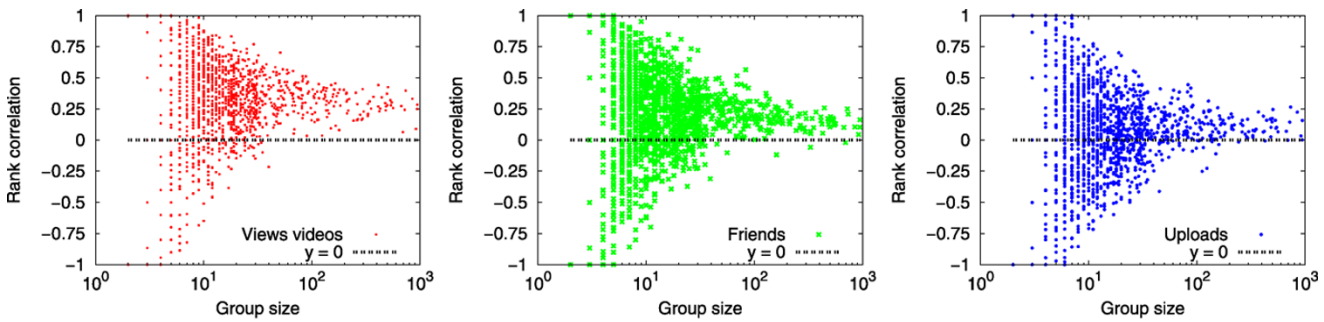


Fig. 5 (Color online) Distribution of correlations between duplicate popularity ranking (in number of views) and duplicate owner ranking according to different user characteristics

In the previous analysis, we cannot observe how the correlation varies with the size of the groups. Small groups have few possible values for the correlation coefficient, suggesting that the group size should also be considered. Accordingly, Fig. 6(a) shows the ranking correlations for different group sizes considering the ranking of users (duplicate owners) in terms of the number of views of all of their other videos. In fact, considering only groups with more than 10 duplicates, 92% of them have positive correlations. Moreover, the correlation is higher than 0.5 for 21% of them. The correlations considering the number of comments and the sum of ratings of all other videos of the duplicate owners are very similar across different group sizes. Considering only groups with more than 10 duplicates, 85% have a positive correlation, and this correlation is higher than 0.5 in 35% of them. Considering the ranking of duplicate owners based on the number of times that their videos were added as favorites in the system, 87% of the groups with more than 10 videos have a positive correlation, and about 32% have a correlation higher than 0.5. These last curves are omitted due to space constraints.

The correlations of rankings considering the number of friends of the duplicate owner for different group sizes are shown in Fig. 6(b). Considering the groups with more than 10 duplicates, around 84% have positive correlation, with about 9% having correlations higher than 0.5. As mentioned before, the correlations are somewhat weaker, particularly for large group sizes, if compared to those measured for the ranking based on the total number of views of the owner's videos.

Lastly, we analyze the ranking of users according to the number of videos uploaded. The correlation of this ranking with the ranking of duplicates, shown in Fig. 6(c), is even more skewed towards small and negative values than the correlations observed for the other metrics, as discussed before. For groups with more than 10 duplicates, 66% have positive correlations, while only 4% of them have correlations higher than 0.5.



(a) Total views from other videos (b) Number of friends (c) Number of videos uploaded

Fig. 6 Correlation between user characteristics and number of views of duplicated videos as a function of group size

This analysis revealed that, in general, users with other popular and qualified videos tend to have the most popular videos within a group of duplicates. In other words, a user with popular videos tends to gain a certain visibility in the system, which might reflect in the popularity of her new videos. Such information is important not only to the association of advertisements to videos, but also for the creation of mechanisms of caching and video prefetching.

5.3 Metadata

One of the most common ways for finding videos in video sharing systems is through search engines. According to Comscore [10], if YouTube was a standalone site, it would be the second largest search engine after Google in terms of search volume. Most current information retrieval mechanisms for video content (search, in particular) rely primarily on metadata (e.g., tags) that users associate with each object, typically to describe its content. The more accurate the description of the video content provided by the user in its associated metadata, the better its chance to be found by other users and thus, the more popular the video may become.

YouTube allows video owners to independently associate three basic types of metadata to their videos, namely, a title, a text describing the video, and tags. Moreover, the user necessarily needs to associate one predefined category to her videos. In the following, we analyze the degree of similarity in the contents of tags, descriptions, and titles of video duplicates (Sect. 5.3.1) as well as the similarity between their assigned categories (Sect. 5.3.2).

5.3.1 Tags, title, and description

We start by providing a brief description of the procedures used to prepare our dataset. First, we took the words that make up the content of tags, title, and description of each duplicate in our dataset, considering only numbers and letters, and discarding special characters such as hyphens and punctuation marks. Moreover, we also reduced each word to its radical, using the well-known Porter stemming method [20] (available at http://tartarus.org/~martin/PorterStemmer). We also filtered stop words, i.e., words with no semantic meaning (e.g., "the", "of", "for", etc.), obtained from a list available on reference [26]. Although the searches performed as part of our crawling strategy were done with random words from an English dictionary, there might be words written in other languages in the metadata fields. For this reason, we considered only videos in which the owners are from the United States, Canada, United Kingdom, and Australia for this analysis. These videos correspond to 33% of the total number of videos in our dataset.

Next, we characterized the contents of each metadata field (i.e., tags, title, and description) associated with the analyzed duplicates with respect to the number of unique terms. We refer to it as the size of the *set of terms* of the metadata field. Figure 7 shows the CDFs of the sizes of the set of terms of tags, title, and description associated with the analyzed duplicates. Titles have a stronger bias towards fewer terms, with a significantly shorter tail. Tags and description have somewhat similar distributions, with significant differences only at the tail. In particular, 87% of the titles have fewer than 6 terms, while around 68% of the duplicates have fewer than 6 terms in their associated tags as well as in their descriptions. Table 3 summarizes these findings, presenting median, average, minimum, and maximum values for each distribution. While titles have on average only 3.4 terms, tags and descriptions have very similar average (5.7 and 5.6, respectively) and median (5) values, although descriptions have a much larger maximum size of set of terms (341 against only 50).

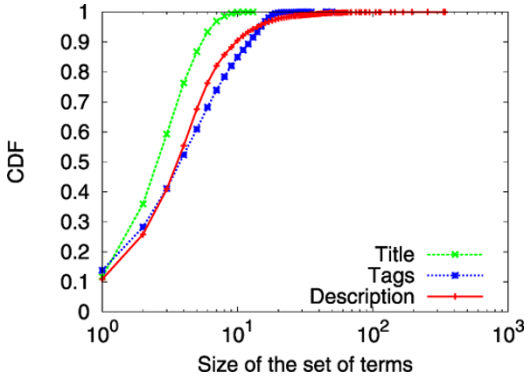


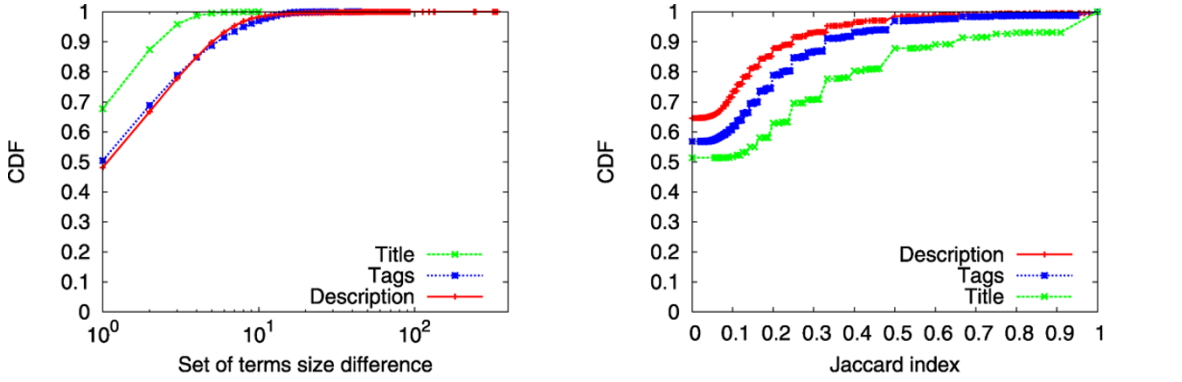
Fig. 7 Distribution of the sizes of the set of terms

Table 3 Statistics on the sizes of the set of terms

Metadata	Median	Avg	Max	Min	Zeros
Tags	5	5.7	50	1	192 (0.8%)
Title	2	3.4	13	1	151 (0.6%)
Description	5	5.6	341	1	3,379 (13.6%)

The last column of Table 3 shows the number of sets with size equal to zero (i.e., no term). We observed that the removal of words with no semantic meaning and containing only special characters, in the preparation of our dataset, reduced some of the metadata of several videos to an empty set. In particular, fewer than 1% of the duplicates have empty sets of terms in their associated tags and titles, while almost 14% of them have no terms in their descriptions. These videos are not considered in the analyses presented in this section. Moreover, before the preparation of our dataset, all duplicates had at least one word in their title, while 24 duplicates had no tags, and 3.112 duplicates had empty descriptions.

We are now ready to analyze the similarity between metadata fields associated with pairs of duplicates of the same



(a) Differences in the sizes of the set of terms (b) Distribution of Jaccard coefficient for tags, description, and title

Fig. 8 Metadata analyses: tags, title, and description

video. We start by analyzing the similarity in the number of terms used by the user in each metadata field across different duplicates in each group. We thus calculate, for each pair of duplicates, the differences between the sizes of the sets of terms in their associated tags, titles, and descriptions. The CDFs of these differences, shown in Fig. 8(a), indicate that both tags and description exhibit similar trends, with a stronger bias towards more significant differences (larger values) if compared to title. In particular, almost 28% of the pairs of duplicates do not differ with regards to the size of the set of terms in their title, while only 19% of the pairs have the same number of terms in either their tags or descriptions. Moreover, only approximately 1% of the pairs of duplicates differ by more than 4 terms in their titles, while this number is approximately 15% for tags and description. Thus, in general, we do observe some degree of dissimilarity in the number of terms used by different users to describe the same content, and this degree is more significant in the metadata fields that typically contain more terms (i.e., tags and description).

Next, we analyze the similarity of the contents (i.e., the set of terms) of the same metadata field across different duplicates. To that end, we use the Jaccard coefficient [2], defined as follows. Let A and B be the sets of terms in a given metadata (say, tags) of two duplicates in the same group. The Jaccard coefficient, $J(A, B)$, between A and B is given by the number of terms in common in A and B divided by the total number of terms in the union of both sets:

$$J(A, B) = |A \cap B|/|A \cup B| \quad (1)$$

A Jaccard coefficient J equal to 0 means that the two sets of tags have no term in common, whereas J close to 1 indicates that both sets share most of the terms.

Figure 8(b) shows the CDFs of the Jaccard coefficients of all pairs of duplicates within the same group, for all groups in our dataset, considering, separately, tags, description, and title of each pair. We note that a significant fraction of the pairs of duplicates have no term in common in their tags (about 56%), and 87% of them have a J value smaller than 0.3. Only 1.2% of the pairs share all tags. We also note that the Jaccard coefficients tend to larger values for title, indicating a stronger similarity in the contents of this metadata field in different duplicate pairs. In fact, 30% of the pairs of titles of duplicates have J values greater than 0.3, and 7% have J values equal to 1. Description presents the lowest levels of content similarity: 7% of the pairs have J values greater than 0.3, and only 0.5% have the entire description in common.

In conclusion we note that, in general, duplicates have low similarity in terms of metadata. Titles present a higher degree of similarity among duplicates, which indicates that different users better agree on the terms chosen for the title to represent the same content. In addition, although tags and description have similar distributions of the sizes of their sets of terms, description presents the lowest degree of content similarity (i.e., Jaccard coefficients), which may be explained by the nature of this metadata field. For tags, users are just supposed to choose a set of (not necessarily) related words to represent the content, while for description users are supposed to write some meaningful structured text describing the content of the video.

5.3.2 *Categories*

YouTube allows users to choose among 14 predefined categories. Table 4 lists these categories as well as the abbreviations used in this paper.

Figure 9 shows the distribution of categories across all duplicates. As we can see, *Music* and *Comedy* are the most popular categories, with 23.8% and 23.6% of the duplicates, respectively. *Entertainment* and *People & Blogs* are the following most popular, covering 16.3% and 9.1%, respectively. The other 10 categories account, in total, for 27% of our collected duplicates.

In order to understand the differences in the categories of duplicates, we compared the categories of each pair of duplicates within a group, for all groups, counting the occurrences of each possible combination of categories. Table 5 shows this distribution. Each line refers to the distribution of videos of one specific category and reports the fractions of their duplicates falling into each of the 14 categories (each line sums up to 100%). As we can see, around 89% of duplicates of videos in category *Travel & Events* are also associated with *Travel & Events*, indicating that users from these categories usually agree on its association with their videos. However, for most of the categories, there is a nonnegligible fraction of the duplicate pairs that are associated with different categories. Particularly, the categories *Howto & Style*, *Education,* and *Nonprofits & Activism* have low fractions of duplicate pairs associated with the same category (12.7%, 5.4%, and 1.5%, respectively), indicating that users do not agree about them. In general, duplicates from these categories are associated with *Comedy, Entertainment*, *Music*, and *People & Blogs*, which are the categories most popular in our database, as shown in Fig. 9. In fact, observing the columns of Table 5, we note that these categories are very popular in terms of occurrences of pairs of duplicates in common.

Generally speaking, our results reflect how users perceive the same content differently. In fact, the categorization of videos is subjective, and a video might be naturally associated with different categories. For example, one of the most viewed videos of all time in YouTube, named "Evolution of Dance," shows a man dancing famous songs which were hits in different decades. This video was associated with the category *Comedy* but could be naturally associated with other categories, such as *Entertainment* and *Music*.

Table 4 List of YouTube predefined categories

Abbreviation	Category name
Com	Comedy
N&P	News & Politics
H&S	Howto & Style
Ent	Entertainment
Edu	Education
S&T	Science & Technology
Mus	Music
A&V	Auto & Vehicles
P&A	Pets & Animals
T&E	Travel & Events
P&B	People & Blogs
F&A	Film & Animation
N&A	Nonprofits & Activism
Spo	Sports

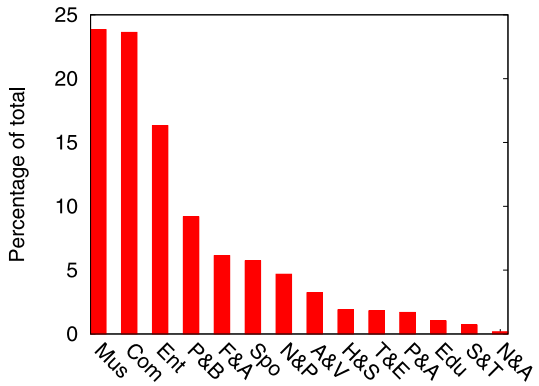


Fig. 9 Distribution of categories

Table 5 Distribution of categories of pairs of duplicates in the same group

	Com	N&P	H&S	Ent	Edu	S&T	Mus	A&V	P&A	T&E	P&B	F&A	N&A	Spo
Com	**61.8**	1.4	0.4	14.4	0.1	0.0	3.3	2.1	2.0	0.4	8.3	4.2	0.0	1.6
N&P	35.4	**17.7**	1.9	10.8	1.7	0.5	3.0	6.7	0.7	1.0	14.1	4.1	0.4	2.0
H&S	21.0	4.0	**12.7**	21.5	1.6	2.2	3.2	10.1	1.1	1.2	11.1	5.5	0.1	4.6
Ent	33.7	1.0	0.9	**33.8**	0.3	0.2	4.9	3.5	1.5	0.5	7.9	8.9	0.0	3.0
Edu	14.4	6.3	2.9	14.5	**5.4**	1.5	5.6	13.5	0.6	0.8	24.4	7.4	0.4	2.3
S&T	7.8	3.3	7.2	17.9	2.9	**45.2**	1.2	3.7	0.1	1.0	5.7	3.2	0.1	0.6
Mus	18.3	0.6	0.3	11.4	0.3	0.0	**50.6**	0.6	1.0	0.3	6.5	8.4	0.0	1.5
A&V	21.6	2.7	1.9	15.4	1.4	0.2	1.2	**38.1**	0.3	2.4	9.8	3.6	0.1	1.2
P&A	38.9	0.5	0.4	11.9	0.1	0.0	3.5	0.5	**30.4**	0.3	5.1	7.5	0.0	0.9
T&E	3.2	0.4	0.2	2.0	0.1	0.0	0.5	2.1	0.1	**89.5**	1.0	0.6	0.0	0.5
P&B	40.8	2.7	1.0	16.8	1.2	0.2	5.8	4.7	1.3	0.5	**15.7**	5.1	0.1	4.1
F&A	27.6	1.1	0.6	25.2	0.5	0.1	10.2	2.3	2.6	0.4	6.8	**20.8**	0.0	1.7
N&A	4.0	22.5	3.1	10.2	5.8	0.8	5.6	14.0	0.5	1.6	23.4	6.3	**1.5**	0.7
Spo	7.4	0.4	0.4	6.1	0.1	0.0	1.3	0.6	0.2	0.2	3.8	1.2	0.0	**78.3**

6 Duplicate content creation

In the previous section we showed that, although duplicates have similar or identical content, they present several contextual differences. Some of these differences are usually associated with the degree of subjectivity and the different perceptions that multiple users may take from the same content either when they watch it or when they associate metadata to it. In this section we focus on possible reasons that lead users to create duplicates.

6.1 Users and their duplicates

Since users can freely create content in online video sharing systems, some accidental creation of duplicates is expected. In fact, most of the videos of the users in our dataset are not duplicates. Only 7% of the users have more than 50% of their videos as duplicates. By computing the ranking of users ordered according to the number of duplicates created (here we consider duplicates from all groups in our dataset), we note that most of the users post few duplicates (98% post less than 3 duplicates), as expected when duplicate creation is accidental. However, there are users who create a large number of duplicates, i.e., the first and second users of the ranking have 714 and 103 duplicates.

So, who are these users who create so many duplicates? We manually inspected the first two users of the ranking and found that all the duplicates created by each of them refer, in fact, to the same content (all duplicates belong to the same group). Moreover, we also found that the first user of the ranking is what we refer to as a *video response promoter*, who exploits the video response feature [5, 6]. The YouTube video response feature allows users to video respond to another user's video contribution. A user willing to promote a certain video may post a large number of video responses to her target aiming to promote it quickly to the lists of most responded videos. The duplicates posted by the first user of the ranking consist of videos of short duration (i.e., less than 5 seconds).

The second user of the ranking, on the other hand, created several duplicates of the same advertisement, but assigned different tags (as well as title and description) to each duplicate. This kind of behavior is known as tag spamming [16, 21], and it is used to fool search engines with nonrelated tags in order to promote some content. The actions of these two users are examples of suspicious (i.e., opportunistic) behavior associated with the creation of duplicates observed in our dataset. In the following, we further investigate the presence of suspicious creation of multiple duplicates by the same user.

6.2 Suspicious duplicate creation

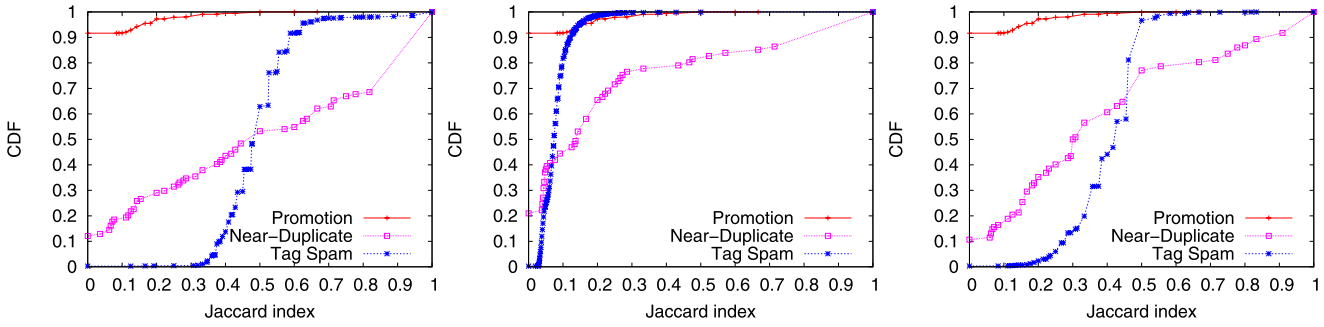
In order to investigate if there are more duplicates of the same content created by the same user, we define a set of suspicious videos as follows. If, for a group of duplicates, the ratio between the number of duplicates to the number of unique duplicate owners is smaller than one, it means that at least one user created two duplicates in that group. We call suspicious videos all duplicates of a group created by a single user. Suspicious users are the owners of suspicious videos.

We found that 92% of the groups of duplicates in our dataset do not have suspicious videos. In total, 2,668 videos created by 608 different users were considered suspicious. For comparison purposes, the rest of the duplicates are referred to as legitimate videos. Since we are interested in studying the metadata of suspicious videos, we again focus on the English language, restricting our dataset to videos in which the owners are from the United States, Canada, United Kingdom, and Australia.

We manually analyzed 1,032 suspicious videos from 71 randomly selected users. As a result, the suspicious videos were divided into three sets: (1) 714 videos used for promotion (these are the videos created by the user who posts more duplicates in our dataset); (2) 159 videos with tag spam, created by 26 different users, including the second of the ranking; and (3) near-duplicates. By near-duplicates we mean similar or identical videos with different quality or with subtitles in a different language, identical videos with different comments embedded, and identical videos with different audio. Interestingly, we noted that most of the near-duplicates complement the original material with additional information, which can be valuable to some users. In total, we have 159 videos from 44 users in the near-duplicate group.

Next, we compare the similarity in the tags, description, and title among pairs of duplicates from the selected suspicious videos. Figure 10 shows the CDF of the Jaccard coefficient for each type of metadata. We compare, for each type of metadata, three sets of duplicates: videos used for promotion, videos with tag spam, and near-duplicates. For each set of videos, we compare all possible pairs of duplicates within the same group.

Observing Fig. 10(a), we note that the set of videos used for promotion presents the lowest degree of tag similarity. As an example, fewer than 8% of the videos have a Jaccard coefficient higher than 0.1. By analyzing the duplicates used for promotion, we noted that each video has only a small set of tags which seem to be generated automatically, since most of them do not exist in an English dictionary and do not seem to have a real meaning. This observation may explain the low degree of similarity encountered. At the other extreme, the set of near-duplicates presents the highest Jaccard coefficients in comparison with other sets of suspicious videos. For instance, 64% and 80% of the videos have Jaccard coefficients higher than 0.3 and 0.1, respectively. In other words, users who create videos with only differences in quality, or with embedded comments or subtitles, tend to assign basically the same tags to the videos. Interestingly, most of the



(a) Tags (b) Description (c) Title

Fig. 10 Similarity between metadata of suspicious videos

videos used for tag spamming present a high concentration of Jaccard coefficient between 0.3 and 0.6 (90% of them). By analyzing some videos with tag spam we noted that most of them have a set of tags which really describes the content of the video and is used for all duplicates of a single tag spam attack.

We also made the same comparison for description and title, depicted in Figs. 10(b) and 10(c), respectively. Generally speaking, the same observations for tags hold for description and title.

7 Concluding remarks

This paper offers a novel analysis of duplicates in video sharing systems. It focuses on understanding the social perception of duplicate content represented by videos marked as duplicate by YouTube algorithms. We analyze duplicated videos from the viewpoints of both the content producer (i.e., creator/owner) and the consumer (i.e., user). We show that owners of high quality and popular videos tend to make their duplicates more popular. We also show that there is a correlation between popularity of duplicates and the number of friendship links a owner has, indicating the importance of the network to attract views to a video. These findings can be useful to support several mechanisms, such as advertising placement schemes and caching approaches.

From the users' viewpoint, we quantify the degree of similarity of different types of metadata associated by different users to videos with the same content. This is important to show that there are differences in the way users view and describe a video content. Video is a rich media full of information, which is hard to describe using only words. Video retrieval mechanisms rely on the metadata that users associate to describe the content of a video, so the low similarity encountered in this work can contribute to support the design of more effective multimedia content retrieval algorithms.

Our results also unveil the existence of opportunistic behavior in the creation of some duplicates. By analyzing a set of specific videos, we found videos that are uniquely used for promotion and also videos containing tag spam. The analysis of the similarity between tags of suspicious videos uncovers intrinsic characteristics of the metadata used in opportunistic activities.

We envision a couple of directions towards which this work can evolve in the future. The first one should focus on a deeper analysis of opportunistic behavior detected in our collection of duplicates. This can be expanded to detect offending or controversial videos. The second one relates to understanding the temporal evolution of the popularity of duplicates. Duplicate analysis is also useful to shed light on understanding the process adopted by users to associate metadata to video objects.

Acknowledgements This work is partially supported by the INCT-Web (MCT/CNPq grant 57.3871/2008-6), and by the authors' individual grants and scholarships from CNPq, FAPEMIG, and CAPES.

References

1. Adar E, Zhang L, Adamic L, Lukose R (2004) Implicit structure and the dynamics of blogspace. In: Workshop on the Weblogging Ecosystem
2. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. ACM/Addison-Wesley, New York/Reading
3. Benevenuto F, Duarte F, Rodrigues T, Almeida V, Almeida J, Ross K (2008) Understanding video interactions in youtube. In: ACM int'l conference on multimedia (MM)
4. Benevenuto F, Rodrigues T, Almeida V, Almeida J, Zhang C, Ross K (2008) Identifying video spammers in online social networks. In: Workshop on adversarial information retrieval on the web (AIRWeb)
5. Benevenuto F, Rodrigues T, Almeida V, Almeida J, Gonçalves M (2009) Detecting spammers and content promoters in online video social networks. In: Int'l ACM SIGIR
6. Benevenuto F, Rodrigues T, Almeida V, Almeida J, Ross K (2009) Video interactions in online video social networks. In: ACM trans on multimedia computing, communications and applications (TOMCCAP)
7. Cha M, Kwak H, Rodriguez P, Ahn Y, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: ACM SIGCOMM conference on Internet measurement (IMC)

8. Cherubini M, Oliveira R, Oliver N (2009) Understanding near-duplicate videos: a user-centric approach. In: ACM int'l conference on multimedia (MM)
9. Comscore (2010) http://www.comscore.com/. June 2010
10. Comscore (2010) Youtube now 25 percent of all Google searches. http://tinyurl.com/4t32l4. June 2010
11. del.icio.us web site (2010) http://www.delicious.com. June 2010
12. Flickr web site (2010) http://www.flickr.com. June 2010
13. Gill P, Arlitt M, Li Z, Mahanti A (2007) Youtube traffic characterization: a view from the edge. In: ACM SIGCOMM conference on Internet measurement (IMC)
14. Golbeck J (2008) Trust and nuanced profile similarity in online social networks. Technical report
15. Hauptmann A, Wu X, Ngo C, Tan H (2009) Real-time near-duplicate elimination for web video search with content and context. IEEE Trans Multimedia 11(2):196–207
16. Heymann P, Koutrika G, Garcia-Molina H (2007) Fighting spam on social web sites: a survey of approaches and future challenges. IEEE Internet Comput 11:36–45
17. Huang Z, Wang L, Shen H, Shao J, Zhou X (2009) Online near-duplicate video clip detection and retrieval: an accurate and fast system. In: IEEE int'l conference on data engineering (ICDE)
18. Ispell (2010) http://www.gnu.org/software/ispell/ispell.html. June 2010
19. Jain R (1991) The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. Wiley, New York
20. Jones KS, Willett P (eds) (1997) Readings in information retrieval. Morgan Kaufmann, San Mateo
21. Koutrika G, Effendi F, Gyöngyi Z, Heymann P, Garcia-Molina H (2007) Combating spam in tagging systems. In: Workshop on adversarial information retrieval on the Web (AIRWeb)
22. Lerman K, Jones L (2007) Social browsing on Flickr. In: Int'l conference on weblogs and social media (ICWSM)
23. Li X, Guo L, Zhao Y (2008) Tag-based social interest discovery. In: Int'l World Wide Web conference (WWW)
24. Marshall CC (2009) No bull, no spin: a comparison of tags with other forms of user metadata. In: ACM/IEEE conference on digital libraries (JCDL)
25. Oliveira R, Cherubini M, Oliver N (2009) Human perception of near-duplicate videos. In: Int'l conference on human-computer interaction (INTERACT)
26. Rijsbergen C (1979) Information retrieval. Butterworth, Stoneham
27. Rodrigues T, Benevenuto F, Almeida V, Almeida J, Gonçalves M (2009) Uma análise contextual de conteúdo duplicado no youtube. In: Simpósio Brasileiro de sistemas multimídia e Web (WebMedia)
28. Suchanek F, Vojnovic M, Gunawardena D (2008) Social tags: meaning and suggestions. In: ACM conference on information and knowledge management (CIKM)
29. Tan H-K, Ngo C-W, Hong R, Chua T-S (2009) Scalable detection of partial near-duplicate videos by visual-temporal consistency. In: ACM international conference on multimedia (MM)
30. Wu X, Hauptmann A, Ngo C (2007) Practical elimination of near-duplicates from web video search. In: Int'l conference on multimedia
31. Zhu J, Hoi S, Lyu M, Yan S (2008) Near-duplicate keyframe retrieval by nonrigid image matching. In: ACM int'l conference on multimedia (MM)
32. Zink M, Suh K, Gu Y, Kurose J (2008) Watch global, cache local: Youtube network traces at a campus network—measurements and implications. In: IEEE multimedia computing and networking (MMCN)