# The Tube over Time: Characterizing Popularity Growth of YouTube Videos

Flavio Figueiredo Fabrício Benevenuto Jussara M. Almeida
{flaviov, fabricio, jussara}@dcc.ufmg.br
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - Brazil

## ABSTRACT

Understanding content popularity growth is of great importance to Internet service providers, content creators and online marketers. In this work, we characterize the growth patterns of video popularity on the currently most popular video sharing application, namely YouTube. Using newly provided data by the application, we analyze how the popularity of individual videos evolves since the video's upload time. Moreover, addressing a key aspect that has been mostly overlooked by previous work, we characterize the types of the referrers that most often attracted users to each video, aiming at shedding some light into the mechanisms (e.g., searching or external linking) that often drive users towards a video, and thus contribute to popularity growth. Our analyses are performed separately for three video datasets, namely, videos that appear in the YouTube top lists, videos removed from the system due to copyright violation, and videos selected according to random queries submitted to YouTube's search engine. Our results show that popularity growth patterns depend on the video dataset. In particular, copyright protected videos tend to get most of their views much earlier in their lifetimes, often exhibiting a popularity growth characterized by a viral epidemic-like propagation process. In contrast, videos in the top lists tend to experience sudden significant bursts of popularity. We also show that not only search but also other YouTube internal mechanisms play important roles to attract users to videos in all three datasets.

## Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Measurement techniques*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

## General Terms

Human Factors, Measurement


*WSDM'11,* February 9–12, 2011, Hong Kong, China.


## Keywords

YouTube, video popularity, popularity growth, referrer

## 1. INTRODUCTION

Understanding content popularity growth on the Internet is of great relevance to a broad range of services, from technological, economical and social perspectives. Such understanding can drive the design of cost-effective caching and content distribution mechanisms as well as uncover potential bottlenecks in system components such as search engines [6]. Moreover, predicting popularity is also important not only for supporting online and viral marketing strategies as well as effective information services (e.g., content recommendation and searching services) [12] but also because it may uncover new (online and offline) business opportunities. From a sociological point of view, a deep study of popularity evolution may also reveal properties and rules governing collective user behavior [10].

Online Social Networks (OSNs) are currently a major segment of the Internet. Considering video sharing OSNs, YouTube[1] is the one with the largest number of registered users [1], who upload and share their videos at a staggering rate. Indeed, it has been reported that the amount of content uploaded to YouTube in 60 days is equivalent to the content that would have been broadcasted for 60 years, without interruption, by NBC, CBS and ABC altogether [2]. Moreover, YouTube has reportedly served over 100 million users only on January 2009 [1], with a video upload rate equivalent to 10 hours per minute [2]. At such unprecedented user and content growth rates, understanding video popularity on YouTube becomes a challenge of utmost importance, as the myriad of different contents make user behavior and attention span highly variable and unpredictable [6].

As argued by Willinger *et al.* [20], most previous analyses of OSNs have treated such systems as static. Most of them focus on analyzing structural properties of single snapshots of relationship networks (e.g., friendship network) that emerge in such systems [3, 5, 15]. However, since OSNs are inherently dynamic, these studies fail to address key properties of the underlying system dynamics. Regarding one such property, namely popularity, a few studies have analyzed YouTube with respect to video popularity characteristics [6, 9, 10] and prediction [14, 19]. However, most of them, despite covering a rich set of popularity properties and their implications for system design, focused on only a single or

[1] `http://www.youtube.com`
[2] `http://www.youtube.com/t/fact_sheet`

at most on a few snapshots of the system, and thus do not deeply analyze the long-term popularity growth patterns for *individual videos* [6, 11]. To the best of our knowledge, the only long-range studies of popularity evolutionary patterns at the granularity of individual videos focus mainly on designing popularity prediction models [10, 19], lacking a discussion on possible sources of video popularity, that is, on mechanisms that attract users to the videos.

In this paper, we analyze popularity growth patterns of YouTube videos with two main goals. First, we intend to analyze how the popularity of individual videos evolves over time, covering the whole period since the video was uploaded to the system. Second, shedding some light into an aspect of popularity that has been mostly overlooked, we aim to investigate how users reach each given video (e.g., by searching on YouTube or following a link in another website), as a means to understand which mechanisms contribute to a video's popularity. Thus, our work is complementary to all previous analyses of YouTube video popularity.

Towards our goals, we crawled YouTube, collecting a new set of statistics available in the system, which provide, for each video: (a) its popularity (according to different metrics) as a function of time, and (b) a set of referrers, that is, links used by users to access the video, along with the number of views for which each referrer is responsible. Given the great diversity of content on YouTube and the multitude of factors that may impact video popularity, we collected data for three different datasets, namely, (a) popular videos that appear on the world-wide top lists maintained by YouTube; (b) videos that were removed from the system due to copyright violation; and, (c) videos sampled according to a random procedure (i.e., random queries). For each collected dataset, and for different video classes defined according to their ages in the system, we characterized popularity growth patterns, correlating popularity with different types of referrers which caught user attention.

We believe the present work provides valuable insights for Internet service and content providers, who can improve the effectiveness of several services, including caching, content delivery networks, searching and content recommendation, by leveraging in these systems not only information on the amount of views a video receives, but also external sources of influence on popularity. They can also help content creators and online marketers to better understand what makes a video popular, driving their future actions.

The rest of this paper is organized as follows. Section 2 briefly discusses related work. Our data collection methodology is described in Section 3, whereas our main results are presented in Sections 4 and 5. Section 6 concludes the paper and offers some directions for future work.

## 2. RELATED WORK

There have been a few studies that address content popularity on OSNs, and, particularly, on video sharing systems. Cha *et al.* [6] presented an in-depth study of two video sharing systems. They analyzed popularity distribution, popularity evolution and content characteristics of YouTube and of a popular Korean video sharing service, and investigated mechanisms to improve video distribution, such as caching and Peer-to-Peer (P2P) distribution schemes. Additionally, they presented the first evidence of the existence of duplicates in user generated content, discussing potential problems they may cause to the system.

Gill *et al.* [11] characterized the YouTube traffic collected at the University of Calgary campus network, comparing its properties with those previously reported for Web and streaming media workloads. In particular, they analyzed daily and weekly patterns as well as several video characteristics such as duration, bit rate, age, ratings, and category. Zink *et al.* [21] also characterized the traffic collected from a university campus. Based on trace-driven simulations, they showed that client-based local caching, P2P-based distribution, and proxy caching can reduce network traffic significantly, allowing faster access to videos.

In common, these studies provide important insights into content popularity and traffic caused by video sharing services. However, they only focused on either a single and static snapshot of the videos and of the traffic generated to them [11, 21] or on at most a few snapshots [6]. Thus, they did not analyze the long-term popularity growth of videos.

A few recent efforts have looked at the time component of video popularity [10, 19]. Crane and Sornette proposed epidemic models to understand how a popularity burst can be explained in terms of a combination of endogenous user interactions and external events [10]. They distinguished four different evolution patterns, which are further discussed in Section 4.3. Szabo and Huberman presented a method for predicting popularity of YouTube and Digg[3] content from early measurements of user accesses [19]. More recently, Lerman and Hogg [14] developed stochastic user behavior models to predict popularity based on early user reactions to new content. They improved on predictions based on simple extrapolations from early votes by incorporating aspects of the web site design, validating their approach on Digg.

Another interesting work on popularity evolution in social media was performed by Ratkiewicz *et al.* [17]. By analyzing traffic logs and data from Google Trends[4], the authors investigated how external events, captured by search volume on Google Trends and local browsing (i.e., university/community traffic), affect the popularity of Wikipedia articles. Although this work, and the aforementioned efforts, provide some insights into the long term evolution of content popularity, there is still little knowledge about what kinds of different external events (e.g., being featured on the front page) and system mechanisms (e.g., search) contribute the most to popularity growth. Thus, our analyses and findings, performed separately for YouTube videos with different characteristics, greatly build on previous efforts, shedding more light into the complex task of understanding content popularity on OSNs.

## 3. MEASUREMENT METHODOLOGY

Since we intend to study video popularity growth on YouTube, we need to collect (a) video popularity as a function of time, and (b) video referrers, i.e., the links that users used to access the videos. In this section, we first introduce a new set of YouTube statistics that provide both types of information. We then describe our data collection methodology as well as the limitations of the collected data.

### 3.1 YouTube Statistics

Recently, YouTube has launched a statistics feature that provides a unique opportunity to study video popularity.

[3] `http://www.digg.com`
[4] `http://trends.google.com`

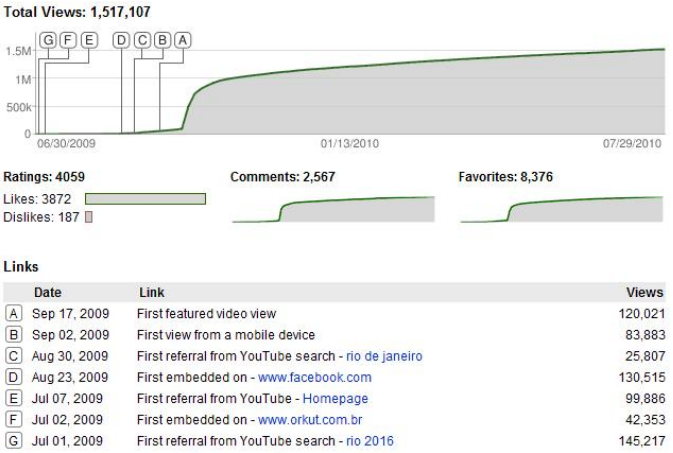



**Figure 1: Example of YouTube statistics.**

Figure 1 shows an example of such statistics for the video theme of the 2016 Olympic Games. There are two sets of valuable information: (1) the cumulative distributions of popularity as a function of time for three popularity metrics, namely, number of views, number of comments, and number of users that marked the video as favorite; and, (2) a list of important referrers that led users to the video containing, for each referrer, the number of views for which it is responsible and the date it was first used to reach the video.

Figure 1 shows that the cumulative growth of the number of views experienced by the video presents three clear distinct phases. Initially, the video stays dormant and unattractive to most YouTube users. The first registered referrer is related to the search for the query *rio 2016* (referrer G). Moreover, before being indexed and available on search results for the query *rio de janeiro* (referrer C), the video first appeared in online social network sites such as Orkut[5] and Facebook[6] (referrers D and F). Finally, still in a dormant phase, the video was featured on the first page of YouTube (referrer A), becoming quickly very popular. Shortly afterwards, popularity growth rate changed once again, remaining very slow through the rest of the curve.

We here analyze YouTube video popularity by exploiting the set of statistics shown in Figure 1. Next, we describe our collection methodology as well as the set of collected videos used in our analyses.

## 3.2 Data Collection Methodology

The graph shown in Figure 1 was plotted by YouTube using the Google charts API[7]. For each popularity metric (number of views, number of comments and number of favorites), YouTube requests the Google charts API, providing on the requested URL one hundred pairs of (x,y) values used to plot the graphs. For each collected video, we gathered these (x,y) values by collecting the URL requested by YouTube. In addition to the popularity growth curves, we also collected all the referrers listed by YouTube [8].

[5]`http://www.orkut.com`
[6]`http://www.facebook.com`
[7]`http://code.google.com/apis/chart/`
[8]Visit `http://vod.dcc.ufmg.br/traces/youtime/` for information on data availability.

Given the diversity of video types on YouTube and the various factors that may influence video popularity, we analyze three different video sets:

**Top**: YouTube maintains several top lists (e.g., most viewed and most commented videos) as a means to help users finding popular content and new trends. Each top list contains one hundred videos. YouTube provides per country and world-wide top lists, and allows users to browse them in different time scales, i.e., top of the day, week, month, and top of all time. We created our Top dataset by collecting all the world-wide top lists available on YouTube, gathering a total of 27,212 unique videos.

**YouTomb**: A second group of interesting videos to be studied is the copyright protected videos. YouTube users may inadvertently or even maliciously introduce in the system unauthorized copies of videos that are protected by the copyright owner's exclusive rights. To our knowledge, there is no previous study on how people find and disseminate these videos. Recently, an MIT project, called YouTomb, started to monitor a large amount of YouTube videos. They register in a public database the identifiers of all monitored videos that are removed from YouTube, along with the reason for which they were removed. We used the YouTomb database in order to obtain videos that had been removed from YouTube due to copyright violation. Surprisingly, we found that we were still able to collect the popularity statistics of such videos. Using the video identifiers provided by the YouTomb database, we collected a total of 120,862 videos that had been removed from YouTube due to copyright violation.

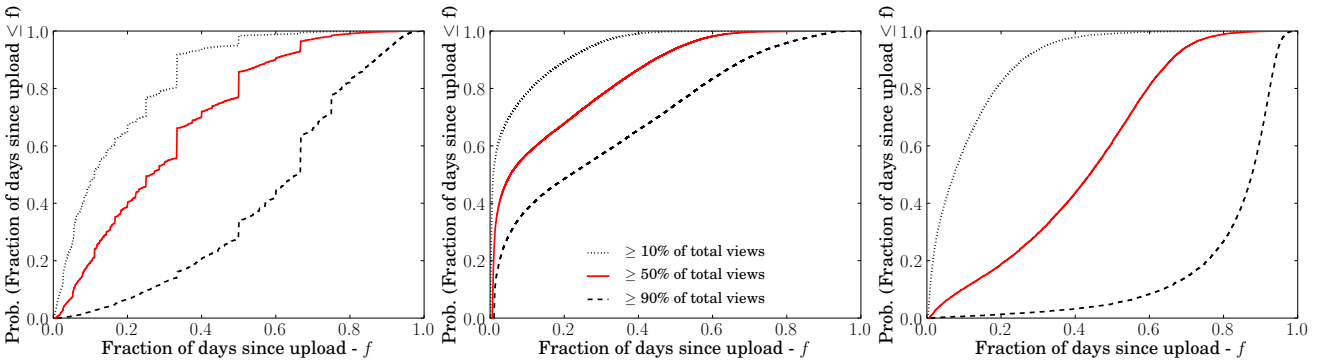
**Random topics**: As basis for comparison, we also want to study popularity growth of a random sample of YouTube videos. Ideally, we would like to have at our disposal the complete set of YouTube videos in order to select a random sample of them. Unfortunately, YouTube does not provide a means to systematically collect all the videos and neither a random sample of them. Instead, we designed a sampling procedure that is based on random topics. First we selected 30,000 entities (i.e., words and proper names) from the Yago lexical ontology [18]. We then used the YouTube search API to retrieve the first result on each selected entity. In total, we collected 24,484 unique videos using this strategy.

For each dataset, our crawler was executed in a single day on April 2010. Throughout this paper, we refer to our datasets as *Top*, *YouTomb* and *Random*[9].

## 3.3 Collected Datasets

We processed our collected datasets to remove: (1) videos with missing or inconsistent information; and, (2) very recent videos (i.e., uploaded on the same day of our crawling). Table 1 provides a summary of each cleaned video datasets, presenting the total and average numbers of views as well as the average video age. Video age, measured in number of days, is defined as the difference between the crawling date (or the removal date, for videos in the YouTomb dataset) and the upload date. We note that, YouTomb videos are on average older than videos in the Top and Random datasets. Moreover, Top videos are, as expected, more popular, on average, than videos in the other two datasets, whereas

[9]Even though we use the term *Random*, we are not claiming to have a truly random sample of YouTube videos, but rather a sample of videos on *random topics*.




(a) Top (b) YouTomb (c) Random

**Figure 2: Cumulative distributions of time until video achieves at least 10%, 50% and 90% of its total views (time normalized by video's lifetime).**

**Table 1: Crawled datasets.**

| Video Datasets | Top | YouTomb | Random |
|---|---|---|---|
| # Videos | 17,127 | 73,257 | 18,095 |
| Avg # of views/videos | 843,001 | 279,486 | 126,056 |
| Average age (# days) | 136 | 627 | 493 |

**Table 2: Number of videos across age ($a$) ranges.**

| | Top | YouTomb | Random |
|---|---|---|---|
| $a \leq 7$ days | 4,609 | 112 | 136 |
| 7 days $< a \leq 1$ month | 4,344 | 14,553 | 7,649 |
| 1 month $< a \leq 1$ year | 6,093 | 249 | 515 |
| $a > 1$ year | 2,081 | 58,343 | 9,795 |

YouTomb videos tend also to attract more views than videos in the Random dataset (on average).

Although this data gives us a unique opportunity to examine video popularity growth, it has some limitations. Each popularity growth curve is registered with at most 100 points, regardless of the age of the video. In order to be able to estimate video popularity on a daily basis, we performed linear interpolation between the 100 points provided for each video. Another limitation is that YouTube does not provide information on every link which led users to the videos, providing only information on the top ten referrers. In total, 64%, 75%, and 65% of views of the videos in the Top, YouTomb, and Random datasets, respectively, do not have the referrer registered.

# 4. POPULARITY GROWTH PATTERNS

In this section, we analyze the popularity growth patterns across our three video datasets, namely Top, YouTomb and Random. This analysis is based on two aspects, namely, (1) the time interval until a video reaches most of its current popularity (measured according to one of the three metrics), and (2) the bursts of popularity experienced by a video in short periods of times (e.g., days or weeks). Inspired by results in [9,10], we also categorize videos according to their temporal popularity evolution dynamics.

We focus our analyses on the number of views as the main popularity metric because: (1) previous analyses of YouTube have found a large correlation between total number of comments (or favorites) and total view count [8], and (2) we have computed the correlations for both pairs of metrics, taken at each point in time (instead of only for the final snapshot, as previously done), finding positive correlations, ranging from 0.18 to 0.24, for all datasets.

We note that, as shown in Table 2, the ages of the videos in each dataset vary significantly. Most videos in the YouTomb and Random datasets are either around a few weeks (under 1 month) old or over 1 year old. In contrast, most Top videos have shorter ages (under 1 year old). Given such variability, we have performed our popularity analyses separately for each age range, in each dataset. However, due to space constraints, we focus on results computed over all videos in each dataset, pointing out significant differences across age ranges when appropriate.

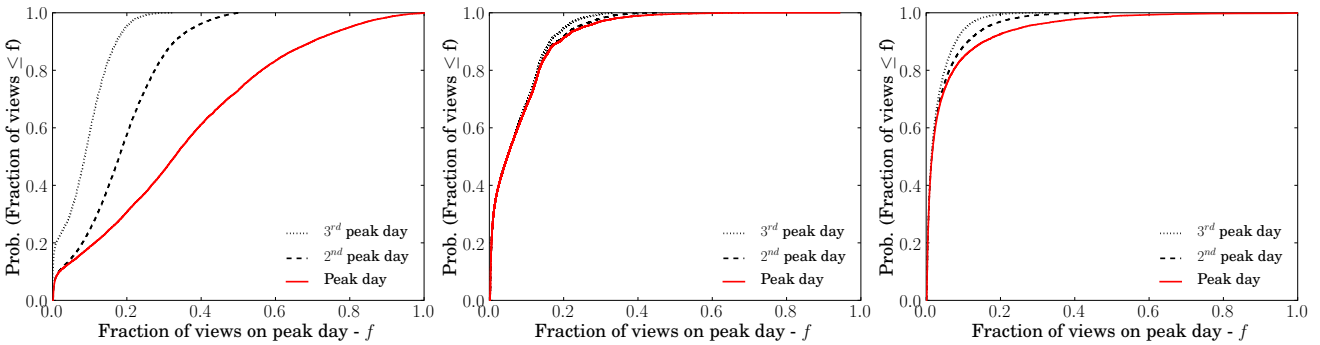
## 4.1 How Early Does a Video Get Popular?

We address this question by plotting, in Figure 2, the cumulative distributions of the amount of time it takes for a video to receive *at least 10%*, *at least 50%* and *at least 90%* of their total views, measured at the time our data was collected. Time is shown normalized by the age of the video, which is here referred to as the video's *lifetime*.

Figure 2 shows that, for half of the videos (y-axis) in the Top, YouTomb and Random datasets, it takes at most 65%, 21% and 87%, respectively, of their total lifetimes (x-axis) until they receive at least 90% of their total views. If we consider at least 50% of their total views, the fractions are 26%, 5% and 43%, respectively, following a similar trend. The same holds for the mark of 10% of the views.

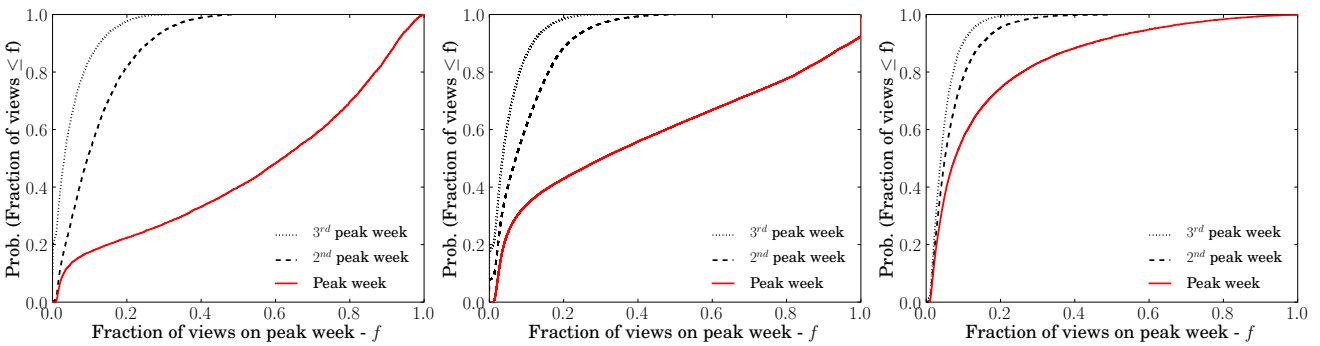
Conversely, around 31% of Top videos take at least 20% of their lifetimes to reach at least 10% of their final popularity. Similarly, around 18% of the Random videos also experience a similar dormant period before starting to receive most of their views. In contrast, the equivalent fraction among YouTomb videos is much shorter, around 10%.

Thus, comparing the results across datasets, YouTomb videos tend to get most of its views earlier in their lifetimes, followed by videos in Top and Random. As videos in the




(a) Top (b) YouTomb (c) Random

**Figure 3: Cumulative distributions of the fraction of total views on the first, second and third peak day.**




(a) Top (b) YouTomb (c) Random

**Figure 4: Cumulative distributions of the fraction of total views on the first, second and third peak week.**

**Table 3: Time until a video achieves at least 90% of its total views, across age ($a$) ranges (time normalized by video's lifetime, mean $\mu$, and standard deviation $\sigma$).**

| | Top | | YouTomb | | Random | |
|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| $a \leq 7$ days | .64 | .10 | .61 | .14 | .61 | .17 |
| 7 days $< a \leq 1$ month | .55 | .19 | .48 | .23 | .66 | .20 |
| 1 month $< a \leq 1$ year | .50 | .27 | .18 | .21 | .79 | .17 |
| $a > 1$ year | .77 | .23 | .31 | .23 | .85 | .11 |

top lists tend to be more popular, the difference between the results for Top and Random datasets are somewhat predictable. Possible reasons as to why YouTomb videos tend to receive most of their views even earlier are: (1) as many of these videos consist of popular TV shows and music trailers, a natural interest on this content closer to when it is uploaded is expected, and (2) being aware that such videos contain copyright protected content, users may seek them quicker after upload, before the violation is detected and they are removed from YouTube.

We note that since lifetime is a normalized metric, these results may be impacted by the distributions of video ages (Table 2). In particular, recall that such distribution is skewed towards older videos in the YouTomb dataset: around 79% of them have at least 1 year of age. This bias may influence the results. However, we also note that 54% of the videos in the Random dataset also fall into the same age range. Yet, in comparison with YouTomb videos, videos in the Random dataset get most of their views much later during their lifetimes.

Thus, to reduce any bias caused by age differences, we repeat our analyses separately for videos in each age range. Due to space constraints, we show, in Table 3, only results for the time until a video achieves at least 90% of its views. We show averages and standard deviations for each age range and dataset. The same aforementioned trend occurs for videos in most age ranges: YouTomb videos reach at least 90% of their views much earlier in their lifetimes Top videos, followed, somewhat behind, by Random videos. The only exception occurs for the youngest videos, for which there is no much difference across the datasets.

## 4.2 Do Videos Experience Popularity Bursts?

We now investigate the bursts of popularity experienced by the videos. We start by analyzing the cumulative distributions of the fraction of views each video receives on its first, second and third most popular (i.e., peak) days, shown in Figure 3.

Figure 3-a) shows that Top videos experience a very distinct (first) peak day: 50% of them receive between 33% and practically 100% of their views on a single (peak) day. In comparison, the fraction of videos receive between 17%

**Table 4: Fractions of *Memoryless* and *Unknown* videos.**

| Time | Top | | | YouTomb | | | Random | | |
|---|---|---|---|---|---|---|---|---|---|
| Granularity | **Total** | *Memoryless* | *Unknown* | **Total** | *Memoryless* | *Unknown* | **Total** | *Memoryless* | *Unknown* |
| Daily | **33%** | 20% | 13% | **98%** | 97% | 1% | **78%** | 77% | 1% |
| Weekly | **60%** | 0% | 60% | **14%** | 13% | 1% | **4%** | 0% | 4% |

and 50% of their views on the second peak day, and between 5% and 34% of their views on their third peak day. Thus, Top videos clearly experience a burst of popularity on a single peak day. This is in sharp contrast with videos in the YouTomb and Random datasets (Figures 3-b and c), for which the fractions of views on the three peak days tend to be more similar. In fact, in both datasets, the three curves are very close to each other and skewed towards smaller fractions of views. While these results might imply different popularity growth patterns, with most videos in Random and YouTomb exhibiting multiple (smaller) daily peaks, we should also note that the interpolation performed over the collected data (see Section 3.3) might introduce distortions in this analysis, particularly considering the large fraction of very old videos in these two datasets.

To cope with these possible distortions, we also analyze the distributions of the fraction of views on the first, second and third peak weeks, shown in Figure 4. Interestingly, we now observe that all three types of videos tend to experience some burst of popularity on a single week. Nevertheless, the general trend is similar to the one observed for daily peaks: the peak week tends to be more significant for Top videos. For instance, 60% of Top videos receive at least 50% of their views on their (first) peak week. In contrast, only 40% of YouTomb videos receive at least as many views on a single week. The peak week is even less significant for videos in the Random dataset, although, in comparison with daily peaks (Figure 3-c), it is more clearly distinguished from the other two peaks. Similar conclusions, for weekly and daily popularity peaks, hold for videos falling in different age ranges.

## 4.3 Temporal Dynamics

As discussed in Section 2, Crane and Sornette [9, 10] proposed epidemic models to understand how popularity growth patterns can be explained in terms of user interactions within the system and external events. They distinguish four different evolution patterns. For the vast majority of videos, popularity dynamics are quite stable, either experiencing little activity or being well described by a simple stochastic process (e.g., a Poisson process). We here refer to such videos as *Memoryless*. In contrast, some videos experience bursts of activity (i.e., popularity), and can be further categorized into: (1) *Viral videos*, which experience precursory word-of-mouth growth resulting from epidemic-like propagation through OSNs; (2) *Quality videos*, which experience a very sudden burst of popularity (due to some external event, such as being featured on the first page of YouTube); and, (3) *Junk videos*, which experience a burst of popularity for some reason (e.g., spam, chance, etc), but do not spread through the social network.

The authors also proposed a simpler intuition to categorize videos that experience bursts of popularity, which consists of grouping videos based on the fraction of views received on the most popular (i.e., peak) day. They found that the aforementioned evolution patterns and the number of views on the peak day are strongly correlated. Thus, more formally, the category of a video can be determined by:

$$\text{Viral} \longrightarrow views_{peak} \leq t \quad (1)$$

$$\text{Quality} \longrightarrow t < views_{peak} \leq (1-t) \quad (2)$$

$$\text{Junk} \longrightarrow views_{peak} > (1-t) \quad (3)$$

where $t$ is the fraction of views on the peak day. Based on empirical evaluation, the authors used $t$=20%

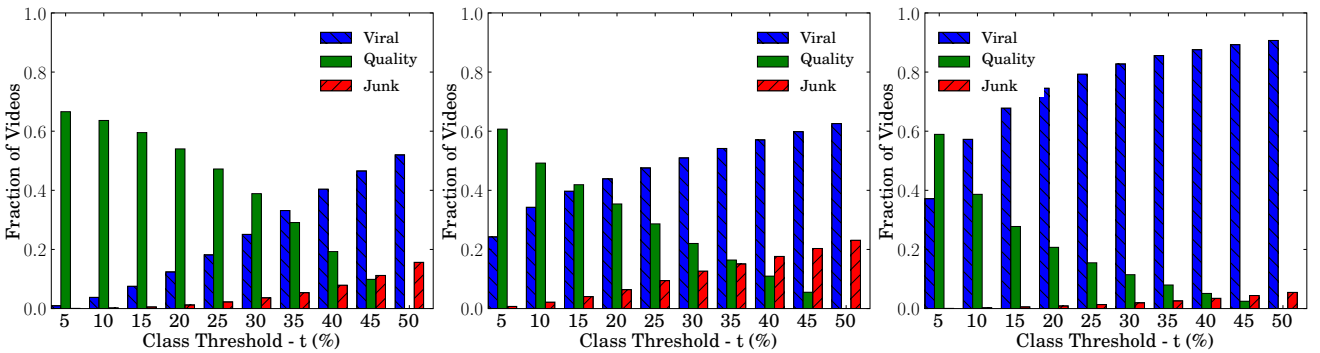
We here apply the ideas presented in [9, 10] to categorize the videos of our three datasets into the 4 classes: Memoryless, Viral, Quality and Junk. As in the previous section, to cope with possible spurious effects of the data interpolation performed, we analyze the evolutionary patterns both at daily and weekly time granularities. The analysis presented here extends the discussion of the previous section: whereas in Section 4.2 we analyzed popularity based on three different points in time, we here characterize its complete time series.

To identify videos falling into the Memoryless category, we used the Chi-Square test to determine whether the time series describing the popularity growth of each video (considering both time granularities) follows a Poisson process. Unfortunately, for several (recently uploaded) videos, the corresponding time series had very few points, subjecting the characterization to too much noise. We experimented with several thresholds for the minimum number of data points, selecting a threshold equal to 4, as larger values had little impact on the number of videos characterized as Memoryless.

Table 4 shows, for each dataset and time granularity, the fraction of videos characterized as Memoryless, the fraction of videos for which the number of points fell below the threshold, thus being characterized as *Unknown*, as well as the total fraction of videos in both groups. As expected, the fractions vary significantly depending on the time granularity used[10]. Note the large fraction of Top videos characterized as Unknown for the weekly based analysis, which is due to the large number of videos with age below 4 weeks.

Recall that, in Section 4.2, we concluded that videos in both Random and YouTomb datasets tend to have multiple (smaller) daily popularity peaks, whereas most Top videos exhibit a single more significant peak at both daily and weekly granularities. The results in Table 4, covering the whole time series, are supported by and extend those findings. Taken at the granularity of days, the vast majority of videos in YouTomb (97%) and Random (77%) experience a popularity growth that follows a Memoryless (Pois-

[10]We experimented with other time granularities ranging from 1 to 30 days, finding similar results for all granularities above 5 days.




(a) Top (Daily) (b) YouTomb (Weekly) (c) Random (Weekly)

**Figure 5: Fractions of *Quality*, *Viral* and *Junk* videos.**

son) process, with no distinct underlying growth pattern. As discussed in [10], the popularity evolution of such videos is largely driven by fluctuations and not bursts of activity. Note that no video in Random was characterized as Memoryless at the granularity of weeks, meaning that, in spite of the somewhat smaller gap between the curves shown in Figure 4-c), the complete time series does exhibit a distinct growth process.

Next, we use Equations (1), (2) and (3) to characterize the remaining videos into Quality, Viral and Junk, respectively. We experimented with various values for class threshold $t$. Figure 5 shows the fractions of videos in each category for various values of $t$. We report results for Top videos according to their daily time series, and for videos in the YouTomb and Random datasets according to their weekly time series. We here focus on such results because, for the other scenarios, the Memoryless and Unknown categories dominate in all three datasets (Table 4). Nevertheless, we note that, for any given dataset, similar qualitative results are obtained for the omitted time granularity. We also note that we did implement the full model and clustering analysis proposed in [10], obtaining similar qualitative results (also omitted).

As shown in Figure 5, most (non-Memoryless) videos in the Random and YouTomb datasets fall into the Viral category for thresholds that are not very restrictive ($t$>15%). For instance, for $t$=20%, the fractions of Viral videos in the Random and YouTomb datasets are 77% and 44%, respectively. The time series of such videos show a slight increase in the number of views up to a peak day, representing an endogenous (word-of-mouth) growth. After the peak, the videos propagate virally through the OSN.

In contrast, the Top dataset, analyzed at the granularity of days, is dominated by Quality videos, except for thresholds that are very restrictive for this category ($t\geq$35%). Indeed, for $t$=20%, 54% of the Top videos are characterized as Quality. Such videos experience a sudden burst of popularity but remain attractive for some time afterwards. For most threshold values, Quality is also the second most frequent category in both Random and YouTomb datasets, whereas Viral is the second most frequent category among Top videos. We note that a small but non-negligible fraction of videos in each dataset are characterized as Junk, particularly for larger threshold values. Such videos, in spite of the sudden popularity burst, do not remain popular for very long. We note that the fraction of Junk videos is much smaller, even for larger thresholds, in the Random dataset, possibly due to the much less significant popularity peaks experienced by these videos (Section 4.2).

We finish this section by noting that the large fraction of videos that could not be characterized (i.e., *Unknown*) because of their small age in the system motivates the design of new models for popularity growth. Such models may exploit, for instance, the results presented in Section 4.1 to estimate how long video popularity stays dormant. In the next section we extend our analysis, focused so far only on temporal data, to investigate how users reach the videos.

# 5. REFERRER ANALYSIS

A number of studies have analyzed the dynamics of word-of-mouth-like information propagation through friends in social networks [4, 7, 13]. However, on YouTube, as well as on several other OSNs, word-of-mouth is not the only mechanism through which information is disseminated. We next address this issue by examining the main referrers that lead users to videos.

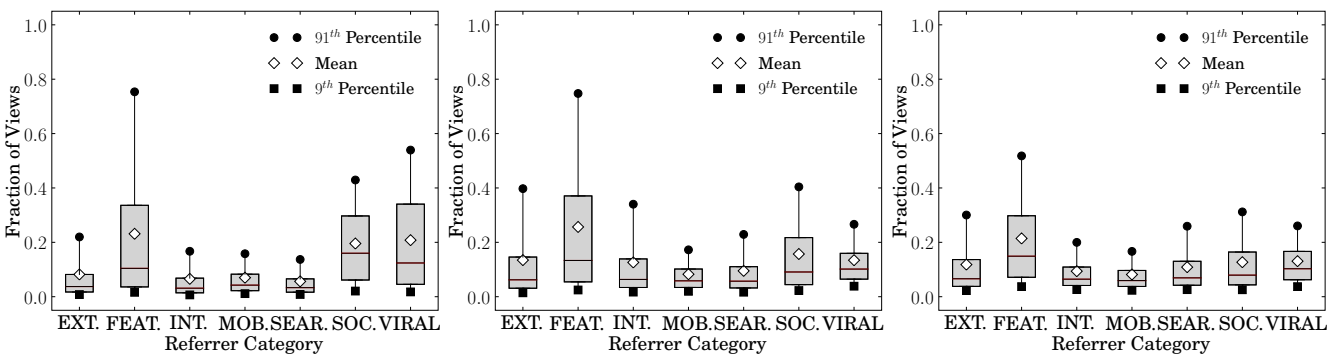
## 5.1 Referrer's Importance

As a first step, we identified 14 types of referrers that appear in our datasets, grouping them into the following categories: External, Featured, Search, Internal, Mobile, Social, and Viral. The *External* category represents websites (often other OSNs and blogs) that have links to YouTube videos. The *Featured* category contains referrers that come from advertises about the video in other YouTube pages or featured videos on top lists and on the front page. On the *Search* category, we group all the referrers from search engines, which comprise only Google services. *Internal* referrers correspond to other YouTube mechanisms, such as the "Related Video" feature, which displays a list of 20 videos that are considered related (according to a YouTube proprietary algorithm) to the watched video. *Mobile* corresponds to all video accesses that come from mobile devices. *Social* referrers consist of accesses coming from the page of the video owner (the channel page) or from users who subscribed to the owner or to some specific topic. Finally, YouTube groups referrers from emails and other sources into a single category, named *Viral*[11].

[11]This type of referrer is not associated with the *Viral* popularity growth model presented in Section 4.3.

| Category | Referrer Type | Top | | | YouTomb | | | Random | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $t_{view}$ | $f_{view}$ | $f_{time}$ | $t_{view}$ | $f_{view}$ | $f_{time}$ | $t_{view}$ | $f_{view}$ | $f_{time}$ |
| EXTERNAL | First embedded view<br>First embedded on<br>First referrer from | 0.57 | 0.11 | 0.35 | 0.81 | 0.16 | 0.41 | 0.07 | 0.08 | 0.22 |
| FEATURED | First view from ad<br>First featured video view | 0.72 | 0.14 | 0.03 | 0.10 | 0.02 | 0.00 | 0.11 | 0.14 | 0.00 |
| INTERNAL | First referrer from YouTube<br>First referrer from Related Video | 1.50 | 0.29 | 0.67 | 1.85 | 0.36 | 0.65 | 0.14 | 0.18 | 0.34 |
| MOBILE | First view from a mobile device | 0.26 | 0.05 | 0.51 | 0.02 | 0.00 | 0.02 | 0.03 | 0.03 | 0.05 |
| SEARCH | First referrer from Google<br>First referrer from YouTube search<br>First referrer from Google Video | 1.05 | 0.20 | 0.36 | 1.80 | 0.35 | 0.52 | 0.29 | 0.37 | 0.41 |
| SOCIAL | First referrer from a subscriber<br>First view on a channel page | 0.36 | 0.07 | 0.35 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.12 |
| VIRAL | Other / Viral | 0.81 | 0.16 | 0.79 | 0.59 | 0.12 | 0.62 | 0.16 | 0.20 | 0.55 |

**Table 5: Referrer categories and statistics ($t_{view}$: number of views (x $10^9$); $f_{view}$: the fraction of views; $f_{time}$: fraction of times a referrer from the given category was the first referrer of a video).**




(a) Top (b) YouTomb (c) Random

**Figure 6: Distribution of the fraction of views for which each referrer category is responsible.**

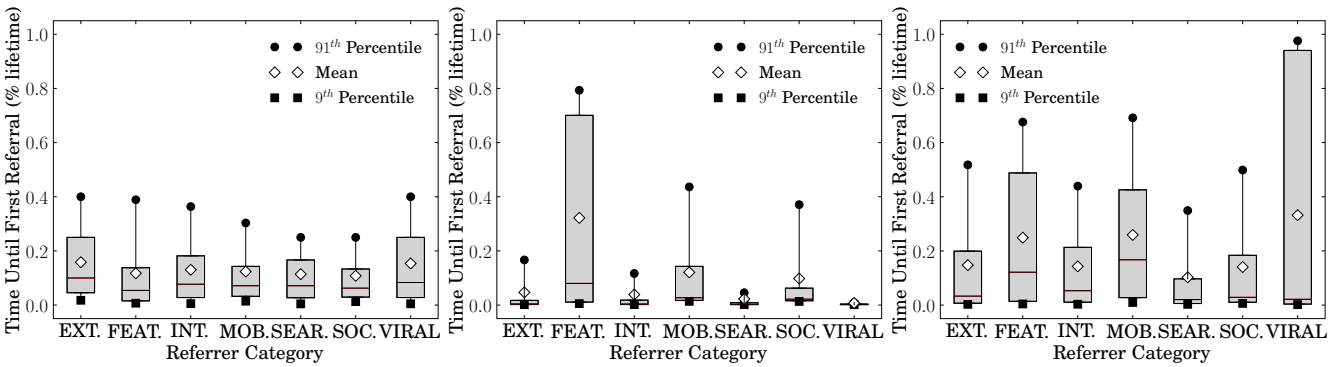
Table 5 displays the list of the 14 types of referrers and 7 referrer categories, showing the number and fraction of views for which each category is responsible. These fractions as well as the following analyses are based on the video accesses from the top ten referrers that we have access to (see Section 3.3). Note that, as shown in Table 5, these ten referrers are responsible for millions of views.

Table 5 shows that search and internal YouTube mechanisms are key mechanisms through which users reach content on YouTube. Interestingly, in a very recent work, Oliveira *et al.* [16] presented the following hypothesis: *video search is the main method for reaching content on video sharing websites.* They verified this through online questionnaires using a large number of volunteers. Whereas our results confirm their hypothesis for videos in the Random dataset, they show YouTube internal features such as "Related Videos" play an even more important role to content dissemination for Top videos. For YouTomb videos, both categories attract roughly the same number of views. We note that YouTube search is responsible for the vast majority of the Search referrers, as less than 1% of the Search accesses come from other Google search mechanisms. Comparing the importance of Search referrers across datasets, we note that search is more important to Random and YouTomb videos, as they are not systematically exposed to users as videos from top lists are. We also note the importance of the Viral referrer category in all three datasets, particularly Random.

We further analyze the importance of each referrer category, by computing the distributions of the number of views for which each referrer category is responsible, considering only videos that received accesses from the given category. Figures 6(a-c) show box plots containing first, second and third quartiles, along with the $9^{th}$ and $91^{th}$ percentiles, and the mean, for each referrer category and each video dataset [12]. Unlike Table 5, which provides aggregated measures for each dataset, these plots allow us to assess the importance of each referrer category for individual videos.

For example, Table 5 shows that Social referrers do not appear to be important for YouTomb dataset as a whole. Nevertheless, considering only copyright protected videos with at least one Social referrer, Figure 6-b) shows that more than 22% of the views come from subscription links for 25% of such videos ($3^{rd}$ distribution quartile). This indicates that users may subscribe to other users that post copyright protected content. The Featured category is a similar case.

[12]For any given referrer category, at least 1000 videos received views for which it is responsible. Thus, these distributions are computed over at least as many videos.




(a) Top (b) YouTomb (c) Random

**Figure 7: Distributions of time spent until the first access from a referrer category (time is normalized by video's lifetime).**

Moreover, we note that the Social, Featured and Viral categories are responsible for more than 30%, 33% and 34%, respectively, of the views for 25% of the Top videos with referrers from each category (Figure 6-a). Finally, according to Figure 6-c), the Featured category plays a dominant role as source of views to videos in the Random dataset: 25% of the videos that received at least on Featured referrer received at least 30% of their views from such referrers.

We note that it is hard to tell whether one referrer might influence the number of views from other referrers. For example, a Top video may experience a popularity growth from Social and Viral referrers *after* being featured in the top list. Conversely, it may first receive a large number of views from Social and Viral referrers, which ultimately leads it to be featured in the top list. Similarly, Viral accesses may greatly contribute for a video to enter a top list; alternatively, videos in a top list may spread much more quickly disseminated via emails. In the following, we study this issue by analyzing how early in a video's lifetime each type of referrer is used.

## 5.2 Referrer's First Appearance

We start by analyzing the referrers that first lead users to a video. The $f_{time}$ columns in Table 5 show the fractions of videos that had the first referrer falling into each given category. We note that, since YouTube provides only the day each referrer was first used and several referrers may appear on a single day, there might be ties, i.e., multiple categories may be listed as containing the first referrer of a video. Thus, the sum of the $f_{time}$ column can exceed 100%.

The first referrer for 79%, 67%, and 51% of the Top videos are from the Viral, Internal, and Mobile categories, respectively. For the YouTomb dataset, Internal, Viral, and Search are the top three categories, containing the first referrers for 65%, 62% and 52% of the videos, respectively. For the Random dataset, the first referrer of 55%, 41%, and 34% of the videos are from the Viral, Search, and Internal categories, respectively. Thus, in general, viral spreading and internal YouTube mechanisms appear as primary forms through which users reach the content for the first time, in all three datasets. Interestingly, mobile devices are also a relevant front door to Top videos, whereas for YouTomb and Random videos, the YouTube search engine accounts for a large fraction of the first referrers.

Figures 7(a-c) show the distributions of the difference between the time of the first referrer access and the time the video was uploaded, measured as a fraction of the video's lifetime. For the Top and YouTomb datasets, referrers (of any category) tend to happen very early: for 75% of the Top and YouTomb videos, most referrer categories have their first appearances during the first quarter of the video's lifetime. In fact, only 9% of the Top videos have their first referrer access (of any category) after 40% of their lifetimes. The exception is the Featured category on YouTomb: those referrers tend to take somewhat longer to appear for the first time. For instance, for 25% of the YouTomb videos, they appear only after 70% of the video's lifetime. This was somewhat expected as YouTube would most likely not feature videos that are suspicious to be copyright protected.

For Random videos, in general, Search, Internal, External, and Social referrers tend to appear earlier than referrers from the other categories. Thus, users are more likely to initially find such videos through social links, search, other YouTube internal mechanisms or some external website, instead of receiving them via e-mail or viewing them on mobile devices.

# 6. CONCLUSIONS AND FUTURE WORK

We characterized the growth patterns of video popularity on the currently most popular video sharing application, YouTube. Using newly provided data by the application, we analyzed how the popularity of individual videos evolve since the video was uploaded, as well as the different types of referrers that most often lead users to the videos.

Comparing the three analyzed datasets, copyright protected (YouTomb) videos tend to get most of their views much earlier in their lifetimes, followed by Top videos and, somewhat behind, videos in the Random dataset. We also found that Top videos tend to experience significant bursts of popularity, receiving a large fraction of their views on a single peak day (or week). As a matter of fact, most (characterized) Top videos have popularity growth patterns falling into the *Quality* category, that is, they experience a sudden burst of popularity remaining attractive for a while. In contrast, videos in the YouTomb and Random datasets tend to experience multiple smaller popularity peaks. Indeed, if popularity growth is analyzed on a weekly basis, most videos

in both datasets fall into the *Viral* category, with a popularity growth following an endogenous word-of-mouth process.

We also identified and quantified the main referrers that led users to videos in each dataset. Particularly, we showed that search and internal YouTube mechanisms, such as lists of related videos, are key mechanisms to attract users to the videos. Whereas search referrers account for the largest fraction of views to videos in the Random dataset, internal YouTube mechanisms play an even more important role to content dissemination for Top and YouTomb datasets.

Our analyses uncover various interesting findings, leading to several possible directions for future work. One such direction is to leverage our findings to build mechanisms for predicting content popularity. Predicting which newly uploaded content will become popular can help companies to maximize revenue through advertise placement tools, and can also help consumers filtering the ever-growing amount of available content. In a system like YouTube, popularity prediction is a huge challenge as it results from the combination of a multitude of factors including complex interactions among users, aspects related to content quality and external events. Such factors are, at least partially, captured by the referrers that are used to reach the content. Indeed, these referrers provide evidence of how the video is being disseminated. Thus, referrer information, along with the popularity growth patterns characterized here, might serve as valuable data sources for predicting content popularity.

Another possible direction is to investigate how referrer and popularity growth patterns can be exploited to improve the effectiveness of content recommendation and search tools. This is particularly interesting given that we found that search and internal YouTube mechanisms are the two most important sources of "hits" for video traffic.

## ACKNOWLEDGEMENTS

This research is partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/INCT Web Grant Number 573871/2008-6), and by the authors' individual grants from CNPq, CAPES and Fapemig. We also thank Elizeu Santos Neto for initial discussions, and members of the YouTomb discussion list.

## 7. REFERENCES

[1] comscore:americans viewed 12 billion videos online in may 2008. `http://goo.gl/2bKmP`. Accessed Nov/2010.

[2] New york times. uploading the avantgarde. `http://goo.gl/S72M8`. Accessed in Nov/2010.

[3] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the World Wide Web Conference (WWW)*, 2007.

[4] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2008.

[5] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)*, 5(4):1–25, 2009.

[6] M. Cha, H. Kwak, P. Rodriguez, Y-Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Network (TON)*, 17(5):1357–1370, 2009.

[7] M. Cha, A. Mislove, and K. Gummadi. A measurement-driven analysis of information propagation in the Flickr social network. In *Proceedings of the World Wide Web Conference (WWW)*, 2009.

[8] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A first step towards understanding popularity on youtube. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2010.

[9] R. Crane and D. Sornette. Quality, and junk videos on YouTube: Separating content from noise in an information-rich environment. In *Proceedings of the AAAI Spring Symposium*, 2008.

[10] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences (PNAS)*, 105(41):15649–15653, 2008.

[11] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from the edge. In *Proceedings of the ACM Internet Measurement Conference (IMC)*, 2007.

[12] M. Gonçalves, J. Almeida, L. Santos, A. Laender, and V. Almeida. On popularity in the blogosphere. *IEEE Internet Computing*, 14:30–37, 2010.

[13] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2003.

[14] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the World Wide Web Conference (WWW)*, 2010.

[15] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the World Wide Web Conference*, 2008.

[16] R. Oliveira, M. Cherubini, and N. Oliver. Looking at near-duplicate videos from a human-centric perspective. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)*, 6(3):1–22, 2010.

[17] J. Ratkiewicz, A. Flammini, and F. Menczer. Traffic in social media I: paths through information networks. In *Proc. of the Int'l Symposium on Social Intelligence and Networking*, 2010.

[18] F. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proceedings of the World Wide Web Conference (WWW)*, 2007.

[19] G. Szabo and B. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[20] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni. Research on online social networks: Time to face the real challenges. *SIGMETRICS Performance Evaluation Review*, 37(3):49–54, 2009.

[21] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: YouTube network traces at a campus network - measurements and implications. In *Proceedings of the IEEE Multimedia Computing and Networking (MMCN)*, 2008.