I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System

Meeyoung Cha*, Haewoon Kwak†, Pablo Rodriguez*, Yong-Yeol Ahn†, and Sue Moon†
*Telefonica Research, Barcelona, Spain
†KAIST, Daejeon, Korea
meeyoung.cha@gmail.com, haewoon@an.kaist.ac.kr, pablorr@tid.es, yongyeol@gmail.com, sbmoon@cs.kaist.ac.kr

ABSTRACT

User Generated Content (UGC) is re-shaping the way people watch video and TV, with millions of video producers and consumers. In particular, UGC sites are creating new viewing patterns and social interactions, empowering users to be more creative, and developing new business opportunities. To better understand the impact of UGC systems, we have analyzed YouTube, the world's largest UGC VoD system. Based on a large amount of data collected, we provide an in-depth study of YouTube and other similar UGC systems. In particular, we study the popularity life-cycle of videos, the intrinsic statistical properties of requests and their relationship with video age, and the level of content aliasing or of illegal content in the system. We also provide insights on the potential for more efficient UGC VoD systems (e.g. utilizing P2P techniques or making better use of caching). Finally, we discuss the opportunities to leverage the latent demand for niche videos that are not reached today due to information filtering effects or other system scarcity distortions. Overall, we believe that the results presented in this paper are crucial in understanding UGC systems and can provide valuable information to ISPs, site administrators, and content owners with major commercial and technical implications.

Categories and Subject Descriptors

C.2.0 [**COMPUTER-COMMUNICATION NETWORKS**]: General

General Terms

Measurement, Design

Keywords

User Generated Content, Power-Law, Long Tail, VoD, P2P, Caching, Popularity Analysis, Content Aliasing

*The data traces used in this paper are shared for the wider community use at `http://an.kaist.ac.kr/traces/IMC2007.html`.

IMC'07, October 24-26, 2007, San Diego, California, USA.

1. INTRODUCTION

Video content in standard Video-on-Demand (VoD) systems has been historically created and supplied by a limited number of media producers, such as licensed broadcasters and production companies. Content popularity was somewhat controllable through professional marketing campaigns. The advent of user-generated content (UGC) has re-shaped the online video market enormously. Nowadays, hundreds of millions of Internet users are self-publishing consumers. The content length is shortened by two orders of magnitude and so is the production time. Wired magazine refers to this small-sized content pop culture as "bite-size bits for high-speed munching" [34].

The scale, dynamics, and decentralization of the UGC videos make traditional content popularity prediction unsuitable. UGC popularity is more ephemeral and has a much more unpredictable behavior. As opposed to the early days of TV where everyone watched the same program at the same time, such strong reinforcement of popularity (or unpopularity) is diluted in UGC. Constant waves of new videos and the convenience of the Web are quickly personalizing the viewing experience, leading to a great variability in user behavior and attention span. Understanding the popularity characteristics is important because it can bring forward the latent demand created by bottlenecks in the system (e.g. poor search and recommendation engines, lack of metadata). It also greatly affects the strategies for marketing, target advertising, recommendation, and search engines. At the same time, a lack of editorial control in UGC is creating problems of content aliasing or copyright infringement, which seriously threatens the future viability of such systems.

To understand the nature and the impact of UGC systems, in this paper we analyze YouTube, the world's largest UGC VoD system. The main contribution of this paper is an extensive trace-driven analysis of UGC video popularity distributions. To this extent, we have collected a large amount of data from YouTube and another UGC system, Daum. Our analysis reveals very interesting properties regarding the distribution of requests across videos, the evolution of viewer's focus, and the shifts in popularity. Such analysis is pivotal in understanding some of the most pressing questions regarding new business opportunities in UGC. Our analysis also reveals key results regarding the level of piracy and the level of content duplication in such systems, which could have major implications in the deployment of future UGC services.

The highlights of our work could be summarized as follows:

1. We compare some prominent UGC systems with other standard VoD systems such as Netflix and Lovefilm. We highlight the main differences between the two systems and point out interesting properties regarding content production, consumption, and user participation patterns.

2. By analyzing the popularity distributions from various categories of UGC services and by tracking the time evolution of it, we show that the popularity distribution of UGC exhibits power-law with truncated tails. We discuss several filtering mechanisms that create truncated power-law distributions, and estimate the potential benefits arising from the hidden latent demand caused by such filtering effects.

3. We provide insights into more efficient UGC distribution systems, namely, caching and peer-to-peer (P2P). Our analysis can be of great value to content providers and site administrators due to the large amount of network traffic generated by UGC.

4. We measure the prevalence of content duplication and illegal uploads in UGC, and their impact in various system characteristics. Content aliasing and illegal uploads are critical problems of today's UGC systems, since they can hamper the efficiency of UGC systems and cause costly lawsuits.

The rest of the paper is organized as follows. §2 describes our trace methodology and the key characteristics of UGC. In §3, we analyze the popularity distribution of UGC and the forces that shape it. §4 investigates how popularity of videos evolve over time. §5 considers the performance potential of server workload and bandwidth savings via caching and P2P. §6 focuses on the level of content duplication and illegal uploads in UGC. Finally, we present related works in §7 and in §8, we conclude.

2. METHODOLOGY AND PROPERTIES

This section introduces our data collection process and the general properties of the measured UGC videos.

2.1 Data Collection

Our dataset consists of meta-information about user-generated videos from YouTube and Daum UGC services. **YouTube**, the world's largest UGC site, serves 100 million distinct videos and $65,000$ uploads daily [6]. **Daum UCC**, the most popular UGC service in Korea, is well-known for its high-quality videos (streaming as high as at 800 kb/s) and serves two million visitors and 35 million views weekly [1].

We crawled YouTube and Daum sites and collected meta information about videos by visiting their indexed pages that link all videos belonging to a category. Due to the massive scale of YouTube, we limited our data collection to two of the categories: 'Entertainment' and 'Science & Technology' (now called 'Howto & DIY'). Throughout this paper, we simply refer to them as `Ent` and `Sci`. For Daum, we have collected video information from all the categories. Each video record contains fixed information (such as the uploader, the upload time, and the length) and time-varying information (such as views, ratings, stars, and links). *Views* and *ratings* indicate the number of times the video has been played or evaluated by users. *Stars* indicate the average score from rating, and *links* indicate the list of external web pages hyper-linking the video. Our traces include multiple snapshots of video information taken daily across six days for `Sci`. These multiple snapshots give insights on the actual request patterns and the popularity evolution of UGC videos. Table 1 summarizes our dataset with basic statistics.

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

2.2 UGC versus Non-UGC

Next, we highlight the key differences and similarities between UGC and non-UGC (or professionally generated contents). For comparison purposes, we use data from three representative non-UGC services. **Netflix**, a popular online video rental store, has made customer ratings for their 17,770 videos publicly available at [4]. We use this data set in our comparison. We additionally crawled the web site of **Lovefilm** [3], Europe's largest online DVD rental store, and **Yahoo! Movies** [5] for meta-information about their movie collections. Our Lovefilm dataset contains the video length and the director. Our Yahoo dataset contains the daily top ten US Box Office Chart from 2004 to March 2007, and their theater gross. Table 2 summarizes the non-UGC dataset.

Table 2: Summary of non-UGC traces

Trace	# Videos	Period	Description
`Netflix`	17,770	Oct 2006	Customer ratings
`Lovefilm`	39,447	Jan 2007	Length and director
`Yahoo`	361	2004 - 2007	Theater gross income

2.2.1 Content Production Patterns

One key characteristic of UGC is the fast content production rate. The scale of production of UGC is strikingly different compared to non-UGC. For example, IMDb, the largest online movie database, carries 963,309 titles of movies and TV episodes produced since 1888 until today [2]. In contrast, YouTube enjoys 65,000 daily new uploads – which means that it only takes 15 days in YouTube to produce the same number of videos as all IMDb movies.

UGC requires less production efforts, compared to non-UGC. Accordingly, the number of distinct publishers is massive for UGC. The average number of posts per publisher, however, is similar for UGC and non-UGC (e.g. 90% of film directors publish less than 10 movies, based on Lovefilm, and similarly 90% of UGC publishers upload less than 30 videos in YouTube). Interestingly, there exist extremely heavy publishers in UGC, who post over 1,000 videos over a few years. In contrast, the largest number of movies produced by a single director scales only up to a hundred movies over half a century.

The length of UGC videos varies across categories. Daum `CF` category shows the shortest median length of 30 seconds, while Daum `Music Video` category shows the longest median length of 203 seconds. Compared with non-UGC, the UGC video length is shorter by two orders of magnitude. The median movie length in Lovefilm is 94 minutes.

Table 1: Summary of UGC traces

Name	Category	# Videos	Tot. views	Tot. length	Data collection period
YouTube	`Ent`	1,687,506	3,708,600,000	15.2 years	Dec 28, 2006 (crawled once)
YouTube	`Sci`	252,255	539,868,316	1.8 years	Jan 14 - 19, '07 (daily), Feb 14, '07, Mar 15, '07 (once)
Daum	All	196,037	207,555,622	1.0 year	Mar 1, 2007 (crawled once)

2.2.2 User Participation

The video popularity and ratings (i.e. the number of viewers who evaluated the video) show a strong linear relationship for both UGC and non-UGC, with the correlation coefficient of 0.8 for YouTube and 0.87 for Yahoo. This is an interesting observation, because it indicates that users are not biased towards rating popular videos more than unpopular ones.

Despite the Web 2.0 features added in YouTube to encourage user participation, the level of active user participation is 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 total views. Other Web 2.0 sites also have reported similar trends on relatively low user involvements [11].

2.2.3 How Content Is Found?

We now examine the Web pages that link to YouTube videos. Based on `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 very significant. We have identified that the top five web sites linking to videos in YouTube `Sci` are `myspace.com`, `blogspot.com`, `orkut.com`, `Qooqle.jp`, and `friendster.com`; four of them from social networking sites, and one on video recommendation.

3. IS UGC POPULARITY POWER-LAW?

Analyzing the exact form of probability distribution not only helps us understand the underlying mechanism, but also helps us answer important design questions in UGC services. This has also been true in other areas [10, 21, 37]. For instance, the scale-free nature of Web requests has been used to improve search engines and advertising policies. The distribution of book sales has also been used to design better online stores and recommendation engines.

The power-law model has been increasingly used to explain various statistics appearing in the computer science and networking applications. A distinguished feature of power-law is a straight line in the log-log plot of views versus frequency. However, there are other distributions (e.g. log-normal) with very similar shape. Hence it is a nontrivial task to determine whether a certain distribution is power-law or log-normal, unless the plot shows a clear straight line across several orders of magnitude [17, 19, 32, 35, 38].

The shape of a distribution reflects the underlying mechanism that generates it. Normally, the power-law distribution arises from *rich-get-richer principle*, while the log-normal distribution arises from the law of proportionate effect[1]. However, in a real-world, the shape of the natural distribution can be affected due to various reasons. In fact, many distributions whose underlying mechanism is power-law fail to show clear power-law patterns, especially at the two ends of the distribution: the most popular and the least popular items. For instance, in the case of movies in cinemas [9], the distortion often comes from the lack of enough movie theaters, where niche content cannot be screened as much as it should. This is a *distribution bottleneck*, which can be removed by bringing the content online.

Yet, there are other bottlenecks that can distort the shape of a distribution. The NetFlix data in Figure 1 shows a pattern for the non-popular videos that is not power-law. In this case, however, it is an *information bottleneck*. This relates to the fact that users cannot easily discover niche content, or content is not properly categorized or ranked[2]. The latent demand for products that cannot be reached due to inefficiencies in system, can have tremendous commercial and technical consequences [10]. No wonder NetFlix recently launched the $1 million netflix prize to improve their recommendation engine [4].

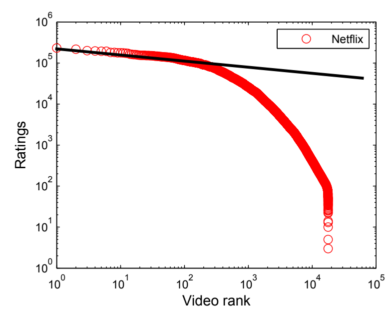


Figure 1: Empirical plot of ranks against ratings, with a synthetic power-law fitted for ranks 1 to 100.

In the rest of this section, we study the popularity distribution properties of YouTube.

3.1 Pareto Principle

The Pareto Principle (or 80-20 rule) is widely used to describe the skewness in distribution. Such skewness tells us how niche-centric the service is. To test the Pareto Principle, we count the number of views for the least r-th popular videos and show it in Figure 2. The horizontal axis represents the videos sorted from the most popular to the least popular, with video ranks normalized between 0 and 100. The graph shows that 10% of the top popular videos ac-

[1]The log-normal distribution is very similar to the normal distribution; the difference is at is multiplicative process, not additive.

[2]Note that we plot customer ratings rather than views since this was the only data available [4]. However, we have observed from other VoD and UGC sites that ratings and views are related by a linear relationship (see §2.2.2). Thus the general distribution presented in this plot should not differ greatly when plotting rank against views.

count for nearly 80% of views, while the rest 90% of the videos account for very requests. Note that Daum data also reveals a similar behavior.

This result is quite surprising, since other online systems show much smaller skew. For instance, analysis of a large VoD system in China, PowerInfo, shows that 90% of least popular VoD files account for 40% of all requests [39]. One would expect that as more videos are made available, users' requests should be better spread across files. However, counter-intuitively, requests on YouTube seem to be highly skewed towards popular files. It is debatable whether such skewed distribution is rooted in the nature of UGC (because people primarily want to see what others have seen before), or whether better recommendation engines would mitigate the strong dominance of popular content and shift the users' requests toward less popular videos.

A nice immediate implication of this skewed distribution is that caching can be made very efficient since storing only a small set of objects can produce high hit ratios. That is, by storing only 10% of long-term popular videos, a cache can serve 80% of requests. We revisit caching in more detail in §5.1.

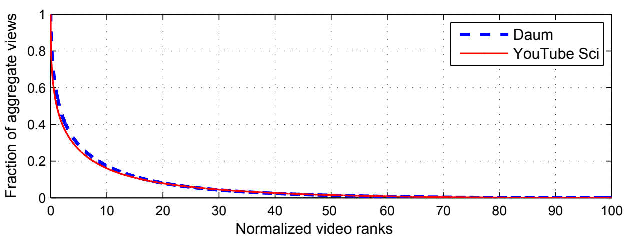


Figure 2: Skewness of user interests across videos

3.2 Statistical Properties

We now delve deeper into the actual statistical properties of UGC popularity, focusing on how users' requests are distributed across popular and non-popular content. To better understand each type of content, we will use two different representations of the popularity distribution: a plot of views against the complementary cumulative number of views (i.e. frequency) and a plot of video ranks against views. With the first representation we can focus on the most popular videos. The second representation shows better the behavior of unpopular videos and has recently been used to understand the so-called "the Long Tail" by Anderson [10]. Note that these two plots are in fact transposed versions of each other and represent the same quantity [37].

3.2.1 Popular Content Analysis

Figures 3(a) and (b) show the popularity distribution of videos for four representative categories of YouTube and Daum. All of them exhibit power-law behavior (a straight line in a log-log plot) across more than two orders of magnitude. The fitted power-law exponents are also shown in the figure. However, YouTube `Sci` and Daum `Food` categories show a sharp decay for the most popular content. To examine the truncation in detail, Figure 3(c) shows the plot of `Sci` with the best-fit curves of power-law, log-normal, exponential, and power-law with an exponential cutoff. A log-normally distributed quantity is one whose logarithm is normally distributed. 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 normal exponential decay.

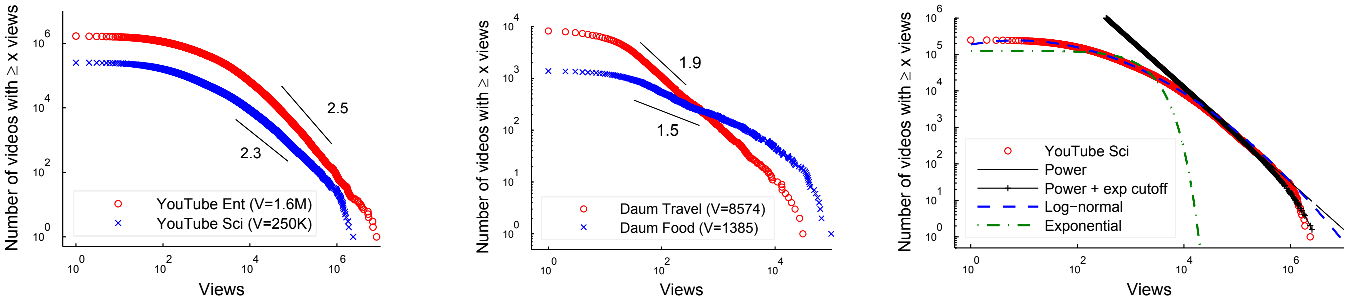
Our fitting result suggests that truncation at the tail follows power-law with an exponential cutoff. However, the exact popularity distribution seems *category-dependent*. For instance, while the distribution of Daum `Food` also showed power-law with exponential cutoff, other Daum categories (not shown here) showed non power-law distributions. Nonetheless, most of them showed *power-law waist*, with *a truncated tail* that fits best by power-law with an exponential cutoff.

There are several mechanisms that generate power-law distributions. The simplest and the most convincing one is the *Yule process* (also rephrased as *preferential attachment* or *rich-get-richer principle*) [12, 30, 40]. In UGC, this process can be translated as follows: if k users have already watched a video, then the rate of other users watching the video is proportional to k. We will now investigate why a power-law distribution can have an exponential cutoff on the most popular content.

To this extent, we first review two models that have been suggested to explain the cause of such truncation and explore whether they are applicable to our scenario. First, Amaral *et al.* suggested that the aging effect can yield truncation [8]. Consider a network of actors, where every actor will stop acting, in time. This means that even a very highly connected vertex will, eventually, stop receiving new links. However, the aging effect does not apply to our case, as videos across all ages show truncated tail. In fact, as we will see later in the paper, our daily trace shows that 80% of the videos requested on a given day are older than 1 month, contradicting the hypothesis of aging effect in our case.

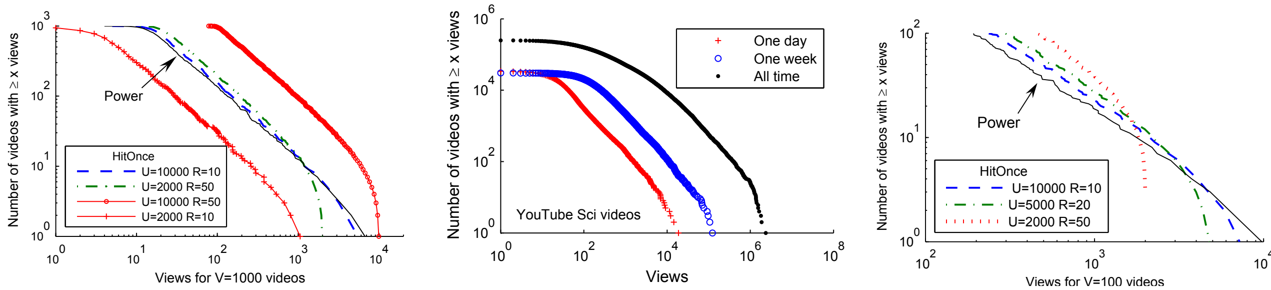
Second, Mossa *et al.* considered a different model to explain the degree distribution of the WWW [36]. Along with the preferential attachment, the model adopts the concept of information filtering, which means that a user cannot regard all the information but receive information from only a fraction or a fixed number of existing pages. Due to this information filtering process, the preferential attachment is hindered and the exponential cutoff appears. The information filtering is surely present also in both UGC and standard VoD services. However, highly popular videos are prominently featured within these VoD services to attract more viewers, and thus it is unlikely that information filtering causes truncation in our case.

Instead, Gummadi *et al.* give us some better hints on the causes for our truncated tail [26]. In their study of file popularity in P2P downloads, they suggest the cause of distortion arises from "fetch-at-most-once" behavior of users. That is, unlike the WWW traffic where a single user fetches a popular page (e.g. CNN) many times, P2P users fetch most objects once. Given a fixed number of users, U, the videos, V, and the average number of requests per user, R, the authors simulate P2P downloads with two types of user populations: *Power* and *HitOnce*. Both user groups make requests based on the same initial Zipf file popularity. However, Power group may request videos multiple times, and HitOnce group, at most once. HitOnce user will make multiple draws until a new item is requested. The resulting popularity graph for HitOnce users appears truncated, as opposed to the straight line behavior of Power users [26].



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

Figure 3: Video popularity distribution of YouTube and Daum follows power-law in the waist, with varying exponent from 1.5 to 2.5. YouTube `Sci` and Daum `Food` exhibit sharp decay in the tail of hot content.



(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

Figure 4: Study on the impact of the "fetch-at-most-once" on tail distribution: synthetic plots, (a) and (c), and empirical plot, (b)

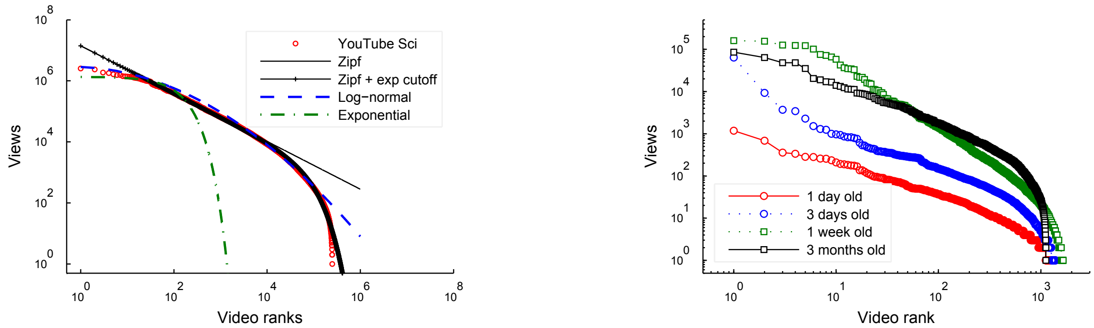
UGC also has "Fetch-at-most-once"-like behavior; since video content does not change (i.e. immutable) viewers are not likely to watch the same video multiple times, as they do for mutable web objects. Expanding on the work in [26] we suggest that other system characteristics such as R and V, in combination with the "fetch-at-most-once", can have a major impact in forming the truncated tail. To verify this, we repeat the simulation described above with varying parameters for U, R, and V. In our setting, the Zipf parameter is set as 1.0 for the initial video popularity.

Figure 4 shows the resulting video popularity in a plot of views against the cumulative number of videos. We make several observations from Figure 4(a). First, compared with Power, all HitOnce scenarios show a truncated tail, as expected. Interestingly, the truncated tail gets amplified as the number of requests per user, R, increases. If R is small, then the "fetch-at-most-once" effect barely takes place. With increased R, the "fetch-at-most-once" effect starts playing a bigger role, since there is a higher chance that a particular user is geared towards the same popular file multiple times. Second, adding more users in the system, U, 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 in the tail truncation. Finally, increasing both R and U (from $U = 2000, R = 10$ to $U = 10000, R = 50$), the tail shape changes in a similar way as when R increases.

Note that larger R and U values 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 the real UGC systems. In fact, our traces also show similar trends. Figure 4(b) shows the popularity distribution of `Sci`, over a short and long-term window. Having a long-term window represents large R and U values. The plot of popularity during one day (i.e. small R) exhibits a clear power-law decay, while for longer terms, the distribution exhibits a truncated tail as in Figure 4(a).

Another factor that can greatly impact the shape of a distribution is the number of videos, V. Figure 4(c) shows the same simulation results, repeating for a smaller number of videos ($V = 100$). If V is small, the "fetch-at-most-once" effect becomes amplified since there is only a small number of videos to choose from. This results in a highly truncated tail, as shown in Figure 4(c) for the case of $U = 2000, R = 50$. We can also empirically verify this from our plots of YouTube and Daum data. Let us revisit the plots in Figures 3(a) and (b). We observe that the tail cutoff is much more pronounced for categories with smaller number of videos, such as `Sci` in the case of YouTube and `Food` in the case of Daum.

So far, we focused on the popularity distribution of popular content and showed, via numerical simulations and empirical validation, that the tail truncation is affected by both the average requests per users and the number of videos in a category. Next, we move on to the non-popular portion of the distribution.



(a) YouTube tail fitting of non-popular videos (b) Popularity distribution of videos with varying ages

Figure 5: Ranks versus views plot for YouTube `Sci` videos.

3.2.2 The Long Tail Analysis

Anderson, in his book "the Long Tail" [10], asserted that there exist huge opportunities in the unlimited number of non-popular items. Here we will 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 in one way or another, and how much benefit the Long Tail can bring for UGC services.

Let us first look into the distribution of the non-popular content. We use a plot of video ranks against views, where unpopular videos are put at the tail. This representation, suggested by Zipf, has been used to observe Zipf's law. Figure 5(a) shows such empirical plot of `Sci` videos, on a log-log scale. The figure shows a Zipf-like waist (a straight line in a log-log plot) with a truncated tail. When we perform goodness-of-fit test with several distributions, the truncated tail fits best with Zipf with an exponential cutoff, as clearly shown in the figure. Log-normal is the second best fit, although it does not follow well in the tail portion.

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 hugely important because it can affect strategies for marketing, target advertising, recommendation, and search engines. Some of the reasons for a truncated tail are as follows:

- **Natural shape**: User-generated content, by definition, varies widely in its quality. One may argue that the natural shape of the popularity distribution of UGC is truncated (e.g. log-normal), since significant fraction of videos in UGC are of low interest to most users. For example, UGC is normally produced for small audiences (e.g. family members), as opposed to professionally generated content.

- **Sampling Biases or pre-filters**: The plot of Netflix in Figure 1 shows a sharp decay in the tail. This can be explained by sampling bias. Even though NetFlix provides an enormous online catalog of DVDs worldwide, their videos are a set of movies that are sampled from all the movies ever made; only a small portion of movies world-wide are made into DVD titles. In UGC services, publishers post videos sampled from the video pool in their possession. Obviously, the sampling is biased toward interesting ones. The following explains the effect of pre-filters: From a complete list of N videos, whose popularity distribution follows Zipf, remove h videos such that the probability of a video removed is proportional to the inverse order of their ranks. Then, the remaining $N - h$ videos will have truncated tail.

- **Information filtering or post-filters**: Search or recommendation engines typically return or favor a small number of popular items [15, 36], steering users away from unpopular ones and creating a truncated tail. This truncation is more apparent over time since old non-popular videos are exposed longer to such post-filtering. Indeed, we are able to observe this in our traces. Figure 5(b) shows the popularity distributions of `Sci` videos of different ages. Videos aged 1 day are clearly less truncated in the tail than older ones.

If Zipf were to be the natural shape and the truncated tail was due to some removable bottlenecks (e.g. post- or pre-filters), then in the system with no bottleneck, the videos in the truncated region would gain deserved views, offering the better chances to discover rare niche videos to users and potential business opportunities to the company. We next estimate the potential benefit from the removal of such bottlenecks. The estimation is calculated as the ratio of aggregated additional views against the existing total views. Table 3 shows the measured benefits for the four UGC video categories. We also present the number of videos that may benefit. 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. While in Daum `Travel` and `Food`, the benefit is reduced due to a small number of videos that benefit. When the number of videos is small, the inefficiencies of the system (due to filtering effects) are smaller since information can be found easier.

Table 3: Potential gain from the Long Tail: additional views and the number of beneficiary videos

	`Ent`	`Sci`	`Travel`	`Food`
Gain	45%(1.2M)	42%(240K)	4%(5K)	14%(400)

Yet, such benefits may not hold if the truncation appears as a result of a natural user behavior. Interestingly, 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 behavior. While Zipf (so as power-law) is *scale-free* in nature, exponential is a distribution that is *scaled* or *limited* in size. Therefore, the two 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 cut-off reveals filtering effects in the system which truncates the tail. Nevertheless, revealing the true nature of the truncated tail calls for further in-depth studies.

4. POPULARITY EVOLUTION OF UGC

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 a much more 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 much more diluted in UGC. Videos come and go all the time, and the viewing patterns also 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 focus, in this section, 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 fast or slow popularity changes for videos of different age, and further test if the future popularity of a video can be predicted. For the analysis, we use daily trace of YouTube `Sci` videos.

4.1 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. Figure 6 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 very young videos (e.g. newer than 1 month), we observe slight increase in the average requests, which indicates viewers are mildly more interested in new videos, than the rest. However, this trend is not very pronounced when we examine the plot of maximum requests. Some old videos too receive significant requests. In fact, our trace shows massive 80% of videos requested on a given day are older than 1 month and this traffic accounts for 72% of total requests. The plot becomes noisy for age groups older than 1 year, due to small number of videos. In summary, if we exclude the very new videos, user's preference seems relatively insensitive to video's age.

While user's interests is video-age insensitive in a gross scale, the videos that are requested the most on any given day seem to be recent ones. To further verify this, we look into the age distribution of top twenty most requested videos. Figure 7 shows the result for a different time-window of 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 views of videos are zero. Over a one day period, roughly 50% of the top twenty videos are recent. However, as the time-window increases, the median age shifts towards older videos. This suggests ephemeral popularity of young videos. To better understand its effect, in the following, we discuss the video popularity evolution over time.

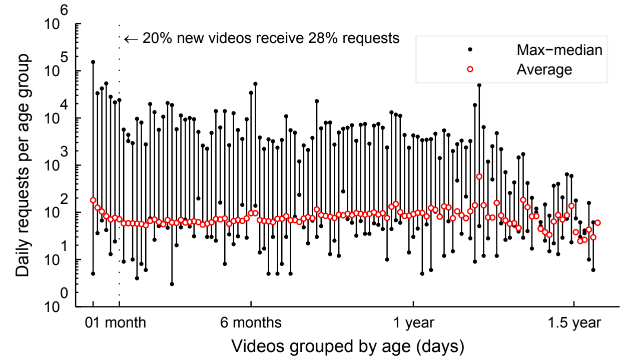


Figure 6: Distribution of request volume across video's age, based on `Sci` daily trace.

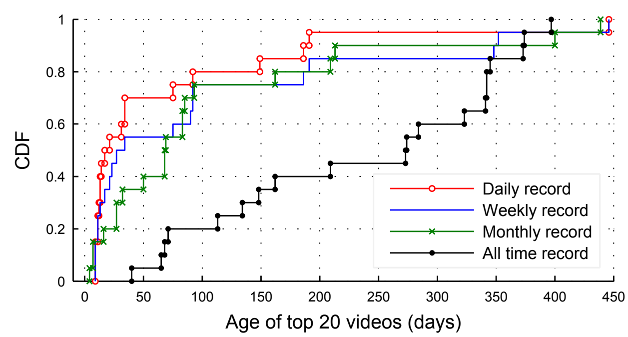


Figure 7: Age distribution of top 20 videos

4.2 Temporal Focus

We now continue our discussion on the video popularity and investigate how the popularity of individual UGC videos evolve over time, how fast or slow it changes, and whether the future popularity of a video can be predicted.

4.2.1 *Probability of Videos Being Watched Over Time*

When a video is posted, it has zero views; gradually videos will gain views over time. To capture this trend in UGC videos, in Figure 8, we show the percentage of videos aged $\leq X$ days having $\leq V$ 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 are watched at least once, and 40% are watched over 10 times. After a longer period of time, more videos gain views, as expected. One noticeable trend in the graph is the consistent deeps at certain times (e.g. 1 day, 1 month, 1 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 horizontal axis, this indicates the probability of a given video to be requested decreases sharply over time. In fact, if we consider the case of

$V = 10$, the probability that a given file gets more than 10 requests over the duration of 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 enough requests during its first days, then, it is unlikely that they 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.

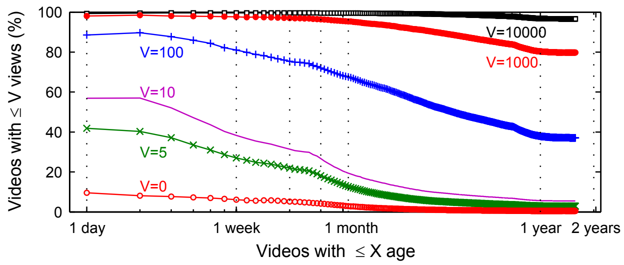


Figure 8: Probability of videos being watched over time, based on YouTube `Sci` trace

4.2.2 *Predicting Near-Future Popularity*

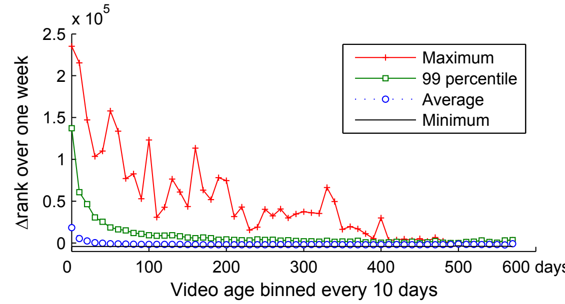
The ability to predict future popularity is immensely useful in many ways, because the service providers may prepopulate these videos within multiple proxies or caches and the content owners may use this fast feedback to better manage their contents (e.g. production companies releasing trailers to predict popularity). We now explore the possibility of using early views records in predicting near-future popularity. We compare the first few days' video views with those after some period of time (i.e. 5, 7, and 90 days). Table 4 shows the correlation coefficient of views for combinations of snapshots. We also present the number of videos used for sampling. Our results show that second day record gives an accurate estimation with a relatively high accuracy (correlation coefficient above 0.8). Using the third day record improves the prediction accuracy, yet, only marginally. Our results also show a high correlation with the second day record even for more distant future popularity (e.g. three months afterwards).

Table 4: Correlation coefficient of video views in two snapshots and the number of videos analyzed

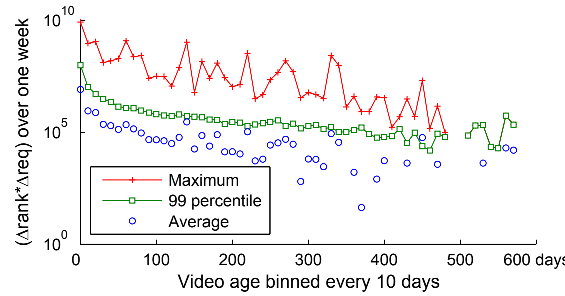
Age (x_0)	x_0+5 days old	7 days old	90 days old
2 days old	0.9665 (5185)	0.8793 (3394)	0.8425 (11215)
3 days old	0.9367 (3394)	0.9367 (3394)	0.8525 (9816)

4.2.3 *Popularity Shifts*

Now we examine how easy or hard it is for new and old videos to 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 Figure 9(a), we use two snapshots from our daily traces of six consecutive days, taken at day zero and day 5, and consider only those videos that appear on both of the snapshots. We group videos by their age (bin in units of ten days) and plot the change in



(a) Popularity distribution based on $\Delta rank$



(b) Popularity distribution based on $\Delta rank \cdot \Delta views$

Figure 9: Changes in ranking and popularity

ranks (i.e. $\Delta rank$) over age. For each age group, we plot the maximum, top 99 percentile, average, and the minimum change in $\Delta rank$. 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.

We make several observations from Figure 9(a). First, 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 effects and good recommendation engines. However, it is hard to conclude this from Figure 9(a) since a few requests may also result in major rank changes. We will revisit this issue at the end of this section.

The gap between the maximum and the top 99 percentile lines reflects that only a few young videos (e.g. less than 1%) make large rank changes, indicating that only a very small percentage of the young videos make it to the top popular list while the rest have much smaller ranking changes. We also see a consistent minimum $\Delta rank$ line at nearly -4000 across all age group. A detailed look at those videos reveals that those videos did not receive any request during the trace period, however their ranking was pushed back as other videos got at least one request. This shows that unpopular videos that do not receive any request will die in the ranking chart at a rate of 2000 positions per day.

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 large number of requests, we propose using the product of $(\Delta rank \cdot \Delta views)$ as in Figure 9(b). Please note that as opposed to Figure 9(a) the vertical axis now is in log scale. Now we observe more drastic popularity shifts for young videos; barely no single old video received a significant number of requests to make major upward shift in the popularity distribution. In short, revival-of-the-dead effect, where old videos are suddenly brought up to the top of the chart, does not happen strongly in our trace.

5. EFFICIENT UGC SYSTEM DESIGN

With the increasing popularity of UGC, YouTube alone is estimated to carry astonishing 60% of all videos online, serving 100 million distinct videos daily [6]. This corresponds to, in our estimation, massive 50 - 200 Gb/s of server access bandwidth on a traditional server-client model. Accordingly network operators are reporting a rise of overall Web traffic and HTTP video streaming [7].

In this section, we explore the benefits for alternate distribution schemes, namely, caching and peer-to-peer (P2P). We provide a rough estimate on the potential savings that caching and P2P can provide to the YouTube servers. Yet, our results are optimistic upper bounds for the benefits that one could expect in real deployment. Throughout this section, we use daily traces of six consecutive days for $263,847$ YouTube `Sci` videos. Our study does not include the network impact of various distribution schemes in ISPs, since our data does not contain geographical locations of the users.

5.1 Better Use of Caching

Caching stores redundant copies of a file near the end user and has been proven to be extremely effective in many Web applications. Several factors affect the caching efficiency: the cache size, the number of users and videos, the correlation of requests, the shifts in popularity, and so on. Here, we hypothesize a virtual global cache system for YouTube and assess with real trace how many hits on YouTube servers can be eliminated. Such cache could be deployed centralized or fully-distributed. In either case, we assume that all user requests are distributed across the global caching system and that caches cooperate, redirecting users to the right video copy. However, we do not make any assumptions about the exact location of caches or their number. Our interest, instead, is at investigating the global cache performance from the server's point-of-view, under massive new uploads and dynamic popularity evolution.

To this extent, we consider the following three conventional caching schemes:

1. A *static finite* cache, where at day zero the cache is filled with long-term popular videos. The cache content is not altered during the trace period.

2. A *dynamic infinite* cache, where at day zero the cache is populated with all the videos ever requested before day zero, and thereafter stores any other videos requested during the trace period.

3. A *hybrid finite* cache, which works like the static cache, but with extra space to store the daily most popular videos.

We populate the static cache with long-term popular videos accounting for 90% of total traffic. This corresponds to 16% of `Sci` videos as in the Pareto Principle (see Figure 2). The dynamic infinite cache simply stores all the videos ever requested. The hybrid finite cache is first populated with the top 16% of `Sci` videos, then the cache also allocates small extra space to store the daily top 10,000 videos.

We perform a trace-driven simulation to assess the cache performance in terms of the required cache size and the cache miss ratio. To do this, we replay the 6-day trace under our three caching schemes and calculate the average hit and miss ratios. We simply use the number of videos cached as the cache size, because the video length and the encoding rate do not vary much across files. We further assume that each time a video is viewed, the full video is stored in the cache (even when the user watched it partially)[3]. Table 5 summarizes the cache performance averaged over the 6-day period. The results indicate that about 40% of the videos that are requested daily are different from the long-term popular videos. However, the corresponding number of requests toward those videos accounts for only about 20% of the total requests. In fact, we can see that a simple static cache that stores the top long-term popular files uses 84% less space than a dynamic infinite cache solution which stores all videos, and still manages to save about 75% of the load in the server. It is worth noting that only about 2% of videos that are requested every day are newly uploaded ones. We should also mention that, by storing the most popular daily requests in addition to the long-term popular videos, a hybrid cache improves the cache efficiency by 10%, compared to the static cache.

Table 5: Synthetic cache efficiency

Type	Size	# missed videos	# missed requests
Static	41,235	115,002 (48.8%)	5,093,832 (26.7%)
Dynamic	256,647	4,683 (1.9%)	648,376 (3.4%)
Hybrid	51,235	94,893 (40.3%)	3,271,649 (17.1%)

5.2 Potential for Peer-Assisted VoD

Now we explore the potential benefits of a P2P technique in UGC distribution based on real trace. We consider a peer-assisted VoD distribution where users stream videos from VoD servers as well as from other online users (or peers). Typically, peers share videos they have watched for a certain period of time. Inherently, a P2P system is effective only when there are enough number of online peers sharing content – this is called a *torrent*. The efficacy of P2P for massive content distribution has been studied in other application [22]. Here we investigate the potential benefits that a P2P technique can bring to YouTube. We first assess the feasibility of peer-assisted VoD distribution of UGC by examining how many files benefit from such an approach. We then perform a trace-driven analysis to measure how much server workload can be lowered when peers assist video delivery, compared to the traditional server-client model.

[3] In fact many caches already download the entire requested object regardless of whether the user terminates the connection early. Such proactive caching strategy can serve future requests more rapidly.

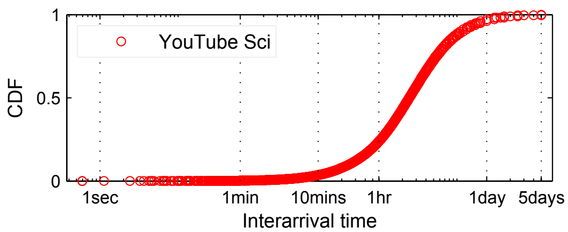


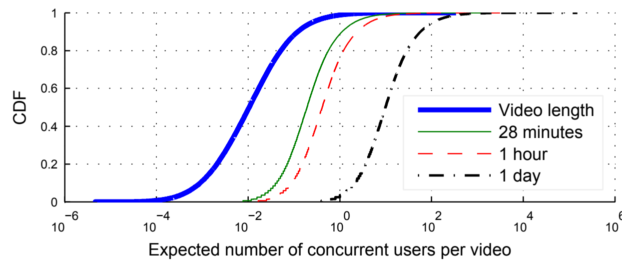
Figure 10: Inter-arrival times of requests

We commence by estimating the inter-arrival times of requests. Our trace provides temporal granularity of requests by the day. Within a single day, we assume that requests are exponentially distributed. Then, the inter-arrival time of requests has a mean of $\frac{1}{\lambda}$, where λ is the intensity of requests (i.e. the number of requests made that day). We later use the inter-arrival time to calculate the number of concurrent online users. Figure 10 shows the CDF of the average inter-arrival time per video. We observe that roughly 95% of the videos are requested once every 10 minutes or longer. This implies that the fraction of files that can benefit from P2P is very small, as most files are requested infrequently.

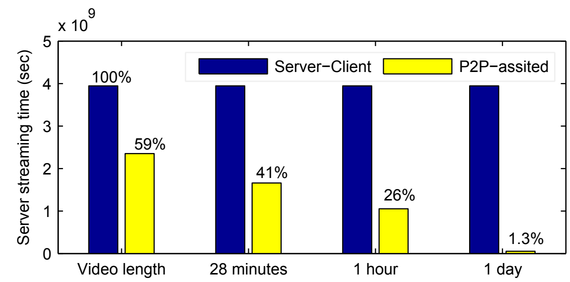
Next we calculate the number of concurrent users per video (e.g. the torrent size). We again assume that users watch the entire video and the users may share the file immediately after they start downloading[4]. The torrent size here depends on how long the users stay on the YouTube site (i.e. session or sojourn time) and how often they visit YouTube (i.e. frequency of sessions). The session time of a user is important because the P2P sharing may happen only when the user is online. We consider the following four P2P session times: 1) a user shares while watching a video (i.e. video length), 2) shares for the duration of time he spends on YouTube, 3) shares for one extra hour after he is done watching, and 4) shares for one extra day. According to Nielsen/NetRatings [13], the average session time of YouTube users is currently 28 minutes. We hence assume users share videos for 28 minutes in our second case. In the last two cases, we consider users sharing videos even when they are no longer in the system. We mention that this may become a reality in the future (e.g. users equipped with always-on set-top boxes that run P2P).

Then for a given P2P system time of a user, t, and the inter-arrival time of requests, $\frac{1}{\lambda}$, the expected number of concurrent users is λt. Note that this value can be less than 1, indicating that there are times within that day with no users watching the video. We consider the P2P approach only when λt is greater than one (i.e. more than one user watched a video). When $\lambda t \leq 1$, we simply apply the traditional server-client model. Figure 11(a) shows the CDF of the average number of concurrent users over the monitoring period per video. We observe that for most of the cases the average number of concurrent users, λt, is less than 1, indicating that only a few videos are to benefit from P2P. However, when users share videos for a longer period of time

[4]We assume that all peers watching the same video have useful data to share with other peers. While we do not discuss in detail how to achieve this efficiency, we note that such a P2P-VoD swarming protocol is feasible (e.g. using coding, proper gossiping, and overlay mesh construction) [25, 33].



(a) The number of concurrent online users



(b) Server workload savings against server-client model

Figure 11: Potentials of a P2P system

(e.g, 1 day), P2P may assist 60% of videos with at least 10 current users all the time.

While the number of files that can benefit from P2P come out relatively small, this does not necessarily mean P2P is inefficient for UGC. As we have seen from the previous sections, UGC requests are highly skewed and temporal. Therefore, we investigate the benefits of P2P by comparing the estimated server workload between traditional client-server and P2P-assisted distribution approaches. In a client-server model, each request is directly served by the server. While in the P2P-assisted model, peers will participate in streaming only when there are concurrent users. As a measure of server workload, we use the total length of the streamed content. Figure 11(b) compares the server workload based on trace-driven analysis. Our results show that the potential of P2P is actually very large. The server workload is reduced by 41% even when users share only videos while they are watching. When users share videos for one day, the server workload reduces by a tremendous 98.7%, compared to a client-server approach.

6. ALIASING AND ILLEGAL UPLOADS

Content aliasing and illegal uploads are critical problems of today's 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 in various system's characteristics.

6.1 Content Aliasing

Traditional VoD services offer differently encoded versions of the same video, typically to support diverse downward 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 from a single view count nor present users with unique videos, instead of numerous identical copies.

To estimate the prevalence of aliases, we have conducted the following experiment. We first sample 216 videos from the top 10,000 videos of YouTube `Ent` category. Then we ask 51 volunteers to view and familiarize themselves with some of those videos. After viewing some from our sample set, volunteers search YouTube using keywords of their choice and flag any videos that deem pertaining to the same event as aliases[5] Our volunteers have identified 1,224 aliases for 184 videos out of original 216. Most videos have 1 to 4 aliases, while the maximum number of aliases is 89. Out of all videos that pertain to the same event, we call the video with the earliest upload time *original*.

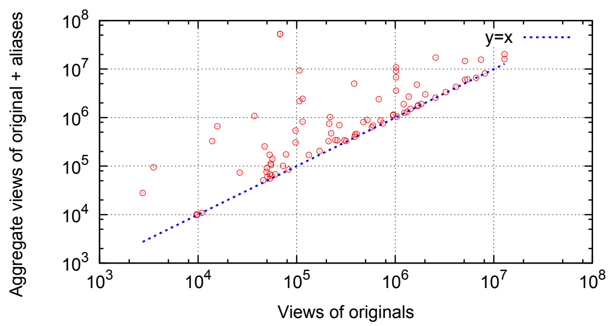


Figure 12: Sum of all views of the original and aliases versus views of original videos

Figure 12 shows the sum of views from all aliases and the original video against the number of views of the original videos. For a few videos, the sum of views from aliases grows more than two orders of magnitude than the views of the original. This clearly demonstrates the popularity dilution effect of content aliasing. Undiluted and augmented by the views of aliases, the original video could have been ranked much higher.

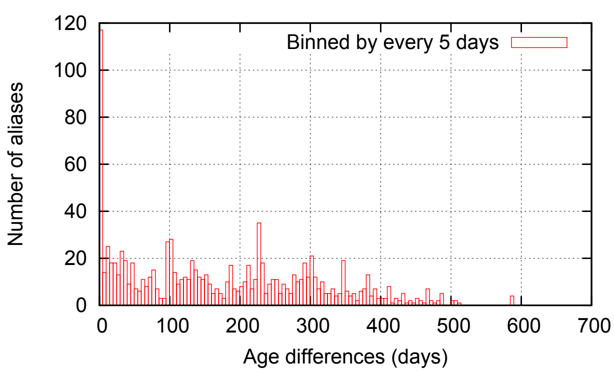


Figure 13: Number of aliases versus the age differences

[5]We have created a webpage `http://beta.kaist.ac.kr` for volunteers to view the video along with the description, and then search for content aliases in YouTube.

Next, we analyze the time intervals between aliases. We plot the age differences between the original video and its aliases in Figure 13 (the bin size is 5 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 has changed, in Figure 14 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 1000 times more views. This particular video was originally listed in the Music category, and later posted on the Comedy category with much more views. We find it rather surprising to see so many aliases still appear 100 or more days after the original video. They are also found to belong to different categories from the original and have been cross-posted over multiple categories. These aliases could be a potential reason for the flattened popularity tail. We leave further investigation behind this delayed popularity for future work.

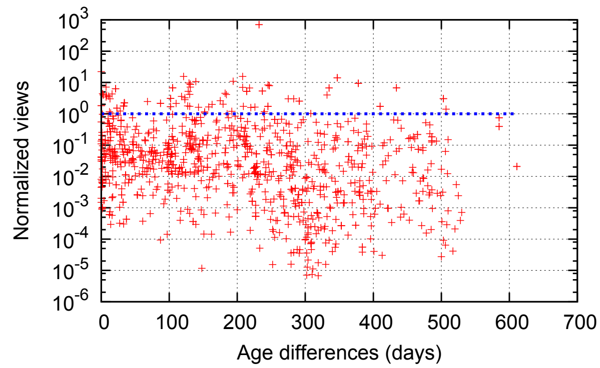


Figure 14: Normalized views versus the age differences

The Pearson correlation coefficient of the plot in Figure 14 is 0.004. It signifies little correlation or no decrease in the number of views over time. With a good number of aliases older than 100 and more views, we discern no clear trend in the aliases and their views over time. Those aliases that turn up 100 days later with much fewer views are likely to serve personal archiving purposes.

Finally, we check for the existence of heavy alias uploaders. Suspecting their strong motivation for online popularity, we have wondered if they could post aliases of already popular videos. 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.

6.2 Illegal Uploads

UGCs derived from copyrighted contents raise a serious legal dilemma for UGC service providers. In a sense, aliases can be considered to a great extent as a form of "video spam." A recent study from Vidmeter [28] suggests that nearly 10 percent of videos in YouTube are uploaded without the permission of the content owner. Vidmeter's report cover only the top ranked UGCs. We augment Vidmeter's work by looking not only at the top ranked videos, but all in `Ent`.

We get the list of all videos at two different times, and compare the two lists. The discrepancy represents the deleted videos. When we follow the links to the 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. From the first set of videos (1, 687, 506), the number of all deleted videos are 6, 843 (0.4%). Only about 5% of deleted videos have violated the copyright law, which is a far smaller number than Vidmeter's.

7. RELATED WORK

We have already incorporated many of the references that closely relate to our work throughout the paper. As our work covers a broad spectrum of topics from popularity analysis to Web caching and P2P streaming, in this section, we briefly summarize the related work on each topic.

VoD service has become extremely popular in the Internet. Especially the demand for user-generated contents has grown explosively. Among the numerous UGC sites, YouTube, MSN, Google Video, Yahoo! Video, and UnCut Video are the notable ones. Despite the excitement, relatively little attempts have been made to understand how these UGC services are fundamentally different from traditional well-explored video distribution services [23]. In contrast, much has been written about traditional VoD services. The first kind of studies is by Griwodz *et al.*, where they use off-line video rental records to study video popularity [24]. Recently, Yu *et al.* conducted an in-depth analysis of access patterns and user behaviors in a centralized VoD system [39].

In the study of popularity distributions, Newman carried out a good comprehensive study of power-law distributions [37]. He examined several examples of power-law: Web hits, copies of books sold, telephone calls, etc. Also a paper by Alderson *et al.* develops an interesting and rich theory for scale-free networks [31]. Power-law distribution with a truncated tail has frequently appeared in the degree distributions of various real-world networks such as WWW, protein networks, e-mail networks, actor networks, and scientific collaboration networks [16, 20, 36].

The concept of peer-assisted video streaming has been extensively explored [14, 22, 27, 29]. Most existing work concentrate on the protocol design under various topological constraints [18, 27]. Our study considers the potential for P2P delivery in large-scale UGC systems, which have unique characteristics in terms of user consumption patterns and video popularity distributions.

8. CONCLUSIONS

In this paper we have presented an extensive data-driven analysis on the popularity distribution, popularity evolution, and content duplication of user-generated video contents. To the best of our knowledge, this work is the first major stab at understanding the explosive growth of UGC and its implications on underlying infrastructures.

We have studied the nature of the user behavior and identified the key elements that shape the popularity distribution (e.g. what shapes the Long Tail, alters the skewness of popularity, or breaks the power-law behavior for very popular contents). Our results indicate that information filtering is the likely cause for the lower-than-expected popularity of niche contents, which if leveraged, could increase the total views by as much as 45%.

We have studied different UGC cache designs, and showed that simple policies that cache the most popular contents can offload server traffic by as much as 50%. Similarly, we have also demonstrated that a distribution system based on a P2P system can have great benefits, despite the diversity of requests and short video length.

Finally, we have tackled the impact of content aliasing and illegal uploads, which could hamper the future success of UGC services. Content aliasing is widely practiced and makes the video ranking difficult. Illegal uploads are more common amongst highly ranked videos. We believe that our work answers very critical and pressing questions, and lies the basis for the design of future UGC systems. Our dataset has been mainly focused on snapshots obtained from two large UGC systems. It will be interesting to see how our analysis results hold in the future and across other UGC systems.

Acknowledgement

We thank Carlos Domingo and all the members at Telefonica Research, Don Towsley, Jon Crowcroft, Christos Gkantsidis, Thomas Karagiannis, anonymous reviewers, and our shepherd, Reza Rejaie, for their valuable comments.

Meeyoung Cha did this work as an intern at Telefonica Research, Barcelona. She was partially supported by the Brain Korea 21 Project through KAIST. Haewoon Kwak, Yong-Yeol Ahn, and Sue Moon were supported by grant No. R01-2005-0001112-0 from the Basic Research Program of Korea Science & Engineering Foundation.

9. REFERENCES

[1] Daum UCC. `http://ucc.daum.net`.

[2] Imdb statistics. `http://www.imdb.com/database_statistics`.

[3] Lovefilm. http://www.lovefilm.com.

[4] Netflix prize. http://www.netflixprize.com.

[5] Yahoo! Movies. `http://movies.yahoo.com`.

[6] YouTube. `http://www.youtube.com`.

[7] Surveys: Internet Traffic Touched by YouTube, January 2006. `http://www.lightreading.com/document.asp?doc_id=115816`.

[8] L. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of Small-World Networks. In *Proc. Natl. Acad. Sci. USA*, 2000.

[9] C. Anderson. A Problem With the Long Tail. `http://www.longtail.com/scifoo.ppt`.

[10] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

[11] E. Auchard. Participation on Web 2.0 Sites Remains Weak, April 2007. `http://www.reuters.com/article/internetNews/idUSN1743638820070418`.

[12] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[13] S. Bausch and L. Han. YouTube U.S. Web Traffic Grows 75 Percent Week over Week, July 2006. Neilsen/Netratings, `http://www.nielsen-netratings.com/pr/pr_060721_2.pdf`.

[14] B. Cheng, X. Liu, Z. Zhang, and H. Jin. A Measurement Study of a Peer-to-Peer Video-on-Demand System. In *Proc. of IPTPS*, 2007.

[15] J. Cho and S. Roy. Impact of Search Engines on Page Popularity. In *Proc. of WWW*, 2004.

[16] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto. Analyzing Client Interactivity in Streaming Media. In *Proc. of WWW*, 2004.

[17] M. E. Crovella and A. Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM ToN*, 5(6):835–846, 1997.

[18] T. Do, K. A. Hua, and M. Tantaoui. P2VoD: Providing Fault Tolerant Video-on-Demand Streaming in Peer-to-Peer Environment. *Proc. of IEEE ICC*, 2004.

[19] A. B. Downey. The Structural Cause of File Size Distributions. In *Proc. of IEEE MASCOTS*, 2001.

[20] T. Fenner, M. Levene, and G. Loizou. A Stochastic Evolutionary Model Exhibiting Power-Law Behaviour with an Exponential Cutoff. *Physica*, (13), 2005.

[21] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical Interests and the Mitigation of Search Engine Bias. In *Proc. Natl. Acad. Sci. USA*, 2006.

[22] C. Gkantsidis, T. Karagiannis, P. Rodriguez, and M. Vojnovic. Planet Scale Software Updates. In *Proc. of ACM SIGCOMM*, 2006.

[23] L. Gomes. Will all of us get our 15 minutes on a youtube video?, *The Wall Street Journal Online*, August 2006.

[24] C. Griwodz, M. Biig, and L. C. Wolf. Long-term Movie Popularity Models in Video-on-Demand Systems. In *Proc. of ACM Multimedia*, 1997.

[25] S. Guha, S. Annapureddy, C. Gkantsidis, D. Gunawardena, and P. Rodriguez. Is High-Quality VoD Feasible using P2P Swarming? In *Proc. of WWW*, 2007.

[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. of ACM SOSP*, 2003.

[27] Y. Guo, K. Suh, J. Kurose, and D. Towsley. P2Cast: Peer-to-peer Patching Scheme for VoD Service. In *Proc. of WWW*, 2003.

[28] B. Holt, H. R. Lynn, and M. Sowers. Analysis of Copyrighted Videos on YouTube.com. `http://www.vidmeter.com/i/vidmeter_copyright_report.pdf`.

[29] C. Huang, J. Li, and K. Ross. Peer-Assisted VoD: Making Internet Video Distribution Cheap. In *Proc. of IPTPS*, 2007.

[30] Y. Ijiri and H. Simon. *Skew Distributions and the Size of Business Firms*. North Holland, Amsterdam, 1977.

[31] D. A. L. Li, J. Doyle, and W. Willinger. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. *Internet Mathematics*, 2(4), 2006.

[32] E. Limpert, W. A. Stahel, and M. Abbt. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, 51(5):341, 2001.

[33] N. Magharei and R. Rejaie. PRIME: Peer-to-Peer Receiver-drIven MEsh-based Streaming. In *Proc. of IEEE INFOCOM*, 2007.

[34] N. Miller. Manifesto for a New Age. *Wired Magazine*, March 2007.

[35] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[36] 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.*, (13), 2002.

[37] M. E. J. Newman. Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46:323, 2005.

[38] V. M. W. Gong, Y. Liu and D. Towsley. On the Tails of Web File Size Distributions. In *Proc. of 39th Allerton Conference on Communication, Control, and Computing*, 2001.

[39] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding User Behavior in Large-Scale Video-on-demand Systems. In *Proc. of ACM Eurosys*, 2006.

[40] G. U. Yule. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Royal Society of London Philosophical Transactions Series B*, 213:21–87, 1925.