This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011 proceedings

Measurements and Analysis of an Unconstrained User Generated Content System

Tianlong Yu   Chen Tian   Hongbo Jiang   Wenyu Liu
Dept. of EIE, Huazhong University of Science and Technology, China

***Abstract*—User-Generated Content (UGC) is overwhelming the Internet with its interactivity and various contents. However, traditional UGC still have constrains on videos' length and size, which block out a wide variety of potential popular contents. In this paper, we present the first experimental measurements and analysis of an Unconstrained User-Generated Content (UUGC) system—a test site (so-called "T" site in this paper) of a leading VOD service provider in China. This test site is a video-sharing portal just like traditional UGC, while its contents are not constrained by either duration or size. As an UUGC system, its most distinguishing characteristics are the various types of contents uploaded (movie, TV episode, TV show, music, documentary, sports, etc.) and the wide range of uploaders, which make it an interesting case study. By matching relative key words in video's index, we classify the contents into several basic types and analyze the statistics of three major types—movie, TV episode and TV show (labeled MVI, TV-E and TV-S). For further study of various contents, we demonstrate the patterns of flash crowd triggering of MVI, TV-E and TV-S with several typical cases. To find out the viewers' consumption pattern, we investigate daily&weekly cycles, as well as grouping the videos by age and exhibiting the popularity evolution. By means of curve fitting with multiple known distributions to video view traces, we show that power law with exponential cutoff best fits the videos' popularity distribution for this UUGC system.**

I. Introduction

User-Generated Content (UGC) is overwhelming the Internet with its interactivity and various contents, attributing to that UGC users are not only content consumers but also content providers. However, due to the upper bound of the video's length and size, video content in traditional UGC services (e.g., YouTube) is still constrained, blocking out a variety of potential popular contents such as movies (especially movie with high definition) and TV episodes.

In this paper, we present specific measurements and analysis of an unconstrained user generated content (UUGC) system—a test site (labeled T site) of a leading VOD services provider in China. Our data consists of run-time traces of all videos on T site from Nov. 21 to Dec. 31, 2009. Different from previous VOD services, T site is a video-sharing portal just like YouTube (traditional UGC), while its contents are not constrained by video length or size, similar to traditional VOD systems. Hence we name it unconstrained User-Generated Content (UUGC) system. As an UUGC system, its most special characteristics are the various types of contents offered (movie, TV episode, TV show, music, documentary, sports, etc.) and the wide range of providers, which make it an interesting case study.

The contributions of our work can be summarized as follows:

1) We classify the contents into several general types and analyze the statistics of the three major types (MVI, TV-E and TV-S).
2) We provide insights on UUGC's various contents by demonstrating the flash crowd triggering for different video types.
3) To find out the viewers' consumption pattern, we investigate daily&weekly cycles, as well as grouping the videos by age and exhibiting the popularity evolution. We show that power law with exponential cutoff best fits the videos' popularity among multiple known distributions.
4) Our work serves as a reference for future's UUGC related system design.

The rest of the paper is organized as follows. Section II discusses the related works. We describe our measurement methodology and various contents of T site in Section III. Section IV further discusses various contents by demonstrating the patterns of correlated videos and the flash crowd triggering for different video types. Section V analyzes the consumption pattern as well as modeling the video popularity. Section VI draws the conclusions.

II. Related Works

As an emerging system, few work has been done on UUGC. We are not aware of any UUGC measurement at the scale we consider in this paper. However, there are many studies focusing on "traditional" UGC or VOD systems.

Phillipa Gill et al. characterized the traffic of YouTube for a three month period [5]. While Meeyoung Cha et al. analyzed the consumption pattern and popularity distribution of YouTube and Daum [3][4]. In [2], a detailed measurement of a large scale live VOD system and an analysis on flash crowd phenomena is performed for the 2008 Beijing Olympics. All are measurements and analysis of "traditional" UGC or VOD systems on a large scale. Our work is complementary, in that it presents specific measurements and analysis of a large scale system of UUGC.

Video popularity modeling and characterization of video content as well as user behavior have also received considerable attention. [6] proposed a model for video popularity evolution that can tend to either a power-law or exponential decay law controlled by a parameter. [1] analyzed a VOD service with various digital newspaper contents and noticed unique characteristics of different content types, but failed

978-1-61284-233-2/11/$26.00 ©2011 IEEE

TABLE I
STATISTICS OF VIDEOS WITH VARIOUS CONTENTS

Category	Num. videos	Total views	Average views	Average lifetime	New release	Production rate
Movie	5145	26817744	5212.4	33.0	1911	0.3714
TV episode	72195	6654734	92.2	30.9	46538	0.6446
TV show	2275	322667	141.8	25.4	1760	0.7736
All	85174	34828512	408.9	30.1	55433	0.6508

to classify the contents. Hongliang Yu et al. analyzed user behavior, content access patterns, and their implications for a VOD system deployed by China Telecom [9].

III. Measurement Methodology And Various Contents

In this section, we introduce our data source and present the characteristics of various video contents on T site. We mainly focus on three types—movie, TV episode and TV show (labeled MVI, TV-E and TV-S), which cover the majority of the videos and occupy most views of T site. Each video type is distinguished by a naming rule when they are uploaded. So we are able to identify them by performing key symbols matching with Perl. Statistics in Table I shows that 93.47% of the videos in T site falls into these three categories, which occupies 97.03% of the total views. The relevant statistics of all the measured videos and the three main types are summarized in Table I. Fig. 1 illustrates the cumulative distribution function(CDF) of video popularity based on view counts.

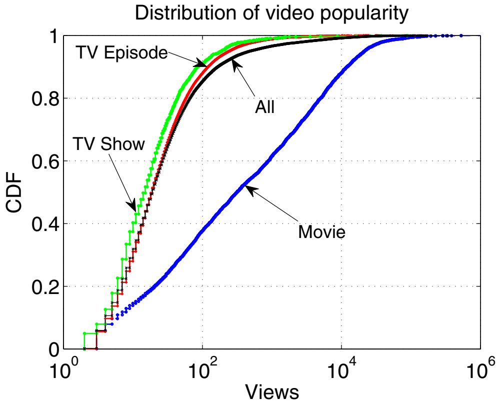


Fig. 1. CDF of video popularity with various contents

A. Data Source Overview

Our data consists of run-time traces of all videos on T site. We are able to access all four load balancer log servers (labeled LBLS) of T site. Each LBLS generates access logs that track altogether 851740 videos' request records in a time span of 41 days. We collect the log files from Nov. 21th to Dec.31th on LBLS. Each request record line of a log file consists of a specified requested video name and its request time. Irrelevant information such as request error processing is omitted, reserving only records of successful views. Since the transmission time between the content servers and the LBLS is ignorable, we consider no daily basis in our measurements[2].

Fig. 1 shows the CDF of views separately for all videos and for the three main types—MVI, TV-E and TV-S. By observing the CDF of all videos (black curve), we notice that it is in a stairs sharp at the initial part, indicating that several portions of videos receive the same few views. Besides, it raises to more than 0.8 even before view counts reaches 10^2. We can deduce that there are many unpopular videos (videos with few views) on T site. When comparing the CDF of MVI, TV-E and TV-S with each other, two observations can be made. First, the CDF of MVI spans nearly 5 orders of magnitude, while TV-E and TV-S only span 3 orders of magnitude, revealing the innate diversity of movie consumers. Second, the median number of MVI's CDF is much greater than that of TV-E and TV-S, indicating that MVI videos are more popular in general.

B. Movie Contents

MVI is obviously the most popular 20% on T site (80-20 rule in Section V-C). Covering merely 6.04% of total videos, it occupies 77.00% of the total views. So it is reasonable that MVI has the highest average views of 5212.4 views per video. On the other hand, MVI sees the lowest production rate of 0.3714 among the three, as 5145 videos only see 1911 new releases in the span of 41 days. This notifies that the popularity of MVI tends to be in long term (not likely to be quickly out of date) and is relatively stable. Another evidence for this notification is that, with an average value of 33.0 days, the lifetime of movie contents is the longest. Here we define the lifetime of a video as the duration from the time it is published to the time when it is put off the site [1].

C. TV Episode Contents

For the following two reasons, TV-E covers the largest number of videos. First, each episode series, though telling one story, is divided into dozens of episodes (several seasons also). Each episode is labeled as a different video and varies in popularity from each other. Second, TV-E engages a considerable amount of producers and audiences, especially in multi-culture frontiers like China (American, Japanese&Korea episodes, along with numerous local ones, occupy a wide range of viewers). TV-E is the least popular 80% of the videos on T site (80-20 rule in Section V-C), as it covers 84.76% of all videos while only receiving 19.11% of all views. The average views of the TV-E is also the lowest—92.2 views per video. The production rate of TV-E, 0.6446, is much higher than movie. This may due to the fact that one episode, not a series, is much easier to produce than one movie. Yet it is still lower than TV-S. The average lifetime of the TV-E, 30.9

[1] This is marked as the video's last viewpoint in the data set.

TABLE II
FLASH CROWDS CASES OF VARIOUS CONTENTS

Date	Video	Category	Event	Rush hour	Daily views contribution
Nov.23	Movie A(HD ver.)	All	Third day of release; highest peak	20:00-23:59	6.96%
Nov.27	Movie B	Movie	Fourth day of release; highest peak	20:00-23:59	1.47%
Dec.12	Episode P	Episode	First day of release; highest peak	23:00-6:00	16.98%
Dec.5	TV Show U	Show	First day of release; second day after broadcast	20:00-23:59	29.56%

days per video, ranks in the middle of the three. As episodes of one series are related with each other, TV-E is more likely to be reviewed than TV-S.

D. TV Show Contents

TV-S covers the least portion of total videos when comparing with movie and TV-E, but it keeps the highest production rate of 0.7736. Among 2275 TV-S videos in the span of 41 days, surprisingly 1760 are new releases. Meanwhile, TV-S's average lifetime, 25.4 days per video, is the shortest. This notifies the common pattern of TV-S—videos of TV-S become very hot soon after their broadcast on TV, then they quickly go out of date. Such a pattern triggers obvious flash crowds like phenomena of TV-S.

IV. Case Studies Of Flash Crowds Triggering

Flash crowds is a phenomenon that a large number of viewers flood in to watch one or a few specific videos, this phenomenon is often triggered by some certain events. Flash crowds is worthy of our concerns, in that it can give rise to large workload and severely disturb the UGC portal. By presenting typical cases of major video types, we provide insights on various contents and system design of UUGC. In this section, we illustrate case studies of various contents for what triggers the flash crowds.

Although traditional VOD and UGC systems do not commonly exhibit flash crowd like phenomena, some video types on T site (e.g., TV-S) shows explicit features of flash crowd. To understand what kind of videos and specific events trigger the flash crowds, we search day by day for rush hours with the most views of the day. Since the magnitude of videos' views in different types differ too much, the rush hour measurement is done separately within different types (movie, TV-E, TV-S and all), so that typical flash crowd featured videos are better identified. For each rush hour, we pick up the top-20 videos in each type and manually find the most common real-world event that may trigger the flash crowd. In this way, we annotate each flash crowd.

In TABLE II, we list four typical cases to illustrate the properties of flash crowds triggering. We record all four flash crowds' occurring date, main cause video, category, event that triggers it, rush hour and daily views contribution to the category. The cases of Movie A (HD ver.) and Movie B illustrate that the flash crowd for long term videos (most MVI) usually arrives at the third or fourth day after release.[2] But for short term videos (episode, show, the third and fourth row), flash crowd is more likely to flood in as soon as the video is released. By the cases of Episode P (TV-E) and TV Show U (TV-S), flash crowds of TV-E and TV-S contribute a greater increment to daily views by percentage (both exceed 10%), indicating that TV-E and TV-S tend to be more attractive to flash crowd. Specially, the flash crowd of TV-S is quite sensitive to broadcast time. Commonly, rush hours arrive between 20:00 and 23:59.

V. Video Consumption Pattern And Popularity Distribution

In this section we focus on the video consumption pattern and the popularity distribution of T site. These are two major aspects of videos' characteristics which answer important design questions of the system. Video consumption pattern reflects video popularity evolution with aging. Popularity distribution reveals the difference in popularity among videos. In the following part, Section V-A, Section V-B illustrate the video consumption pattern, while Section V-C, and Section V-D discusses the popularity distribution.

A. Daily And Weekly Cycle

Daily and weekly cycle reflect the variations of viewers' consuming pattern.

1) Daily Cycle: Fig. 2 displays 24 hours' average views of all videos. 7:00 sees the minimum views, then it begins to raise across the daytime until reaches a local maximum at 16:00. By 18:00 the popularity raises again until it reaches the global maximum at 22:00, then it continually decreases across midnight until 7:00. Hours from 20:00 to 23:00 receive the most views, while 04:00 to 08:00 receive the least. The difference can be as much as tenfold.

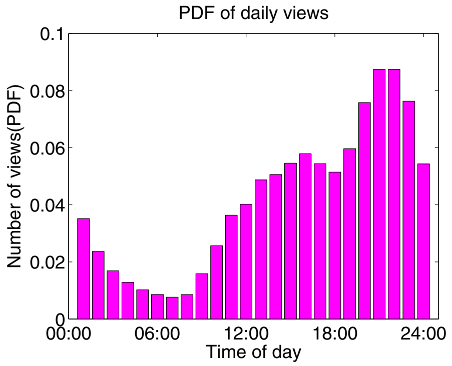
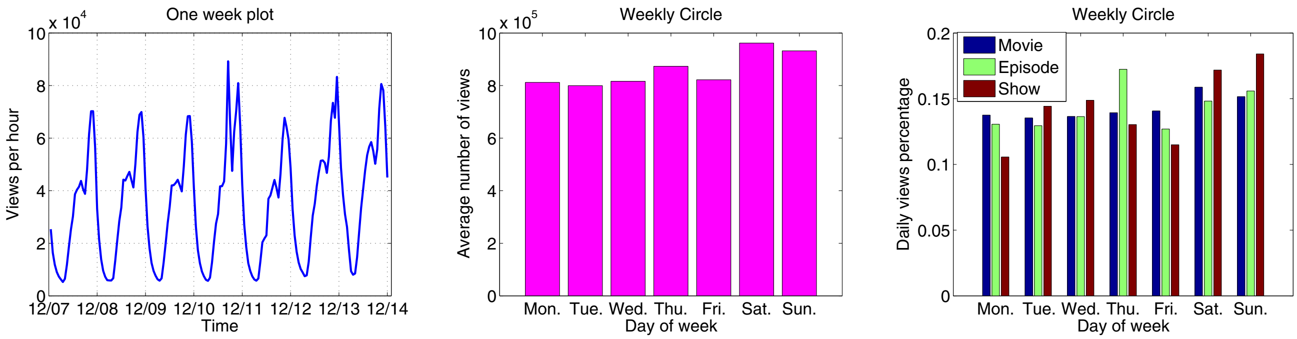


Fig. 2. Daily cycle

[2] Other similar experiments also support the conclusions, but limited by space, we just list four cases here as illustrations.



(a) Hourly views of a sample week

(b) Weekly cycle

(c) Weekly cycle with various content

Fig. 3. Weekly cycle

2) Weekly Cycle: Fig. 3(a) displays hourly views of T site in a sample week from Dec. 7th to Dec. 13rd. We notice that the difference of hourly views can be as much as fifteenfold, much greater than the threefold of Diggs[4], another UGC portal with a low reviewing rate. A low reviewing rate implies that viewers are likely to pile up in a short duration. Yet still T site sees a greater difference, indicating that daily and weekly cycle play a more important role in T site's consumption pattern than Diggs. Fig. 3(b) illustrates the average number of views from Monday to Sunday in a 5-week duration from Nov. 23rd to Dec. 27th. By the graph, Saturday and Sunday receive the highest number of views, followed by Thursday. Fig. 3(c) shows that movie and TV-S receive the highest number of views on Saturday and Sunday, while TV-E's peak arrives on Thursday.

In summary, the viewers of T site are more active on Saturday and Sunday, while YouTube on Tuesday and Wednesday, and Daum (a Korea VOD portal) on Sunday and Monday[5]. This difference occurs as the region of viewers differs, T site in mainland China, Daum in Korea and YouTube worldwide. The possible reason for Chinese viewers to be more active on Saturday and Sunday can be that, Chinese prefer to surf UGC sites at the weekends when they are enjoying their spare time at home.

B. Video Consumption Evolution

To demonstrate how the videos' popularity evolves with aging, we group the videos by age(binned by day) and calculate the maximum, median and average views for each age group. Fig. 4 displays the result. The vertical axis is in log-scale. The graph shows that, apart from small fluctuations, the maximum views tends to be consistent, indicating that popular videos scatter in all age groups evenly. Meanwhile, the median and average views are decreasing on the whole, illustrating that videos with shorter lifetime receive higher daily views. In general, viewers' consumption pattern seems relatively sensitive to video age[3] in short term(41 days). Recently published videos receive higher daily views than older ones. This is quite different from YouTube's conclusions for long term[3].

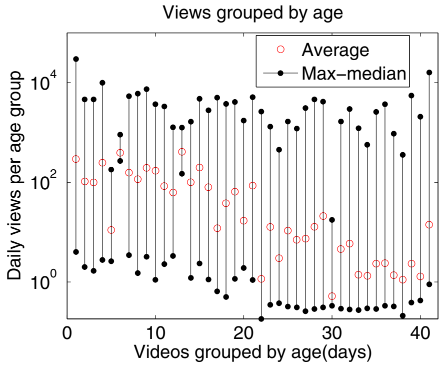


Fig. 4. Popularity evolution with age

C. Pareto Principle

The Pareto Principle (or 80-20 rule), is the most frequently applied rule to describe the skewness of user interest distributions. To test if the Pareto Principle can be applied to the actual views of T site, we calculate the fractions of cumulative views as the upper bound of cumulative set increases from the least popular ones to the most. Fig. 5 displays the result in a complementary manner, with the x axis being the normalized video ranks[4] and y axis being the fraction of cumulative views. The result shows that 20% of the top popular videos counts for 96.3% of views, while the rest 80% of videos counts for only 3.7% views, moreover, 80% of views is contributed by merely 1.747% of videos.

This result is significant, since other online systems show much smaller skew. Even for YouTube (UGC), which is prone to skewness, 10% of videos is still required to contribute 80% of views[4]. This skewness property illustrates some immediate applications in the system design. For instance, the efficiency of caching can be greatly improved, since by storing only 2% of the popular videos, a server can serve 80% of the requests.

D. Power-Law Test

In this part, we perform graph fitting to actual video views with multiple known distributions, in order to infer

[3] Video age counts from the publish time of the video.

[4] Sorted from the most to the least; ranks normalized between 0 and 100.

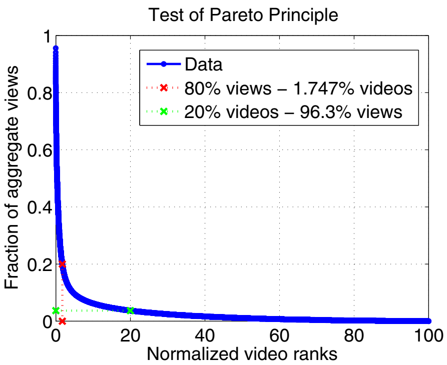


Fig. 5. Test of Pareto Principle

the intrinsic properties of UUGC. Determining the shape of video popularity distribution is a vital problem in discussing a VOD system's statistic properties, because it answers important design questions for the system. The shape of the distribution reflects underlying mechanisms that generate it and exhibits the system's distinguishing features. We mainly focus on two models that are widely applied in modeling web traffic in today's computer science—Power-Law model (with exponential cutoff) and Log-Normal model.

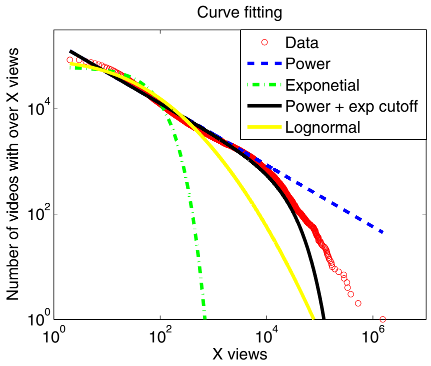


Fig. 6. Popularity distribution test

Fig. 6 demonstrates our graph fitting to actual video views by power law, exponential distribution, power law with exponential cutoff and log-normal distribution. The graph illustrates that power law with exponential cutoff best fits our data. Power law is in a straight line in the log-log plot of views versus frequency. But when a exponential cutoff is added, the shape then turns into a straight line shape in the waist part of the distribution, followed by a curved tail. Comparing with YouTube, our data is better fitted by power law with exponential cutoff at the straight line part[3]. The possible reason is that UUGC removes constrains on videos' length and size, which matches the scale-free property of power law.

VI. Conclusions

In this paper, we present specific measurements and analysis on a test site of the emerging UUGC service. By matching relative key words in video's index, we classify the contents into several general types and analyze the statistics of the three major types—MVI, TV-E and TV-S. Results show that MVI contributes most views, TV-E covers most videos, while TV-S is flash crowds-featured.

For further study of various contents, we provide insights on UUGC's various contents by demonstrating the flash crowd triggering for different video types. For long term videos (e.g., MVI), flash crowds usually arrive at the third or fourth day after release, while for short term videos (e.g., TV-S), flash crowds flood in as soon as the video is released.

To find out the viewers' consumption pattern, we investigate daily&weekly cycles, as well as grouping the videos by age and exhibiting the popularity evolution. Our results reveal that recently published videos receive higher daily views than older ones. By performing graph fitting with multiple known distributions to actual video views, we show that power law with exponential cutoff best fits the videos' popularity distribution.

Acknowledgment

The authors would like to thank all the members of NEST research group of Dept. of EIE, Huazhong University of Science and Technology for many helpful advises, and the anonymous reviewers for many helpful comments. This work was supported in part through National Natural Science Foundation of China (No.60803115, No.60873127, No.61073147), National Natural Science Foundation of China - Microsoft Research Asia (No.60933012).

References

[1] Xabiel G.Paneda, R.Garcia, D.Melendi, M.Vilas and V.Garcia, "Popularity analysis of a video-on-demand service with a great variety of content types. Influence of the subject and video characteristics", In *Proc.of ACM AINA. 2006*

[2] H.Yin, X.Liu, F.Qiu, N.Xia, C.Lin, H.Zhang, V.Sekar and G.Min, "Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics", In *Proc.of ACM IMC, 2009*

[3] M.Cha, H.Kwak, P.Rodriguez, Y.-Y.Ahn, and S.Moon, "I Tube, YouTube, Everybody Tubes: Analyzing the Worlds Largest User Generated Content VideoSystem", In *Proc.of ACM IMC, 2007*

[4] M.Cha, H.Kwak, P.Rodriguez, Y.-Y.Ahn, and S.Moon, "Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems", In *IEEE/ACM Trans. on Netw. 2009*

[5] Phillipa Gill, Martin Arlitt, Zongpeng Li and Anirban Mahanti, "YouTube Traffic Characterization: A View From the Edge", In *Proc.of ACM IMC, 2007*

[6] Z.Avramova, S.Wittevrongel, H.Bruneel and D.D.Vleeschauwer, "Analysis and Modeling of Video Popularity Evolution in Various Online Video Content Systems: Power-Law versus Exponential Decay", In *INTERNET 2009*

[7] W.Tang, Y.Fu, L.Cherkasova, and A.Vahdat, "Modeling and Generating Realistic Streaming Media Server Workloads", In *Computer Networks, vol. 51, 2007, pp. 336-356.*

[8] L.Guo, E.Tan, S.Chen, Z.Xiao and X.Zhang, "The Stretched Exponential Distribution of Internet Media Access Patterns", In *PODC'08*

[9] Hongliang Yu et al., "Understanding User Behavior in Large-Scale Video-on-Demand Systems", In *EuroSys 2006*

[10] Siddharth Mitra et al., "Characterizing Web-based Video Sharing Workloads", In *WWW 2009*