Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems

Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon

***Abstract*—User generated content (UGC), now with millions of video producers and consumers, is re-shaping the way people watch video and TV. In particular, UGC sites are creating new viewing patterns and social interactions, empowering users to be more creative, and generating new business opportunities. Compared to traditional video-on-demand (VoD) systems, UGC services allow users to request videos from a potentially unlimited selection in an asynchronous fashion. To better understand the impact of UGC services, we have analyzed the world's largest UGC VoD system, YouTube, and a popular similar system in Korea, Daum Videos. In this paper, we first empirically show how UGC services are fundamentally different from traditional VoD services. We then analyze the intrinsic statistical properties of UGC popularity distributions and discuss opportunities to leverage the latent demand for niche videos (or the so-called "the Long Tail" potential), which is not reached today due to information filtering or other system scarcity distortions. Based on traces collected across multiple days, we study the popularity lifetime of UGC videos and the relationship between requests and video age. Finally, we measure the level of content aliasing and illegal content in the system and show the problems aliasing creates in ranking the video popularity accurately. The results presented in this paper are crucial to understanding UGC VoD systems and may have major commercial and technical implications for site administrators and content owners.**

***Index Terms*—Interactive TV, human factors, exponential distributions, log normal distributions, pareto distributions, probability, copyright protection.**

I. Introduction

VIDEO content in standard video-on-demand (VoD) systems has historically been created and supplied by a limited number of media producers such as licensed broadcasters and production companies. The advent of user-generated content (UGC) has re-shaped the online video market enormously. Nowadays, hundreds of millions of Internet users are not only content consumers, but also publishers. The length of UGC videos is shortened by two orders of magnitude than traditional videos and so is the production time. Wired magazine refers to this small-sized content pop culture as "bite-size bits for high-speed munching" [1].

The scale, dynamics, and decentralization of UGC videos make their content popularity more ephemeral and unpredictable. As opposed to the early days of TV when everyone watched the same program at the same time—for instance, the biggest hit of 1953, "I Love Lucy", was watched by 70% of TV households—such strong reinforcement of popularity (or unpopularity) is absent in UGC. Unlimited choice of content and the convenience of the Web have quickly personalized the viewing experience, leading to a great variability in user behavior and attention span. Understanding the popularity characteristics of UGC is important because they can be used to estimate the latent demand that may exist due to bottlenecks in the system (e.g., poor search and recommendation engines, missing metadata). Bottlenecks greatly affect the strategies for marketing, target advertising, recommendation, and search. At the same time, a lack of editorial control in UGC is creating problems for content aliasing and copyright infringement, which seriously threatens the future viability of such systems.

To understand the nature and the impact of UGC systems, we analyzed YouTube, the world's largest UGC VoD system, and Daum Videos, a popular UGC service in Korea. The main contribution of this paper is an extensive trace-driven analysis of UGC video popularity distributions. For this, we have collected information about millions of videos from YouTube and Daum websites, which we share for the wider community to use.[1] Our analysis reveals very interesting properties about how users of these systems request UGC videos. Based on a static snapshot of video view counts, we investigate whether video popularity can be modeled as a power-law and what characteristics of the system influence the shape of the distribution. Based on video views observed over multiple consecutive days, we examine non-stationary properties of the UGC video popularity. Our analysis further reveals the level of piracy and content duplication, which has major implications in the deployment of future UGC services.

The highlights of our work are summarized as follows:

1) We outline the high-level characteristics of UGC systems by comparing them with standard VoD systems. We find that the two systems show stark differences in their content production and consumption patterns.
2) We analyze the popularity distributions of UGC videos. We find that video popularity follows a power-law distribution

[1]Datasets are made available at http://an.kaist.ac.kr/traces/IMC2007.html

Manuscript received December 23, 2007; revised July 03, 2008; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor P. Barford. First published March 16, 2009; current version published October 14, 2009. M. Cha did this work as an intern at Telefonica Research, Barcelona. The work of H. Kwak and S. Moon was supported by Grant A1100-0801-2758 from the IT R&D program of MKE/IITA in Korea.

M. Cha is with Max Planck Institute for Software Systems (MPI-SWS), Networked Systems Research Group, Saarbruecken, Saarland D-66123, Germany (e-mail: meeyoung.cha@gmail.com).

H. Kwak and S. Moon are with the Computer Science Department, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305-701 Korea.

P. Rodriguez is with the Telefonica Research Lab, Internet and Multimedia, Barcelona 08021, Spain.

Y.-Y. Ahn is with the Center for Complex Network Research, Northeastern University, Boston, MA 02108 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2008.2011358

1063-6692/$26.00 © 2009 IEEE

TABLE I
SUMMARY OF USER-GENERATED VIDEO TRACES

Name	Category	Num. videos	Total views	Total length	Data collection period
YouTube	Ent	1,687,506	3,708,600,000	15.2 years	December 28, 2006 (crawled once)
YouTube	Sci	252,255	539,868,316	1.8 years	January 14-19, 2007 (daily), February 14, March 15, 2007 (once)
Daum	All	196,037	207,555,622	1.0 year	March 1, 2007 (once)
YouTube	Pop*	2,091	avg. 31,689	med. 186 sec	January 13 - February 5, 2007 (daily)

*For globally popular videos in YouTube, we show the average number of views and the median length of videos.

with an exponential cutoff. We discuss several mechanisms that generate such a distribution. Assuming the underlying distribution is Zipf, we show that 45% more views can be obtained by removing bottlenecks.

3) We study the evolution of video popularity over time. We investigate the relationship between video age and request rate and measure the ephemeral lifetime of the most popular videos. We demonstrate that popularity is mostly determined at the early stage of video age.
4) We estimate the prevalence of content duplication and find that the total view counts from multiple copies of a single video can grow more than two orders of magnitude of the original video. Our findings indicate that content duplication can hamper the system's efficiency, especially in ranking video popularity accurately.

The rest of the paper is organized as follows. We describe our data collection methodology and datasets in Section II. We conduct an empirical comparison between UGC and non-UGC in Section III. Section IV presents our analysis of the popularity distribution of UGC and the forces that shape it. Section V investigates how popularity of videos evolves over time. Section VI focuses on the level of content duplication and illegal uploads in UGC. Finally, we discuss related work in Section VII and conclude in Section VIII.

II. MEASUREMENT METHODOLOGY

In this section, we describe our data collection methodology and introduce our datasets. To examine how UGC is different from non-UGC, we also introduce datasets of professionally generated content that we use later for comparison.

A. UGC Videos

Our dataset consists of metadata of videos from YouTube and Daum services. Table I summarizes the basic statistics.

YouTube, launched in February 2005, is credited for jump-starting the UGC boom [2]. YouTube serves over 100 million distinct videos daily, growing with over 65,000 new uploads per day. YouTube provides an easy-to-use platform for uploading and sharing. Users may watch videos without logging in to YouTube and no additional client programs are required for viewing. YouTube streams videos through the Adobe Flash Player plug-in in web browsers, which is available on 90% of Internet-connected computers [3]. To upload, rate, or comment on a video clip, users must log in.

We crawled the YouTube website and collected meta information about all the videos in two of its categories: "Entertainment" and "Science & Technology" (now called "Howto" and "Style"). Throughout this paper, we refer to them as Ent and Sci. To get the complete list of videos, we exploited the indexed URL structure of YouTube, which was available at the time of this study. To examine the request patterns over time, we crawled the entire Sci category for six consecutive days. To understand the characteristics of globally popular videos, we monitored YouTube's list of the 100 daily most popular videos (from all categories), collecting 2,091 unique videos over 24 days. We refer to this dataset as Pop.

Daum Videos, launched in late 2006, is the most popular video sharing service in Korea and serves two million visitors and 35 million views weekly [4]. Unlike the stand-alone YouTube service, Daum Videos is an add-on service of the main Daum portal site—a major provider of e-mail, blog, and search services in Korea. Like YouTube, Daum also streams videos via Adobe's Flash Player. However, Daum uses a codec that allows users to upload higher-quality videos (streaming at 800 kb/s). We used the indexed URL structure and crawled video information from all of its 18 categories, and recorded detailed video information.

We implemented a Python web crawler to access YouTube and Daum video pages and parsed their HTML codes using Beautiful Soup HTML/XML parser. By specifying the category ID and the page counter in the indexed URL, we repeatedly accessed YouTube and Daum web pages showing multiple thumbnails of videos and obtained the complete set of videos belonging to a category. Each video record contained static information (e.g., the uploader, the upload time, the length) and dynamic information (e.g., ratings, links). *Views* and *ratings* indicate the number of times the video has been played or evaluated by users. *Links* indicate the list of external web pages hyper-linking the video. Each video's views, ratings, and links are publicly visible in YouTube.

Our traces do not contain information about individual user requests. However, our focus is on video popularity evolution, aggregated request distribution, and other statistics that do not require detailed knowledge of an individual user's behavior.

B. Non-UGC Videos

To examine the characteristics of UGC, we compare UGC with professionally generated content. While there are numerous real-world VoD systems, there exist only a handful of publicly available large-scale analyses. One such analysis is the extensive study on the PowerInfo system [5], a major video streaming service in China. The study is based on server logs of its first year of service from 2004 to 2005, covering over 6,700 movies and TV series. We also cite statistics from **IMDb** [6], the largest online movie database, and use real traces collected from other movie databases:

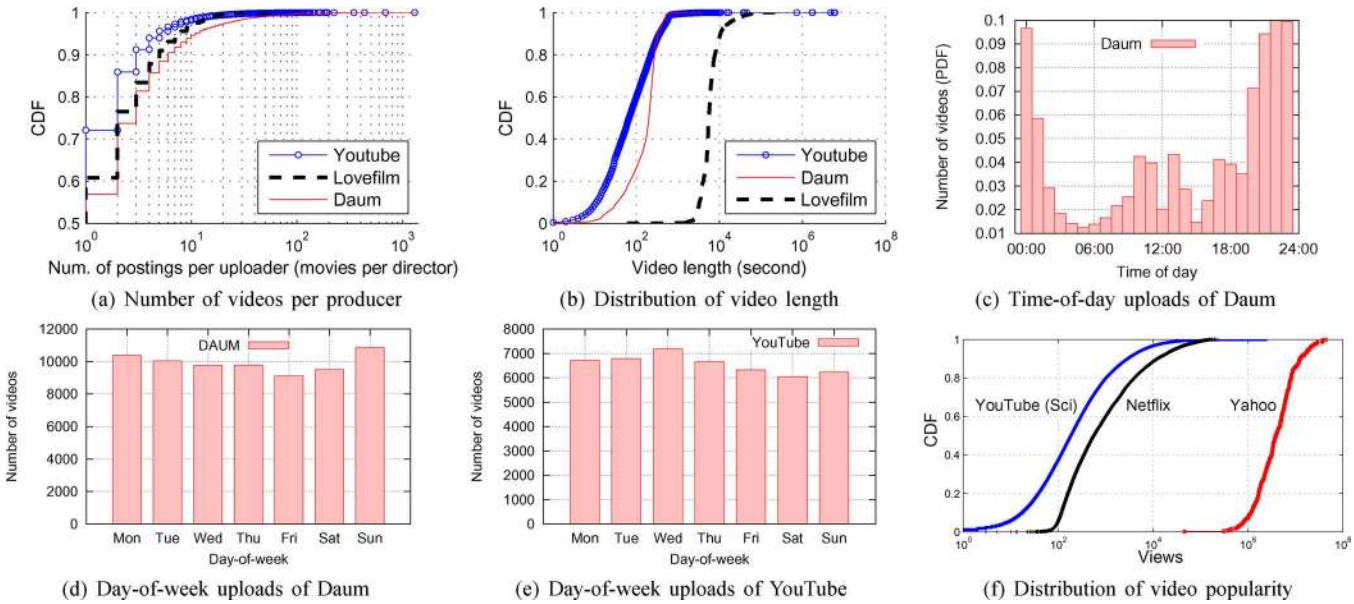


Fig. 1. Characteristics of UGC video uploads and comparison against non-UGC videos: content production and consumption patterns.

TABLE II
SUMMARY OF NON-UGC TRACES

Name	# Videos	Period	Description
Netflix	17,770	Oct 2006	Customer ratings
Lovefilm	39,447	Jan 2007	Length and director
Yahoo! Movies	361	2004–2007	Theater gross income

Netflix, a popular online video rental store, has made customer ratings for their 17,770 videos publicly available [7].

Lovefilm is Europe's largest online DVD rental store, carrying 60,000 movie titles [8]. For crawling, we started from a random set of movies and repeatedly visited all the movies following the links to "director" and "starring". We collected the length and the director information of 39,447 titles.

Yahoo! Movies provides the daily top 10 Box Office Chart in the United States, from 2004 to 2007, and the theater gross of each film on the list [9].

Table II summarizes our non-UGC traces.

III. UGC VERSUS NON-UGC

We describe the high-level characteristics of UGC services by contrasting them with traditional VoD (non-UGC) services. We compare the two systems in terms of content production and consumption patterns. We use traces from YouTube Sci and Daum Movies categories as representative UGC datasets.

A. Content Production Patterns

One key characteristic of UGC is the fast content production rate. As of June 9th, 2008, the largest online movie database IMDb carries 1,039,447 movies and TV episodes that were produced during the past 120 years.[2] In contrast, YouTube has 65,000 daily new uploads. This means that it only takes 15 days in YouTube to produce the same number of videos as are listed in IMDb.

1) Content Producers: To compare the production rate, we plot the cumulative distribution function (CDF) of the number of videos (or movies) posted per uploader (or director) in Fig. 1(a). UGC requires less production effort and accordingly has many distinct publishers. The average number of posts per publisher, however, is similar for UGC and non-UGC. 90% of film directors publish fewer than 10 movies, based on Lovefilm. Similarly, 90% of UGC publishers upload fewer than 30 videos in YouTube. The difference in production rate becomes evident when we focus on heavy producers. In UGC, there are extremely active publishers, who post over 1,000 videos over a few years. In contrast, the largest number of movies produced by a single film director only reaches 100 movies over half a century.

2) Video Length: Fig. 1(b) shows the distribution of video length for Daum, YouTube, and Lovefilm. UGC videos are shorter than non-UGC by two orders of magnitude. The length of UGC videos varies across categories. We have checked this for Daum, for which we have information across all its categories. The median video length ranged from 30 seconds (for advertisements) to 203 seconds (for music videos). 99% of videos are under 10 minutes in YouTube. Short video length may be due to the fact that UGC sites often cap video lengths. Both YouTube and Daum had a 100 MB file size limit at the time of this study. Daum increased its limit to 500 MB in early 2008. In contrast, the median movie length in Lovefilm is 94 minutes. Some traditional VoD systems also carry medium length videos such as 30 to 60 minute long TV series [5].

On a separate note, we did not see a correlation between the video popularity (e.g., view counts) and video length. The Pearson correlation coefficients between the two distributions were significantly small: -0.0001 for Daum and 0.0190 for YouTube. The correlation coefficients for the top 100 and the top 1,000 videos in YouTube were -0.0443 and -0.1452. In fact, we found that very short videos (e.g., 3 seconds) also appear in the list of 100 most popular YouTube videos.

3) Content Uploading: Fig. 1(c) shows the distribution of the number of new videos uploaded by the hour in Daum. While videos are uploaded throughout the day, 50% of total uploads are concentrated between 8 PM and 2 AM. This is in sharp contrast

[2]IMDb statistics are provided at http://www.imdb.com/database_statistics The oldest film listed in IMDb is "Roundhay Garden Scene" shot by Louise Le Prince in 1888.

to the peak usage hours of business applications. We are not able to conduct the same analysis for YouTube, due to lack of information on the exact upload time and geographical location of uploaders.

Fig. 1(d) and (e) shows the day-of-week upload patterns for Daum and YouTube. The vertical axis represents the total number of uploads for all videos in Daum and for recent five weeks for YouTube. New videos are uploaded relatively evenly throughout the week. A similar pattern has been reported in [5]. We see a subtle increase in uploads from Monday through Wednesday for YouTube, and from Sunday through Tuesday for Daum. Peak upload days are Sunday for Daum and Wednesday for YouTube. Monday shows moderately heavy uploads for both services. We reason that cultural differences may cause Daum uploaders to be more active on Sundays, while making it an off-peak day for YouTube users. Also, high broadband penetration in Korea may have facilitated sharing videos as a major pastime activity.

B. Content Consumption Patterns

To examine the different patterns of how content is consumed by users, we compare the popularity distributions of UGC and traditional VoD services. For UGC services we also analyze user participation and content discovery patterns. No comparable data was available for non-UGC services.

1) Scale of Popularity: Fig. 1(f) shows the CDF of video popularity based on view counts for YouTube, Netflix, and Yahoo! Movies. We consider the Sci video category for YouTube. For Netflix, we do not have information about views, so we use customer ratings instead. We expect the actual number of rentals to be significantly larger than the number of ratings. The plot on Netflix is, therefore, a lower bound on the number of rentals per movie. Finally, for Yahoo! Movies, we inferred the number of viewers per movie based on the reported box office earnings [9]. We divided the box office earning by an approximate price per ticket of $10. Note that Yahoo! Movies only contains extremely popular movies. Consumers in Yahoo! Movies and Netflix are regionally limited to the United States, while YouTube is used internationally.

From this popularity data, we make several observations. First, YouTube has 1,782 videos in the Sci category that had zero views, while all the movies in Netflix and Yahoo! Movies have been watched by at least one viewer. Second, the median number of views for YouTube (182) is much smaller than those of Netflix (561) and Yahoo! Movies (3,843,300), indicating that there are many unpopular videos in UGC. Finally, the scale of consumers per video is very different for UGC and non-UGC. The views distribution of YouTube spans more than 6 orders of magnitude, while the number of ratings per movie in Netflix and Yahoo! Movies span about 4 orders of magnitude. This illustrates the innate diversity in UGC producer and consumer population.

2) User Participation: The video popularity and ratings (i.e., the number of viewers who evaluated the video) show a strong positive linear relationship for both UGC and non-UGC, with the correlation coefficient of 0.8 for YouTube and 0.87 for Yahoo! Movies. This indicates that users are not biased towards rating popular videos more often than unpopular ones. Despite many features of YouTube that encourage user participation, the level of active user participation is still very low. While 54% of all videos are rated, the aggregate ratings only account for 0.22% of the total views. Comments, a more active form of participation, account for mere 0.16% of the total views. Other Web 2.0 sites have also reported relatively low user involvement [10].

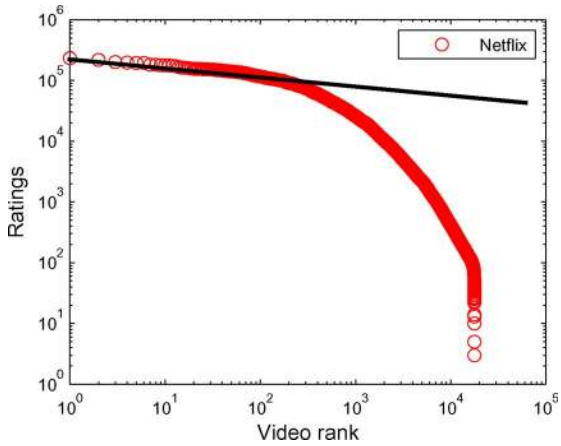


Fig. 2. Empirical plot of Netflix movie popularity based on customer ratings (denoted by circles) and a synthetic power-law distribution fitted for movies ranked between 1 and 100 (denoted by a straight line).

3) How Content is Found: We examine the external web pages that embed YouTube videos. Based on the Sci trace, 47% of all videos have incoming links from external sites. The aggregate views of these linked videos account for 90% of the total views, indicating that popular videos are more likely to be linked. Nevertheless, the total clicks derived from these links account for only 3% of the total views, indicating that views coming from external links are not significant. We have identified the top five websites linking to videos in YouTube Sci: myspace.com, blogspot.com, orkut.com, qooqle.jp, and friendster.com—four of them are social networking sites and one is a video recommendation site.

IV. Popularity Distribution

In this section, we examine static snapshots of video view counts to investigate statistical properties of video popularity. We analyze the shape of the video popularity distribution, and determine which probability distribution best fits the data. Analyzing the exact form of the probability distribution helps us understand the underlying mechanism that generated this distribution [11], and also helps us answer important design questions for UGC services. Analysis of probability distributions has proven fruitful in other domains. For instance, the scale-free nature of Web requests has been used to improve search engines and advertising policies [12]. The distribution of book sales has also been used to design better online stores and recommendation engines [13].

Normally, the shape of a distribution reflects the underlying mechanism that generates it and a distinguishing feature, in case of the power-law distribution, is a straight line in the log-log plot of views versus frequency. Although the power-law is commonly used to explain the frequency or the popularity distributions observed in the real world, it is a nontrivial task to

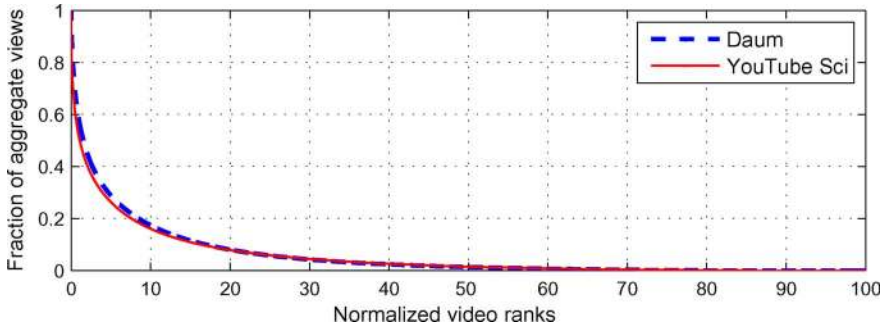


Fig. 3. Testing if Pareto Principle applies to UGC video popularity. For both YouTube and Daum, 10% of the most popular videos account for nearly 80% of views, while the remaining 90% account for total 20% of views.

determine whether a certain distribution is power-law or not [14]–[18]. This is because many other distributions exhibit a similar shape. For example, the log-normal distribution likewise has a straight line shape in the waist part of the distribution, followed by a curved tail. Further, the processes yielding the log-normal and the power-law may be very similar as discussed in [14].

Even when the true distribution is power-law, there may be other factors that distort the shape, particularly at the two ends of the distribution: the most popular and the least popular items. For example, the Netflix data in Fig. 2 shows a pattern for the non-popular videos that is not power-law. In this case, the cause might be *information bottleneck*, which arises when users cannot easily discover niche content or the content is not properly categorized or ranked. The latent demand for products that cannot be reached due to inefficiencies in the system can have tremendous commercial and technical consequences [13]. No wonder Netflix launched the $1 million Netflix prize to improve their recommendation engine [7].

In this section, we conduct a set of analyses that provide a holistic view of the UGC video popularity distribution. We first examine how skewed users' requests are across videos. We then examine how requests are distributed across popular and non-popular content. We use two different representations of the popularity distribution: one to focus on the most popular videos, and the other, on the least popular ones. We perform graph fitting of actual data with multiple known distributions to infer the intrinsic properties of UGC popularity distribution. Although the overall distribution fits a power-law well, this is not the case for either the most popular or the least popular content. For each of these two extremes, we discuss the shape of the distribution, and hypothesize possible mechanisms that could have generated the observed distribution.

A. Pareto Principle

The Pareto Principle or the 80–20 rule is widely used to describe the degree of skew in distribution. The skew tells us how niche-centric the service is. To test if the Pareto Principle applies, we count the number of views for the rth least popular videos (Fig. 3). The horizontal axis represents the videos sorted from the most popular to the least popular, with video ranks normalized between 0 and 100. The figure represents a cumulative plot on the horizontal axis, i.e., a value of 50 represents view counts from the less popular half of all videos. The graph shows that 10% of the top popular videos account for nearly 80% of views, while the remaining 90% of videos account for total 20% of requests. Daum videos show a similar pattern. This reinforces our finding in Section III-B that there are many unpopular videos in UGC services.

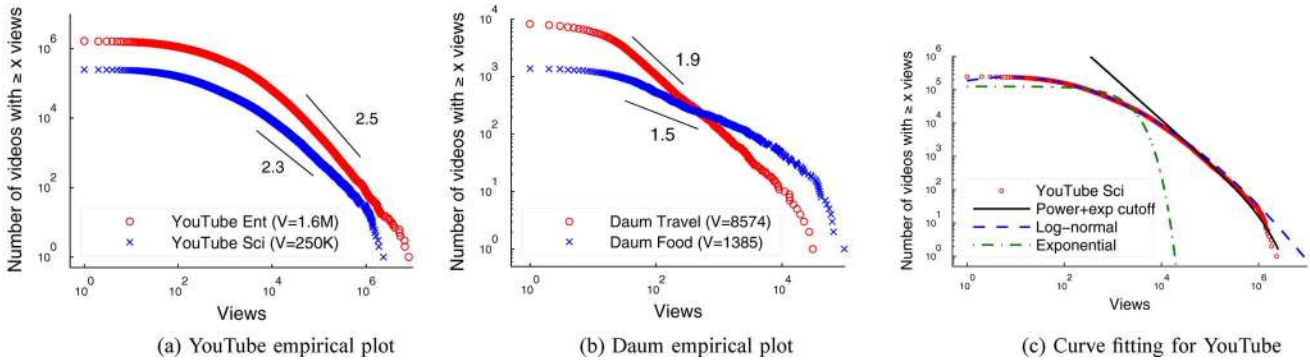
The strong skew seen in UGC services is quite surprising, since other VoD systems show a much smaller skew. For instance, analysis of the PowerInfo system showed that 90% of the least popular videos accounted for 40% of all requests [5]. One would expect that as more videos are available, users' requests should be better spread across files. However, requests on YouTube are highly skewed towards popular files. It is debatable whether such a skewed distribution is rooted in the nature of UGC (i.e., the "intended" audience of user generated content is small), or whether better recommendation engines would mitigate the strong dominance of popular content and shift users' requests toward less popular videos.

One immediate implication of the strong skew is the potential for caching. Caching can be made very efficient when storing only a small subset of objects can produce high hit ratios. By storing only 10% of long-term popular videos, a cache can serve 80% of requests. While we do not show the data here, video requests on a smaller time scale (during a day) similarly showed a skewed distribution. Another implication of the skewed distribution is the potential for peer-to-peer (P2P) distribution of popular content. In an earlier version of this paper, we explored multiple caching policies and P2P efficacy for YouTube videos [19] and found a huge potential for more efficiently delivering content. According to Huang *et al.*, strong locality (i.e., skew) suggests high potential for a viable P2P distribution [20].

B. Statistical Properties

We delve deeper into the statistical properties of UGC popularity and examine how users' requests are distributed across popular and non-popular content. To better understand each type of content, we use two different representations of the popularity distribution: (a) a frequency graph showing the number of views a video received plotted against the number of videos falling into that bin and (b) a plot of video ranks against the number of views a video received. The first representation lets us focus on the most popular videos, and the second representation, on the non-popular videos. These two plots are, in fact, transposed versions of each other and represent the same quantity [11].

1) Popular Content Analysis: Fig. 4(a) and (b) show the popularity distributions for four representative video categories in



(a) YouTube empirical plot (b) Daum empirical plot (c) Curve fitting for YouTube

Fig. 4. Video popularity distributions of YouTube and Daum videos follow a power-law distribution in the waist with exponents between 1.5 and 2.5. YouTube Sci and Daum Food exhibit decays in the tail of their distributions, which represents the most frequently viewed content.

YouTube and Daum. All of them exhibit a straight line—a characteristic behavior of the power-law distribution—for more than two orders of magnitude. The fitted power-law exponents are also shown in the figure. Interestingly, YouTube Sci and Daum Food data show a sharp decay from the straight line for the most popular videos. To examine this in detail, Fig. 4(c) shows the plot of Sci videos with the best-fit curves of log-normal, exponential, and power-law with an exponential cutoff. The best-fit power-law graph, which is not shown for clarity of other curves, has the same exponent as the power-law with an exponential cutoff curve, but its tail falls above the log-normal curve. A log-normally distributed quantity is one whose logarithm is normally distributed. The power-law with an exponential cutoff has an exponential decay term $e^{-\lambda x}$ that overwhelms the power-law behavior at large values of x. For $x < \frac{1}{\lambda}$, it is almost identical to a normal power-law, and for $x > \frac{1}{\lambda}$, to a typical exponential decay.

The results of our curve fitting suggest that the decay at the tail is best fit by the addition of the exponential cutoff to the power-law distribution.[3] However, the exact popularity distribution seems *category-dependent*. For instance, while the distribution of Daum Food also fit a power-law with an exponential cutoff, other Daum categories showed non power-law distributions. Nonetheless, most other Daum categories showed a power-law waist, with a decaying tail that is best fit by a power-law with an exponential cutoff.

As mentioned before, the shape of a distribution reflects the underlying mechanism that generates it. Several mechanisms have been proposed for power-law distributions. The most well-known explanation is the *Yule process* (also rephrased as *preferential attachment* or *rich-get-richer principle*) [21]–[23]. In UGC, this process can be translated as follows: if k users have already watched a video, then the rate at which other users watch the video is proportional to k. This explanation does not, however, explain the decay observed in the tail of the distribution. Three models have been proposed to explain the cause of such a decay. Here we review these models and explore whether they are applicable to our scenario.

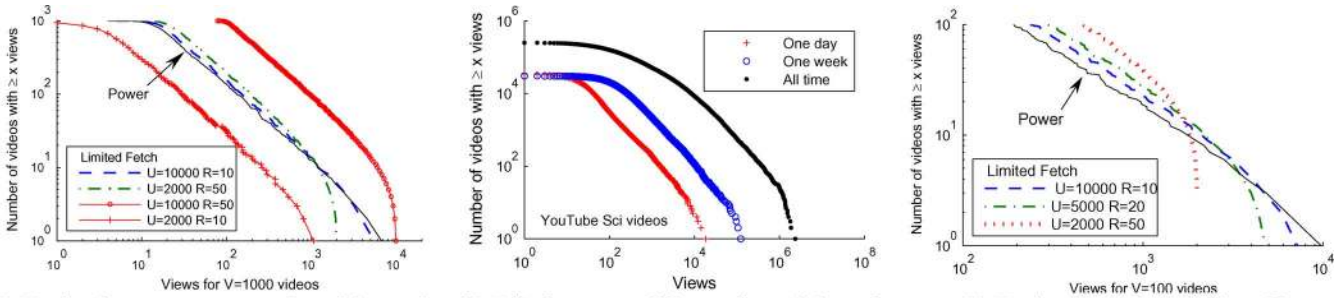
First, Amaral *et al.* suggested that the aging effect can yield a decay [24]. Consider a network of actors, whose nodes represent actors and whose edges represent movies made together by the actors. Every actor will stop acting, in time. This means that even a very highly connected node will, eventually, stop receiving new links. However, the aging effect does not apply to our case, as videos of all ages show a decaying tail. In fact, as we will see later in this paper, old videos are not necessarily inactive in YouTube: 80% of requests on a given day are towards videos older than one month and some old videos even appear in the most popular list.

Second, Mossa *et al.* considered a different model to explain the degree distribution of the WWW [25]. Along with preferential attachment that generates the power-law, the model proposes the concept of information filtering. In UGC VoD systems, this means that a user cannot receive information about all available videos, but receives information from only a fraction or a fixed number of existing pages. Due to such information filtering, preferential attachment is hindered and the exponential cutoff appears. Information filtering is surely present in both UGC and standard VoD services. However, highly popular videos are prominently featured within VoD services to attract more viewers, and thus it is unlikely that information filtering causes a decay in our case.

Gummadi *et al.* provides a better explanation of why the tail of the graph is curved [26]. In a study of file popularity in P2P downloads, they suggest that distortion arises from "fetch-at-most-once" behavior. That is, unlike the WWW traffic where a single user fetches a popular page (e.g., CNN) many times, P2P users typically fetch each object only once. Given a fixed number of users U, the videos V, and the average number of requests R per user, the authors simulate P2P downloads with two types of user populations: *Power* and *FetchOnce*. Both groups request files based on the same initial Zipf popularity. However, the Power group may request videos multiple times, whereas the FetchOne group can request videos at most once. The resulting popularity distribution (based on the number of total requests) for FetchOnce users appears curved, as opposed to the straight line observed for Power users.

UGC also has fetch-at-most-once-like behavior: since video content does not change (i.e., immutable), viewers watch the same video once or a small number of times, not at the high rates that they visit popular web pages (e.g., thousands of times over a lifetime). We call this phenomenon *limited fetch*. Expanding on the work in [26] we suggest that system characteristics such as R and V, in combination with the limited fetch behavior, can

[3]The best-fit distribution was determined by the goodness-of-fit statistics.



(a) Varying the requests per user, R, and the number of users, U (b) Tail shape over different time windows for YouTube Sci videos (c) Varying the number of videos, V

Fig. 5. Study on the impact of the limited fetch behavior on tail distribution: synthetic plots of (a) and (c), and empirical plot of (b).

result in a decaying tail. To verify this, we repeat the simulation described above with varying parameters for U, R, and V. In our setting, the Zipf parameter for initial video popularity is set to 1.0.

Fig. 5 shows the resulting video popularity in a plot of views against the cumulative number of videos. A solid line denoted "Power" represents a scenario in which users select videos based only on the initial Zipf popularity. In contrast, limited fetch scenarios, in Fig. 5(a), yield a curved tail. Interestingly, the decay in tail gets amplified as the number of requests R per user increases. This is because when R is small, the limited fetch effect barely has any impact. With increased R, the limited fetch effect plays a larger role since there is a higher chance that a particular user chooses the same popular file multiple times. The limited fetch user makes multiple draws until a new item is requested. Adding more users U in the system increases views per videos (shifting the plot in the x-axis). However, the overall shape of the graph does not change, indicating that U has little impact on the shape of the tail. Finally, increasing both R and U (from $U = 2{,}000$ and $R = 10$ to $U = 10{,}000$ and $R = 50$), the tail shape changes in a similar way as when R increases.

Note that larger values for request rate R and users U represent the case where new users are added to the system and old users make more and more requests (thus R increases). This intuitively captures what happens in real UGC systems. In fact, our traces also show similar trends. Fig. 5(b) shows the popularity distribution of the Sci category, over a short and a long time window. Having a long time window represents large R and U values. The plot of popularity during a single day exhibits a clear power-law shape, while for longer terms, the distribution exhibits a decaying tail as in Fig. 5(a).

Another factor that can greatly impact the shape of the distribution is the number of videos V. Fig. 5(c) shows the same simulation results for a smaller number of videos ($V = 100$). If V is small, the limited fetch effect is amplified since there are only a small number of videos to choose from. Likewise, Fig. 5(c) shows a highly decaying tail for $U = 2{,}000$ and $R = 50$. We can also empirically verify this from our plots of YouTube and Daum data. Revisiting the plots in Fig. 4(a) and (b), we observe that the decay in tail is much more pronounced for categories with smaller number of videos, i.e., Sci for YouTube and Food for Daum.

So far, we have examined the popularity distribution of popular content and showed, via numerical simulations and empirical validation, that tail decaying is affected by both the average number of requests per user and the number of videos in a category. Next, we focus on the non-popular portion of the distribution.

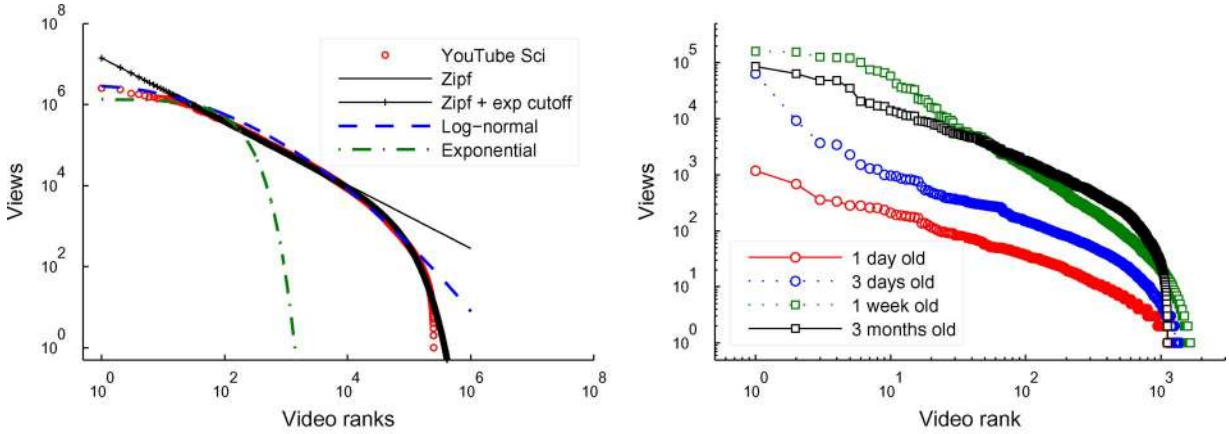
2) The Long Tail Analysis: Anderson, in his book "The Long Tail" [13], talks about huge opportunities of revenue in the unlimited number of non-popular items (e.g., by bringing more content online, enriching metadata). Here we investigate the Long Tail opportunities in UGC services. In particular, we try to answer the following questions: what is the underlying distribution of non-popular items, what shapes the distribution, and to what degree can UCG services benefit from the presence of the Long Tail? For this, we use the transposed representation of the graph: a plot of video ranks against the number of views a video received.

Fig. 6(a) shows such a plot of the Sci videos, on a log-log scale. The figure shows a straight line waist with a decaying tail. When we perform a goodness-of-fit test with several distributions, the decaying tail fits best with Zipf (or power-law) with an exponential cutoff, as clearly shown in the figure. Log-normal is the second best fit, although it does not fit well in the tail of the graph.

However, as stressed before, it is hard to decide whether a distribution is Zipf and is modulated by a bottleneck, or is just a natural log-normal distribution. Identifying the true nature of the distribution is important because it can affect strategies for marketing, target advertising, recommendation, and search engines. We discuss the two reasons for a decaying tail below:

First, one may argue that the natural shape of the UGC popularity distribution is curved (e.g., log-normal, exponential). Indeed, there are distributions that are naturally curved, for instance, particle size distributions for nano-scale fumed silica follows a log-normal distribution. Nevertheless, Zipf popularity distribution is overwhelmingly prevalent in the real world [11]. User-generated content varies widely in its quality and a significant fraction of videos may be of low interest to most users. For example, UGC is typically produced for small audiences such as family members, as opposed to professionally generated content.

Second, the natural shape of the distribution is Zipf and the decaying tail may be due to bottlenecks in the system such as information filtering or post-filters. Search or recommendation engines typically return or favor a small number of popular items [25], [27], steering users away from unpopular ones and creating a decaying tail. This decay is more apparent over time since old non-popular videos are exposed longer to such post-filtering. In-



(a) YouTube tail fitting of non-popular videos

(b) Popularity distribution of videos with varying ages

Fig. 6. Ranks versus views plot for YouTube Sci videos.

TABLE III
POTENTIAL GAIN FROM THE LONG TAIL WHEN ASSUMING THE UNDERLYING VIDEO POPULARITY FOLLOWS A ZIPF DISTRIBUTION

	Ent	Sci	Travel	Food
*Total potential gain**	45%	42%	4%	14%
*Num. beneficiary videos***	1.2M	240K	5K	400

*Percent increase in total views obtained from removing bottlenecks.
**The number of videos whose views will increase by removing bottlenecks.

deed, we are able to observe this in our traces. Fig. 6(b) shows the popularity distributions of the Sci videos of different ages. Videos aged one day are clearly less curved in the tail than older ones.

If Zipf was the natural shape and the decaying tail was due to removable bottlenecks (e.g., post-filters), then in a system with no bottleneck, the videos in the curved region would gain the deserved views. This offers users a better chance to discover rare niche videos, and also offers copyright holders and companies like YouTube and Daum potential business opportunities. To estimate the potential benefit from removal of such bottlenecks, we calculate the ratio of aggregated additional views in the best-fit Zipf curve against the existing total views. Table III shows the estimated benefits for the four UGC video categories. YouTube Ent and Sci show great opportunities in the Long Tail economics (42%–45% potential improvement), due to the large number of videos that can benefit. In Daum Travel and Food, in contrast, the total number of videos is smaller, and so the benefit is reduced. When the number of videos is small, the inefficiencies of the system (due to filtering effects) are smaller since information can be found easier.

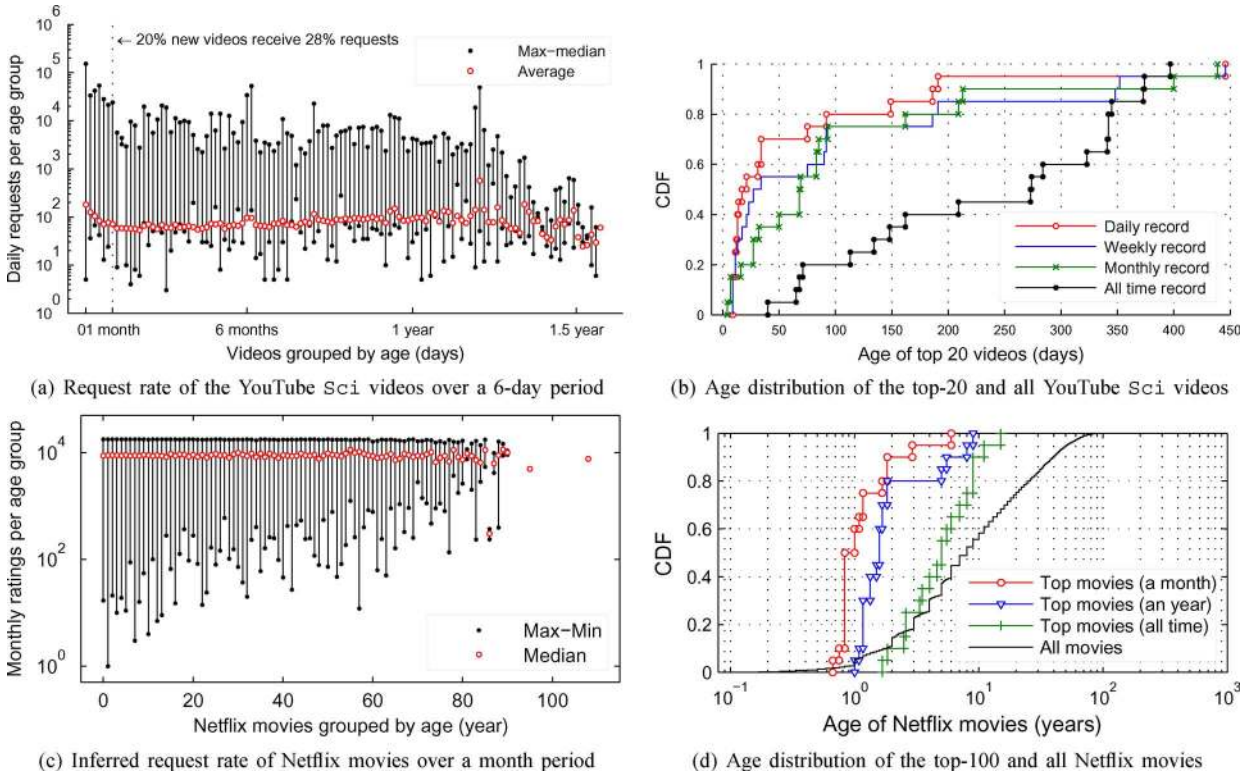
However, Zipf may not be the natural shape and the true distribution may lie between the empirical plot and Zipf. In this case, improvements from removing bottlenecks (e.g., post-filters) will not be as large, and the gains listed in Table III may be an overestimate. For most of our UGC data, goodness-of-fit suggests Zipf with an exponential cutoff as the best fit, rather than a log-normal. This makes a stronger case for filtering effects rather than a natural shape. While Zipf as well as power-law is *scale-free* in nature, exponential is a distribution that is *scaled* or *limited* in size. Therefore, the two (i.e., scale-free and scaled distributions) will rarely appear coherently and naturally as a single mechanism. Rather, a more likely explanation is that the underlying mechanism is Zipf and the exponential cutoff reveals filtering effects in the system which decays the tail. Nevertheless, revealing the true mechanism that generates the decaying tail calls for further in-depth studies.

V. POPULARITY EVOLUTION OVER TIME

As opposed to standard VoD systems where the content popularity fluctuation is rather predictable (via strategic marketing campaigns of movies), UGC video popularity can be ephemeral and has unpredictable behavior. Similarly, as opposed to the early days of TV when everyone watched the same program at the same time, such temporal correlation is diluted in UGC. Viewing patterns fluctuate based on how people get directed to such content through RSS feeds, web reviews, blogs, e-mails, or other recommendation web sites. To better understand this temporal pattern, we analyze the UGC video popularity evolution over time. Our analysis is conducted from two different angles. We first analyze whether requests concentrate on young or old videos. We then investigate how quickly popularity ranks change for videos of different ages, and further test if the future popularity of a video can be predicted. For this analysis, we use the daily trace of YouTube Sci videos.

A. Popularity Distribution Versus Age

To examine the age distribution of requested videos, we first group videos by age (binned every five days) and count the total volume of requests for each age group. More videos belonged to younger age groups than older ones. Fig. 7(a) displays the maximum, median, and the average requests per age group. We only consider videos that are requested at least once during the trace period. The vertical axis is in log-scale. For videos newer than one month, we see a slight increase in the average requests, which indicates viewers are mildly more interested in new videos. However, this trend is not very pronounced in the plot of maximum requests. Some old videos also receive significant requests. In fact, our trace showed that 80% of videos requested on a given day are older than one month and this traffic accounts for 72% of the total requests. The plot becomes noisy for age groups older than one year, due to the small number of videos. In summary, if we exclude the very new videos, users' preference (or the request rate) seems relatively insensitive to the video's age, amongst those videos that were watched within



(a) Request rate of the YouTube Sci videos over a 6-day period

(b) Age distribution of the top-20 and all YouTube Sci videos

(c) Inferred request rate of Netflix movies over a month period

(d) Age distribution of the top-100 and all Netflix movies

Fig. 7. Relationship between the request rate and video age for YouTube and Netflix videos (a),(c) and the age distribution of the most popular videos (b),(d).

the given time duration. Collectively, young videos obtained more views than older ones, because there were more videos in the younger age group.

While the users' interest is video age-insensitive on a gross scale, the videos that are requested the most on any given day seem to be the recent ones. To verify this, we look into the age distribution of the 20 most requested videos. Fig. 7(b) shows the result for four different time windows: a day, a week, a month, and all time. For each plot, we used two snapshots, taken the corresponding periods apart, and ranked videos based on the increase in their views. For the plot of "all time", we assume the initial number of views is zero. Over a day period, roughly 50% of the top 20 videos are younger than three weeks. However, as the time window increases, the median age shifts towards older videos. This suggests the ephemeral popularity characteristic of young videos.

Next, we repeat the above analysis for a standard VoD system to identify UGC specific characteristics. We use the Netflix trace and examine the number of customer ratings received per movie across the production year. Fig. 7(c) shows the maximum, median, and the minimum ratings received per movie age (grouped yearly), for those movies that were rated at least once during the one month period of December 2005. The median number of ratings is insensitive to movie age, similar to YouTube. In contrast to YouTube, the Netflix trace shows unique patterns: (a) the maximum number of ratings is strictly movie age-insensitive and (b) the minimum number of ratings is larger for old movies than for newer ones. The trace showed that 18% of ratings were made on movies released after 2003 (which accounts for 10% of the Netflix dataset), and the remaining 82% of ratings were made on older movies.

Fig. 7(d) shows the age distribution of the top 100 Netflix movies rated over a month, a year, and all time, and the age distribution of all movies in Netflix. The average age of the most popular movies (based on the customer ratings) increase as we increase the time window from one month to all time. This indicates that the most popular movies in a given month are slightly biased towards newer ones, as in YouTube. The age of the top 100 movies of the entire trace spans from nearly one to nine years, compared to the age of all videos, which spans from one month to 109 years. To better understand the dynamic popularity characteristic of UGC videos, in the following section, we discuss how video popularity evolves over time.

B. Temporal Focus

We now investigate how the popularity of individual UGC videos evolves over time, how fast or slow popularity changes, and whether the future popularity of a video can be predicted.

1) Probability of Videos Being Watched Over Time: When a video is posted, it has zero views; gradually videos gain views over time. To capture this trend in UGC videos, in Fig. 8, we show the percentage of videos aged up to X days that had no more than V views. We provide several view points by considering a range of V values from 0 to 10,000. The graph shows that after a day, 90% of videos have been watched at least once, and nearly 60% have been watched up to 10 times. After a longer period of time, more videos gain views, as expected. One noticeable trend in the graph is the consistent dips at certain times (e.g., one day, one month, one year). These points seem to coincide with the time classification made by YouTube in their video categorization. From this plot, we can see that the slope of the graph seems to decay as time passes. Noting the log-scale in the

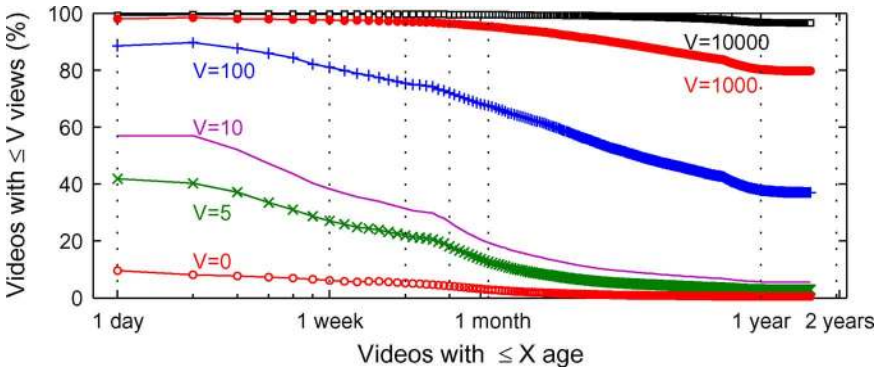


Fig. 8. Probability of videos being watched over time for YouTube Sci.

horizontal axis, this indicates the probability that a given video will be requested decreases sharply over time. In fact, if we consider the case of $V = 10$, the probability that a given video gets more than 10 requests over the first 24 hours, 6 days, 3 weeks, and 11 months, is 0.43, 0.18, 0.17, and 0.14, respectively. This indicates that if a video did not get multiple requests during its first day, it is unlikely that it will get many requests in the future. Based on these observations, we will next test if it is possible to predict a video's future popularity.

2) Predicting Near-Future Popularity: The ability to predict future popularity is useful in many ways, because the service providers may pre-populate these videos within multiple proxies or caches and the content owners may use this fast feedback to better manage their content (e.g., production companies releasing trailers to predict popularity). To explore the possibility of using early views records in predicting near-future popularity, we compare the first few days' video views with later views. If the two sets of views have a correlation of 100%, this means perfect predictability. However, if the two have a low correlation (typically lower than 0.8), then this means we are not able to predict future popularity based on the views from early age of videos.

The correlation coefficients between views of videos after seven days of upload with the views within 24 hours of upload, after one day, and after two days were 0.5885, 0.8793, and 0.9367, respectively. The first day's views are unstable as some of the videos are exposed to the system for a very short period of time (e.g., few hours or minutes). Our results show that the second day record gives an estimation with a relatively high accuracy (correlation coefficient close to 0.9). Using the third day record improves the prediction accuracy only marginally. When compared with video views after 90 days of upload, video views at the second day and third day showed correlation coefficients of 0.8425 and 0.8525, indicating a high correlation even for more distant future popularity.

3) Popularity Shifts: Now we examine the likelihood that new and old videos will become very popular as a function of their age. To observe this, we will first look at how the video rank changes over a range of video ages. In Fig. 9(a), we use two snapshots from our daily traces of six consecutive days, taken at day zero and day five, and consider only those videos that appear on both of the snapshots. We group videos by their age (binned in units of ten days) and plot the change in ranks (denoted $\Delta rank$) over age. For each age group, we plot the maximum, the 99th percentile, and the average change in $\Delta rank$ values. The vertical line plot shows the range of average to the maximum change of ranks. The vertical axis ranges from $-4{,}059$ to 235,132, which indicates that some videos decreased in their ranks by 4,059 during the trace period, while some jumped up 235,132 ranks.

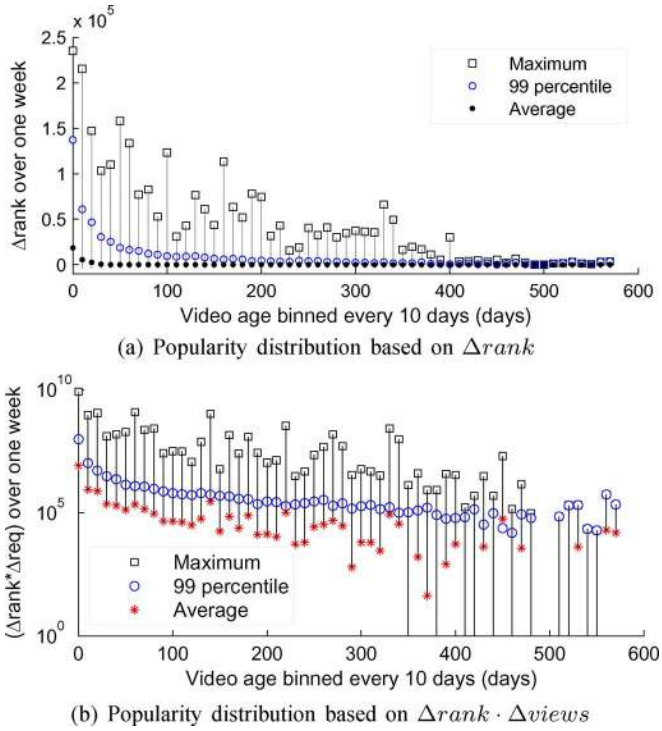


Fig. 9. Changes in ranking and popularity.

Young videos can change many rank positions very fast, while old videos have a much smaller rank fluctuation, indicating a more stable ranking classification for old videos. Still, some of the old videos also increased their ranks dramatically. This could indicate that old videos are able to ramp up the popularity ladder and become popular after a long time, e.g., due to the Long Tail effect and good recommendation engines. However, it is hard to conclude this from Fig. 9(a) since only a few requests can result in major rank changes.

The gap between the maximum and the 99th percentile lines indicates that only a few young videos (e.g., less than 1%) make large rank changes. This means that a small percentage of the young videos make it to the most popular list while the rest

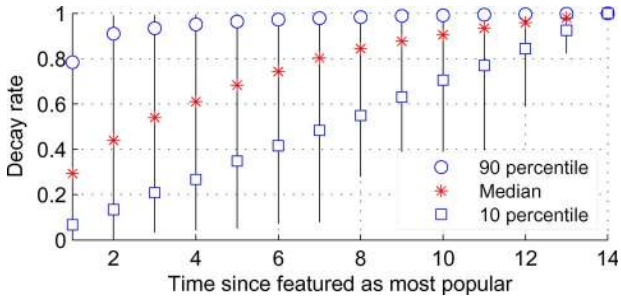


Fig. 10. Popularity decay rate of the most popular videos (first two weeks).

have much smaller ranking changes. We also see a consistent minimum $\Delta rank$ line at nearly $-4{,}000$ across all age groups. A detailed look at those videos reveals that they did not receive any requests during the trace period, however their ranking was pushed back as other videos received at least one request. This shows that unpopular videos that do not receive any requests will die in the ranking chart.

As discussed before, when it comes to identifying major shifts in the popularity distribution, considering the actual change in views or ranks is not enough. Videos can get many requests but make a minor rank change, and vice versa; a large rank change could be due to a very few requests (e.g., from zero to five requests). To identify videos that made dramatic rank changes as well as received significant requests, we propose using the product of rank changes and increment in views $(\Delta rank \cdot \Delta views)$ as in Fig. 9(b). As opposed to Fig. 9(a), the vertical axis now is in log scale. We show the range of average to maximum values as the vertical line plot. The plot does not show data point for 500 days in the horizontal axis because all videos in that group had a decrease in rank, resulting in negative values for effective change in ranks. In this graph, we observe more drastic popularity shifts for young videos; hardly any single old video received enough number of requests to make a major upward shift in the popularity chart. In short, the revival-of-the-dead effect, where old videos are suddenly brought up to the top of the chart, happens infrequently in our trace.

C. Time Evolution of the Most Popular Videos

Finally we perform a case study with a 24-day trace of the 100 "daily most popular" videos on YouTube, which we refer to as Pop. There are 2,091 unique videos in this dataset. Our study of the Pop trace allows us to better understand the evolution of popularity and the characteristics of the most requested videos. We first measure how quickly the Pop list is refreshed. When we examine the overlap of videos in any given day's list with that of the previous day, we see that, with some variability across days, on average 12% of the items are common and 88% of videos are new videos each day.

We examine how the popularity transition from hot to warm happens. In Fig. 8, we showed the popularity evolution for all videos (regardless of their popularity) as a function of their age in the system. We now focus on the most popular videos, and again follow their popularity evolution over time. Because the Pop list changes rapidly, we followed up on any video that once appeared in the Pop list and monitored their views daily from January 13th to February 5th, 2007. During the 24 days, 958 videos consistently made to the hotlist for at least two weeks, while the remaining videos appeared in the hotlist for a shorter period of time. For these 958 videos, we examine how the daily request volume changes during the first 14 days after a video has been featured as the "daily most popular". For this, we calculate the CDF of the request rate change over two weeks as follows. Let $r_i(t)$ be the request volume of video i at day t, and let T be the monitoring period. We define the *growth rate* of video i at time x as

$$d_i(x) = \frac{\sum_{i=0}^{x} r_i(t)}{\sum_{i=0}^{T} r_i(t)}. \tag{1}$$

As x reaches T, the value of $d_i(x)$ reaches 1.0. This growth rate reflects the popularity transition: videos quickly losing popularity will result in a convex shape, while a linear or concave shape will occur due to a video maintaining or increasing in popularity.

Fig. 10 shows the resulting trends of decay rate, $d_i(x)$, for 958 videos. We plot the 90th percentile and the median values of $d_i(x)$ across 14 days. The plot shows the existence of ephemeral popularity (e.g., a video receiving 98% of its requests in the first day) and videos with dramatic popularity growth (e.g., receiving less than 1% of requests on the first day). Within the first week, the median number of videos reaches 80% of their popularity for the two-week long period. Following the 90th percentile lines and the median line, we observe that most videos receive at least 30% of their requests on the first day, and then the popularity decays gradually.

VI. Aliasing and Illegal Uploads

Content aliasing and illegal uploads are critical problems in UGC systems, since they can hamper the efficiency of UGC systems as well as cause costly lawsuits. In this section, we study the prevalence of content duplication and illegal uploads in UGC, and their impact on various system's characteristics.

A. Content Aliasing

Traditional VoD services offer differently encoded versions of the same video, typically to support diverse streaming bandwidths. In UGC, there often exist multiple identical or very similar copies for a single popular event. We call this group of videos *aliases* and this new phenomenon *content aliasing*. Multiple copies of video for a single event dilute the popularity of the corresponding event, as the number of views is distributed over multiple copies. This has a direct impact on the design of recommendation and ranking systems, as it is no longer straightforward to track the popularity of an event.

To estimate the prevalence of aliases, we conducted an experiment for a subset of the top 10,000 Ent videos in YouTube. We created a web page with an interface to watch YouTube videos, search for similar videos in YouTube using any keyword, and flag videos as aliases (by clicking on check boxes given along with the search results). We recruited 51 volunteers and assigned them with non-overlapping sets of videos. Our testers watched and familiarized themselves with a total of 216 videos

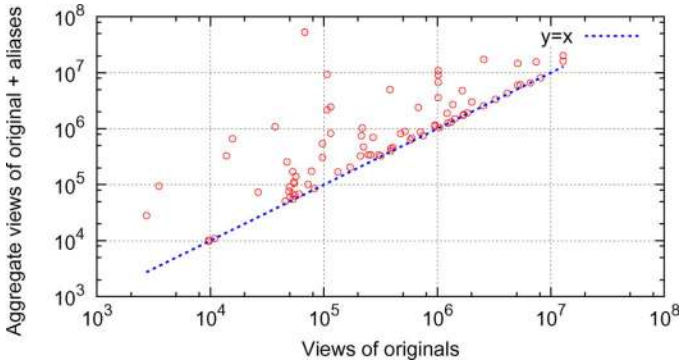


Fig. 11. The level of popularity dilution due to content duplication.

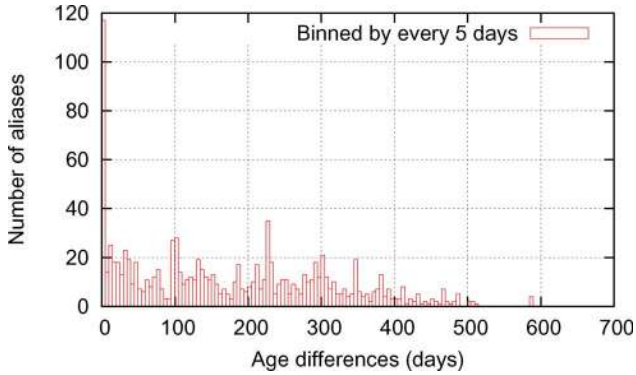


Fig. 12. Number of aliases versus the age differences.

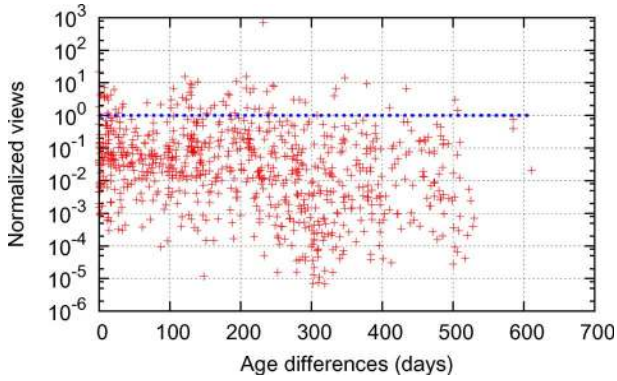


Fig. 13. Normalized views versus the age differences.

and searched for aliases using keywords of their choice through the experiment web page. Our testers were instructed to consider videos (a) taken from a different camera angle or (b) containing non-overlapping parts of longer than 1 minute, as a new video rather than an alias. Our testers identified 1,224 aliases, covering 184 out of the 216 videos in our sample set. Most videos had from 1 to 4 aliases, and the video with the most aliases had 89. For each set of aliases for one video, we called the alias with the earliest upload time the *original*.

Fig. 11 shows the sum of views from all aliases including the original against the views of the original. For some videos, the total views from aliases is more than two orders of magnitude greater than that of the original. This clearly demonstrates the popularity dilution due to content aliasing. Undiluted, the original video would be ranked much higher.

Next, we analyze the time intervals between aliases. We plot the age differences between the original video and its aliases in Fig. 12 (binned every five days). A large number of aliases are uploaded on the same day as the original video or within a week. To examine how the number of views changes over time, in Fig. 13, we plot the views of aliases normalized against that of the original versus their age difference. One conspicuous point represents an alias that showed up more than 200 days later than the original and received almost 1,000 times more views. This particular video was originally listed in the Music category, and later posted on the Comedy category. We find it rather surprising to see so many aliases cross-posted over multiple categories that appear 100 or more days after the original video. These aliases could be a potential reason for the flattened popularity tail. We leave further investigation into this delayed popularity for future work.

We do not find any correlation between the upload time of an alias and its significance in the normalized views against the original. The Pearson correlation coefficient of the plot in Fig. 13 is 0.004. This demonstrates little correlation or no decrease in the number of views over time. With aliases that appear after more than 100 days of the original, we discern no clear trend in the aliases and their views over time. Those aliases that turn up 100 days later with many fewer views are likely to serve personal archiving purposes.

Finally, we check for the existence of heavy alias uploaders. We wondered if some users might habitually post aliases of already popular videos in order to increase their own online popularity. Our data, however, shows that over 80% of all aliases are by one-time uploaders and the maximum number of aliases by one uploader is 15.

B. Illegal Uploads

UGCs derived from copyrighted content raise a serious legal dilemma for UGC service providers. A recent study from Vidmeter suggests that nearly 10% of videos in YouTube are uploaded without the permission of the content owner [28]. Vidmeter's report is based on the measurements taken every six hours of the most popular videos (i.e., daily, monthly, and all-time most-viewed videos). We augment Vidmeter's work by looking not only at the most popular videos, but also all of the videos in the Ent category.

For checking copyright violations, we focus on deleted videos. For all deleted videos, YouTube offers a notice about the reason behind deletion. Possible reasons are: removed by users, terms of use violation, copyright claim, and restricted access. To identify deleted videos, we compared a later list with an earlier list of videos. The discrepancy represents the deleted videos. Based on the first set of 1,687,506 Ent videos, 6,843 or 0.4% of videos were deleted. Only about 5% of deleted videos have violated the copyright law, which is a far smaller number than the 10% Vidmeter found [28]. This reflects the higher frequency of illegal uploads among highly ranked videos.

VII. Related Work

UGC VoD services have become extremely popular in the Internet. Among numerous UGC sharing websites that serve videos covering a broad range of topics, YouTube, MSN Video, Google Video, Yahoo! Video, AOL's UnCut Video, Grouper,

Guba, iFilm, Metacafe, Revver, and Veoh are notable. There are also specialized video services for specific topics like poker, breaking news, bicycling, lacrosse, photography, vegetarian cooking, fine wine, horror films, obscure sitcoms and Japanese anime [29].

There are a few in-depth studies on traditional VoD services. One of the first studies is by Griwodz *et al.*, where they use off-line video rental records to study video popularity [30]. Cherkasova and Gupta collected workload traces from their corporate media servers at HP and analyzed the evolution of traffic access patterns [31]. Yu *et al.* conducted an extensive analysis of access patterns and user behaviors in a centralized VoD system [5]. More recently, Huang *et al.* used the VoD server logs of MSN video and characterized the VoD traffic to explore the potential for peer-assisted VoD service [20].

Compared to these works, relatively few attempts have been made to understand how these UGC services are fundamentally different from traditional well-explored video distribution services [32]. Gill *et al.* analyzed YouTube traffic generated by a particular collection of clients, providing a detailed view of local UGC service usage [33]. In this paper and an early version of this work (which appeared in the ACM Internet Measurement Conference in 2007 [19]), we have presented a comprehensive analysis of the popularity distribution and the time evolution of UGC video requests and their implications. Our work provides a complementary "global view" by crawling metadata of complete sets of video categories from two of the major UGC systems.

In a study of popularity distributions, Newman carried out a comprehensive study of power-law distributions [11]. He examined several domains that are well-specified by a power-law: Web hits, copies of books sold, telephone calls, etc. Also, Alderson *et al.* developed an interesting and rich theory for scale-free networks [34]. The power-law distribution with a decaying tail has frequently appeared in the degree distributions of various real-world networks such as the WWW, protein networks, e-mail networks, actor networks, and scientific collaboration networks [25], [35], [36].

VIII. Conclusion

In this paper, we presented a data-driven analysis of the static popularity distribution, dynamic popularity evolution, and content duplication of user-generated content (UGC). For our study, we collected traces from two large UGC video websites, YouTube and Daum Videos. We demonstrated how UGC services are different from traditional VoD services, based on an empirical comparison. The ability for anyone to create content has led to unique production patterns for UGC such as the massive content, short video length, and heavy publishers. The convenience of the Web and the infinite choice have led to two distinct consumption patterns: certain UGC videos become extremely popular and reach tens of millions of viewers, while others are less popular but have a good chance of reaching niche audiences.

We studied the nature of user behavior in UGC video services and the key elements that shape the popularity distribution—namely, what shapes the Long Tail, alters the skewness of popularity, or breaks the power-law pattern for very popular content. Our results indicated that, assuming Zipf is the underlying popularity distribution, lower-than-expected popularity of niche content could be explained by information filtering (e.g., poor search and recommendation engines, missing metadata). We estimated that leveraging such latent demand could increase the total views by as much as 45%. However, if the underlying distribution is naturally curved, there will not be such a huge potential in the Long Tail. Revealing the true nature of the curved tail thus is important and calls for further in-depth studies.

Our study of popularity evolution over time showed that videos older than one month account for more than 70% of requests (this pattern is also true in Netflix). Also, content popularity is mostly determined at the early stage of video age, explaining why it is rare for non-popular, old videos to be suddenly brought up as hits. Our study of the most popular videos captured the ephemeral lifetime of the daily hotlist.

Finally, we tackled the impact of content aliasing and illegal uploads, which could hamper the future success of UGC services. Using a small set of randomly chosen videos and their aliases, we demonstrated that content aliasing is widely practiced and that it makes video ranking difficult. Also, we found that illegal uploads are more common amongst highly ranked videos. Recently, YouTube announced its Video Identification Beta tool which automatically compares and differentiates any two videos' content, to combat illegal aliases of copyrighted content [37].

In summary, we believe that our work provides a basis for the design of future UGC systems. There are several directions that we wish to pursue as future work. The first is to study the impact of various features embedded in websites in making videos popular. For example, we are interested in knowing how featuring on the front page, links from external websites, changes in the YouTube user base, and massive uploads affect video popularity. Second, using detailed HTTP server logs, we would like to design practical caching and P2P distribution strategies that can reduce the video server load for UGC distribution. We would like to determine which cache replacement policy is the best for UGC and how many video requests can be served from peers within the same geographical location. Finally, it would be interesting to see to what extent our results in this paper hold in the future and across other UGC systems.

Acknowledgment

The authors thank J. Jung, C. Howell, R. Hoberman, and the anonymous reviewers for many helpful comments, the volunteers who have participated in content duplication testing, D. Towsley, J. Crowcroft, C. Gkantsidis, T. Karagiannis, R. Rejaie, and C. Domingo, and members at Telefonica Research for their comments on an early version of this work.

References

[1] N. Miller, "Manifesto for a new age," *Wired Magazine*, Mar. 2007.
[2] YouTube. [Online]. Available: http://www.youtube.com
[3] Adobe Flash Player Version Penetration. [Online]. Available: http://www.adobe.com/products/player_census/flashplayer/version_penetration.html
[4] Daum User Created Content. [Online]. Available: http://ucc.daum.net
[5] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in *Proc. ACM Eurosys*, 2006.

[6] IMDB Statistics. [Online]. Available: http://www.imdb.com/database_statistics
[7] Netflix Prize. [Online]. Available: http://www.netflixprize.com
[8] LoveFilm. [Online]. Available: http://www.lovefilm.com
[9] Yahoo! Movies. [Online]. Available: http://movies.yahoo.com
[10] E. Auchard, "Participation on Web 2.0 sites remains weak," 2007 [Online]. Available: http://www.reuters.com/article/internetNews/idUSN1743638820070418
[11] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, p. 323, 2005.
[12] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Topical interests and the mitigation of search engine bias," in *Proc. Natl. Acad. Sci.*, 2006.
[13] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*. New York: Hyperion, 2006.
[14] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Mathematics*, vol. 1, no. 2, pp. 226–251, 2004.
[15] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.
[16] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues," *BioScience*, vol. 51, no. 5, pp. 341–352, 2001.
[17] A. B. Downey, "The structural cause of file size distributions," in *Proc. IEEE MASCOTS*, 2001.
[18] W. Gong, Y. Liu, V. Misra, and D. Towsley, "On the tails of web file size distributions," in *Proc. 39th Allerton Conf. Communication, Control, and Computing*, Univ. Illinois, 2001.
[19] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM IMC*, 2007.
[20] C. Huang, J. Li, and K. W. Ross, "Can internet video-on-demand be profitable?," in *Proc. ACM SIGCOMM*, 2007.
[21] G. Yule, "A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S.," *Royal Soc. London Philosoph. Trans. Ser. B*, vol. 213, pp. 21–87, 1925.
[22] Y. Ijiri and H. Simon, *Skew Distributions and the Size of Business Firms*. Amsterdam: North Holland, 1977.
[23] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.
[24] L. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," in *Proc. Natl. Acad. Sci.*, 2000.
[25] S. Mossa, M. Barthélémy, H. E. Stanley, and L. A. N. Amaral1, "Truncation of power law behavior in "Scale-free" network models due to information filtering," *Phys. Rev. Lett.*, vol. 88, no. 13, p. 138701, 2002.
[26] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, "Measurement, modeling, and analysis of a peer-to-peer file-sharing workload," in *Proc. ACM SOSP*, 2003.
[27] J. Cho and S. Roy, "Impact of search engines on page popularity," in *Proc. WWW*, 2004.
[28] B. Holt, H. R. Lynn, and M. Sowers, "Analysis of copyrighted videos on YouTube.com," [Online]. Available: http://www.vidmeter.com/i/vidmeter_copyright_report.pdf
[29] IP TV Evangelist, "Riding the Long Tail, an interview with Chris Anderson," [Online]. Available: http://www.iptvevangelist.com/2006/12/riding_the_tail_an_interview_w.html
[30] C. Griwodz, M. Biig, and L. C. Wolf, "Long-term movie popularity models in video-on-demand systems," in *Proc. ACM Multimedia*, 1997.
[31] L. Cherkasova and M. Gupta, "Analysis of enterprise media server workloads: Access patterns," *IEEE/ACM Trans. Networking*, vol. 12, no. 5, pp. 781–794, Oct. 2004.
[32] L. Gomes, "Will all of us get our 15 minutes on a YouTube Video?," *The Wall Street Journal Online*, Aug. 2006.
[33] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," in *Proc. ACM IMC*, 2007.
[34] D. A. L. Li, J. Doyle, and W. Willinger, "Towards a theory of scale-free graphs: Definition, properties, and implications," *Internet Mathematics*, vol. 2, no. 4, 2006.
[35] T. Fenner, M. Levene, and G. Loizou, "A stochastic evolutionary model exhibiting power-law behaviour with an exponential cutoff," *Physica*, no. 13, pp. 641–656, 2005.
[36] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto, "Analyzing client interactivity in streaming media," in *Proc. WWW*, 2004.
[37] YouTube Video Identification Beta. [Online]. Available: http://www.youtube.com/t/video_id_about

Meeyoung Cha received the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2008.

She is a postdoctoral researcher at the Max Planck Institute for Software Systems (MPI-SWS) in Germany. Her research interests are in the design and analysis of large-scale networked systems. Her recent work has focused on multimedia streaming systems and online social networks.

Dr. Cha won the Best Paper Award at the ACM Internet Measurement Conference 2007 for her work characterizing the YouTube workload.

Haewoon Kwak received the B.S. and M.S. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2006 and 2007, respectively, where he is currently pursuing the Ph.D. degree in computer science.

He is working on social networks and user behaviors in web 2.0 services. Recently, he studied recommender systems as an intern at Telefonica Research, Barcelona, Spain. His advisor is Dr. Sue Moon.

Pablo Rodriguez received the Ph.D. degree from the Swiss Federal Institute of Technology and the M.S. degree in telecommunications engineering from the University of Navarra, Spain.

Currently, he is the Scientific Director at Telefonica Research, Barcelona, Spain, leading the Internet research area. Previously, he was at Microsoft Research, Cambridge, UK, where he led projects in P2P networks and distributed systems and at Bell Laboratories, NJ, USA, as a researcher. He also worked as a software architect for Inktomi and Tahoe Networks, Silicon Valley start-ups. He holds over 25 patents in the Internet and wireless related technologies and he frequently consults for networking and Internet start-up companies.

Yong-Yeol Ahn studied physics at the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. He received the Ph.D from KAIST in 2008. His advisor was Dr. Hawoong Jeong.

He joined CCNR at Northeastern University, Boston, MA, in June 2008. His research interests include evolution, optimized structure of living organisms, relationship between structure and dynamics in complex networks, robustness of genetic networks, and pattern of online social interactions.

Sue Moon received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively, in computer engineering. She received the Ph.D. degree in computer science from the University of Massachusetts at Amherst in 2000.

From 1999 to 2003, she worked in the IPMON project at Sprint ATL, Burlingame, California. In August 2003, she joined the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, as an Assistant Professor. Her research interests are in network performance measurement and monitoring of diverse network types and their security and anomalous aspects.