Reproducing Popularity Distribution of YouTube Videos

Noriaki Kamiyama, *Member, IEEE,* Masayuki Murata, *Member, IEEE*

***Abstract*—To provide video streaming of user-generated contents (UGCs) with high quality and at low cost by maximizing the effect of CDN, CDN providers are required to adequately design CDN cache servers by accurately estimating the UGC view-count distribution. To achieve this goal in a practical time frame, we need to construct a simple time-series model that captures the transition of UGC popularity. Therefore, in this paper, we first analyze the daily view count (DVC) of YouTube videos over nine months and find that the DVC of YouTube videos obeys a lognormal distribution. As a simple time-series model of the DVC of each YouTube video, we propose the grouped MPP (gMPP), extending the multiplicative process (MPP) which is widely known as a simple time-series model generating a lognormal distribution. We also propose reproducing the DVC distribution of YouTube videos by using a superposed gMPP (SgMPP) that aggregating multiple gMPPs. The SgMPP can accurately reproduce the DVC distribution of YouTube videos with a low computational overhead, so we can expect to use the SgMPP as the input for computer simulations for designing various network components that require the popularity distribution of UGC, e.g., cache capacities.**

***Index Terms*—popularity distribution, reproduction, multiplicative process**

I. Introduction

Services that stream user-generated content (UGC) have been spreading widely on the Internet. Many UGC streaming services use content delivery networks (CDNs), which deliver content from cache servers deployed at the edge nodes of the networks close to the requesting users [4][32][38]. Moreover, as a new network architecture efficiently delivering content, information-centric networking (ICN), which stores content at routers and forwards packets on the basis of the content name, has gathered a lot of attention recently [5][14][25]. The storage capacities of cache servers and memories are finite, so the effect of CDN and ICN strongly depends on the location of cached content [42].

To improve the cache hit ratio and maximize the effect of CDN and ICN, various methods for estimating the future popularity of each content item have been investigated [1][11][22][27][29][39][40]. For example, Gursun et al. classified the change pattern of the view count of YouTube videos into two types, frequently accessed and rarely accessed, and proposed estimating the future demand of each YouTube video by estimating the change pattern of the principal components extracted by PCA for the former type and applying the change pattern of each cluster of videos classified by using the hierarchical clustering method for the latter type [22]. Moreover, Szabo et al. found a correlation between the initial popularity and long-term popularity in Digg and YouTube content, and they proposed estimating the long-term popularity of each content item by using its initial popularity [39].

However, unlike VoD services, in which major content holders provide content as commercial services, UGC is generated by a variety of users, so the change pattern of the popularity of UGC is complex and diverse [22], and the computational overhead in estimating the future popularity of each video is high. For example, the method of Gursun et al. used the autoregressive moving average (ARMA) model, which requires a large computational overhead, and the number of days with one or more views within one year needed to be recorded for each video [22]. The method of Szabo et al. needed to repeatedly calculate the regression coefficient in the linear model from the training data set [39]. Unlike VoD services, UGC is generated by a huge number of users, and the catalogue set, i.e., the set of content items, widely changes over time [10]. Therefore, although it is desirable to frequently repeat the estimation process of the future demand of each content item, estimating the demand of a huge number of content items within a short time interval is difficult for existing estimation methods, which require a large computational overhead. Although Xu et al. proposed a lightweight approach to forecast the future video popularity by utilizing the contextual information on social networks, the future popularity was just roughly forecasted over a limited number of popularity levels, e.g., low, medium, and high popularities [43].

In this paper, we construct a simple time-series model that captures the dynamics of the daily view count (DVC) of YouTube videos, which is one of the most popular types of UGC[1]. First, we analyze the DVC of YouTube videos over nine months and find that it obeys a lognormal distribution. This finding agrees with the result obtained by analyzing the DVC of YouTube videos, which was done by Borghol et al. [7]. The multiplicative process (MPP) is known as a simple time-series model that generates a lognormal distribution [31], so we model the dynamics of the DVC of YouTube videos by using the MPP. The MPP is a discrete-time stochastic process giving X_j, a random variable at time j, in $X_j = F_j X_{j-1}$. Here, F_j is an independent and identical arbitrary distribution, and we call this a *multiplicative value (MPV)* in this paper. The logarithm of X_j always obeys a normal distribution because of the central limit theorem. In this case, day is a discrete

N. Kamiyama is with the Faculty of Engineering, Fukuoka Universitry, Fukuoka, 814-0180 JAPAN. M. Murata is with the Department of Information Science, Osaka University, Osaka 565-0871, JAPAN. e-mail: kamiyama@fukuoka-u.ac.jp, murata@ist.osaka-u.ac.jp

[1]A shorter version of this manuscript was presented in [26].

time step, and the MPV is the magnification of the DVC of a YouTube video on the j-th day against its DVC on the previous day. We try to reproduce the DVC distribution of YouTube videos by using the superposed MPP (SMPP) aggregating multiple MPPs.

However, the magnification of the DVC of each YouTube video against its DVC on the previous day strongly depends on the magnitude of the DVC, so we cannot accurately reproduce the DVC distribution of YouTube videos by using the SMPP. Therefore, we propose capturing the dynamics of the DVC of each YouTube video by using the grouped MPP (gMPP), which gives the MPV distribution for each DVC group on the basis of its magnitude, and we also propose reproducing the DVC distribution of YouTube videos by using the superposed gMPP (SgMPP) aggregating multiple gMPPs. The contribution of this paper is summarized as the following two points.

- By analyzing the DVC data of YouTube videos over nine months, we clarify that the generated video count (GVC), defined as the video count newly uploaded on each day, the initial view count (IVC), defined as the view count on the uploaded date of each video, the DVC of each video, and the DVC of all videos on one day obey a lognormal distribution.
- We model the dynamics of the DVC of each YouTube video by using the gMPP and reproduce the DVC distribution of YouTube videos by using the SgMPP aggregating multiple gMPPs. The proposed SgMPP can accurately reproduce the DVC distribution of YouTube videos with a low computational overhead.

We can expect to use the proposed SgMPP as the input for computer simulations for designing various network components, which requires the popularity distribution of UGC, e.g., cache capacities. Table I summarizes the definition of symbols, and Tab. II summarizes abbreviations used in this paper.

After giving an overview of related works in Section II, we describe in detail the properties of the dataset of YouTube DVC used in this paper in Section III. In Section IV, we describe in detail applying the MPP to the time-series model of the YouTube DVC with the numerical results, and we describe the gMPP and SgMPP as well as the numerical results in Section V. We give an application example of the proposed SgMPP in Section VI and conclude this manuscript in Section VII.

TABLE I
DEFINITION OF SYMBOLS.

Symbol	Definition
$g(x)$	DVC group to which DVC x is classified
N_k	Number of MPPs at k-th time step
n_k	Number of newly added MPPs at k-th time step
Ω	Distribution of MPV
Ω_g	Distribution of MPV of DVC group g
$r_{i,k}$	MPV of MPP i at k-th time step, i.e., magnification of DVC on k-th day against that on previous day
$r_v(d)$	Magnification of DVC of video v on day d against that of previous day
Θ	Distribution of GVC approximated by lognormal distribution
U_v	Upload date of video v
Υ	Distribution of IVC approximated by lognormal distribution
$X_{i,k}$	State of MPP i at k-th time step, i.e., DVC of video i on k-th day
$x_v(n)$	DVC of video v on day n
$\tilde{x}_v(k)$	DVC of video v on k-th day from uploaded date
$y_v(n)$	Cumulative request count of video v on day n from uploaded date

TABLE II
ABBREVIATIONS.

ADVC (average DVC)
DVC (daily view count)
gMPP (grouped MPP)
GVC (generated video count)
IVC (initial view count)
LL (life length)
MPP (multiplicative process)
MPV (multiplicative value)
NDVC (normalized daily view count)
SgMPP (superposed gMPP)
SgMPP-G (SgMPP with G DVC groups)
SMPP (superposed MPP)

II. Related Works

To clarify the tendencies of the demand dynamics and the popularity distribution of YouTube videos, various results obtained by analyzing the access log of YouTube videos have been reported [6][9][10][13][17]. Arvidsson et al. revealed the periodicity of user requests [6], and Broxton et al. analyzed the change pattern of content popularity on social networks [9]. Moreover, Cha et al. compared the statistical tendency of the popularity distribution of YouTube videos with those of VoD content items [10], Cheng et al. investigated various properties, e.g., video length and bit rate, of YouTube videos by crawling YouTube videos [13], and Figueiredo et al. compared the change patterns of content popularity among selection mechanisms, i.e., external link and search, or video types, i.e., top-rank videos and illegal videos [17].

We can also find reports on the tendencies of the spatial pattern of demand on YouTube videos [8][16][45]. Duarte et al. compared the distribution of the view count of YouTube videos among three areas, i.e., the USA, South America, and others, by using randomly sampled YouTube videos [16]. Brodersen et al. analyzed the locality of demand and its change pattern by investigating the viewing history of YouTube videos over one year [8]. Zink et al. revealed geographical tendencies of the popularity of YouTube videos, e.g., low correlation between the global popularity and local popularity of YouTube videos [45]. Moreover, Dernbach et al. investigated the effect of considering the geographical locality of movie-content popularity in selection policy of cached content by using the MovieLens dataset giving the ratings of 4,000 movies [15]. Although we can obtain the DVC distribution of YouTube videos by analyzing the time- and spatial-change patterns of popularity, the obtained results are limited to a specific period and area, and we cannot generically use the results for various periods and areas. To estimate the DVC distribution of YouTube videos in a generic manner, it is desirable to model the change pattern of the DVC of YouTube videos by using a simple time-series model.

Therefore, to clarify the factors changing the popularity of each YouTube video, models capturing the change pattern of the view count of YouTube videos have been proposed [19][20][34][37][41]. Traverso et al. proposed modeling the transition of the request count of each YouTube video by using a short-noise model (SNM) obtained by aggregating multiple Poisson processes that represent each of the six groups in which YouTube videos were classified on the basis of the total request count and life length [41]. Moreover, Garetto et al. also proposed to capture the dynamics of content popularity by ON-OFF traffic model [19]. However, they focused on the time interval of requests within a short time scale. i.e., one day, so the change pattern of the popularity of YouTube videos over days or months was not considered.

Ghimire et al. modeled the popularity transition of each YouTube video by using a Markov chain [20]. Soysa et al. focused on the high correlation between the viewing frequency and the sharing ratio on Facebook, and they modeled the spread of interest on each YouTube video by using the fast threshold spread model (FTSM) [37]. Moreover, Ratkiewicz et al. revealed that the change ratio of the content popularity of Wikipedia and websites showed a power law distribution by analyzing the change pattern of external links, and they reproduced the discontinuous change of popularity due to external factors by using the ranking-shift model [34]. However, all three of these models focused on the change process of a single piece of UGC and did not consider the popularity distribution of the catalogue set of a large amount of UGC.

We can also find works reproducing the popularity distribution of UGC [3][7]. Adamic et al. revealed that the distribution of the number of users who visited each website in one day showed a power law property and theoretically showed that the power law distribution of the user count can be reproduced by using the MPP as the transition model of the number of users who visited each website in one day [3]. However, they focused on the number of users who visited each website instead of the DVC of YouTube videos. Borghol et al. proposed a method for reproducing the view count in one week of YouTube videos [7]. However, they reproduced the weekly view count by classifying YouTube videos into three phases, i.e., peak demand day, before the peak day, and after the peak day, and combining the distributions of view count for each of the three phases. Therefore, the time transition of the DVC of each video was not considered in [7]. Moreover, they assumed a fixed number of videos, and they did not consider change of video catalog, i.e., addition of new titles. In this paper, on the other hand, we propose a method of accurately estimating the DVC distribution at any time instance in future when YouTube videos are added dynamically.

III. YouTube Data Set

A. Procedure for Measuring Daily View Count

Using the YouTube Data API [21], which provides various statistical data on YouTube videos, we collected the DVC data of YouTube videos for 267 days, starting from April 9, 2013 to December 31, 2013. Hereafter, we indicate the date by using the elapsed day count from the initial day of measurement, i.e., April 9. For example, day 1 corresponds to April 9, and day 267 corresponds to December 31. Once every minute, we obtained the IDs of *recently uploaded videos*, i.e., videos newly uploaded in the latest one minute, by inquiring for this information from YouTube by using the API, and we generated a list of video IDs as well as the upload date for each of the 1,440 minute in a day. For example, in the list of 14:28, the IDs and the upload date of the videos uploaded within one minute from 14:28 were added day by day. These 1,440 lists of video IDs continued to increase day by day, and in total, 52,269 videos were added to one of the 1,440 lists.

Moreover, at every minute, we obtained the cumulative number of viewing requests from the upload day for each video included in the ID list of the corresponding time by inquiring of YouTube it using the API. By repeating this procedure every day, we obtained $y_v(n)$, the cumulative request count of each video v on each day n from the uploaded date, at the identical time. Let $x_v(n)$ denote the DVC of video v on day n and U_v denote the upload date of video v. We can calculate $x_v(n)$ from $y_v(n)$ as $x_v(n) = y_v(n) - y_v(n-1)$ for $U_v < n \leq 267$ and $x_v(n) = y_v(n)$ for $n = U_v$. The YouTube Data API provides the cumulative view count generated from users in all over the world for each video, so we cannot obtain DVC data on the basis of requests generated from a limited area or country. We leave the modeling of the popularity dynamics of YouTube videos in consideration of locality as future work.

B. Properties of DVC Data Set of YouTube Videos

In this section, we show the results of evaluating the properties of the DVC data of 52,269 YouTube videos mentioned in Section III-A. In addition to the DVC and IVC, as the properties which can be obtained from the YouTube data set, we also define the LL (life length) as the number of elapsed days of each video from the uploaded date until the day on which the last view was observed and the ADVC (average DVC) as the average DVC over LL days of each video. Table III summarizes the mean, median, standard deviation (STD), and maximum of the five properties of the YouTube data set, GVC, LL, IVC, DVC, and ADVC. We calculated the GVC for all 267 days, the LL, IVC, and ADVC for all of the 52,269 videos, and the DVC for all samples greater than or equal to unity of all 52,269 videos over all 267 days. We confirmed that the last view was observed around the last day, i.e., day 267 for almost all of the videos. Except for videos removed by YouTube due to copyright issues and those removed by the users who uploaded them, a large part of the videos seemed to remain in the video servers of YouTube. The LL of many YouTube videos was much larger than the length of the measurement period, 267 days, so it is difficult to analyze the LL of YouTube videos by using this DVC data set. To evaluate the LL of YouTube videos, we need a DVC data set with a much longer measurement period. We leave the analysis of the LL of YouTube videos as future work.

TABLE III
PROPERTIES OF YOUTUBE VIDEOS

	Mean	Median	STD	Maximum
GVC	198.7	186.0	66.3	508.0
LL	136.2	143.0	77.5	263.0
IVC	9.018×10^4	1.628×10^4	3.576×10^5	1.002×10^7
DVC	3.650×10^3	109.0	6.841×10^4	9.056×10^7
ADVC	6.287×10^3	557.2	5.909×10^4	5.859×10^6

C. Generated Video Count on Each Day

Figure 1(a) plots the GVC against each day. We observed no weekly periodicity in GVC, and we found that the difference of GVC among days of the week was small. However, we observed an increase and decrease trend on the scale of several tens of days after about day 100, and the GVCs in the initial about 80 days tended to be larger than those in the later days. Figure 1(b) shows the complementary cumulative distribution (CCD) of the GVC of the YouTube data set as well as the lognormal distribution, whose mean and STD were matched with those of the YouTube data set, i.e., mean of 198.7 and STD of 66.3. We confirmed that the lognormal distribution coincided with the distribution of the GVC.

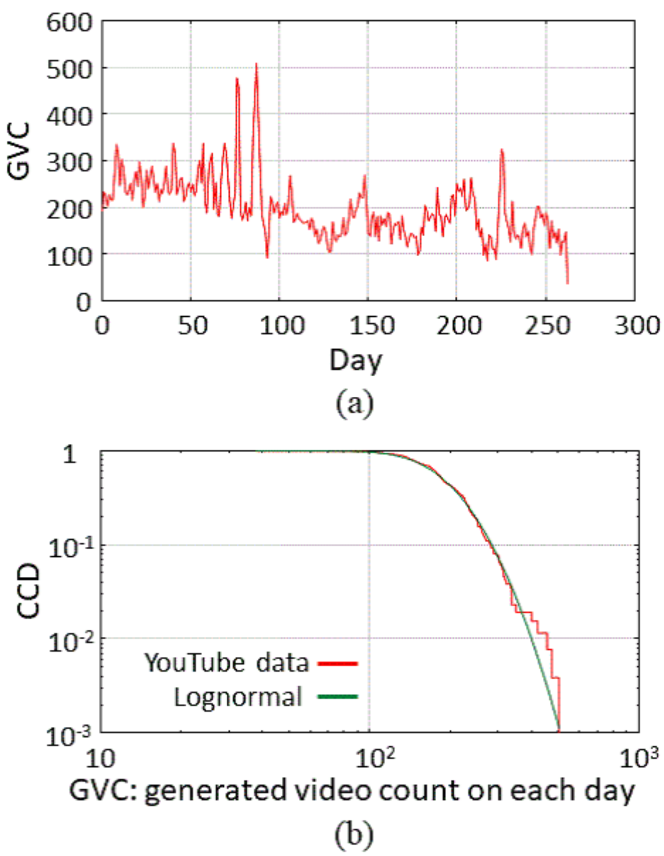


Fig. 1. Time series and CCD of video count uploaded each day

D. Initial View Count

Next, we investigate the tendency of the IVC, i.e., the number of views for each video v on the upload date. Figure 2(a) plots the average IVC of each day d. We observed that the average of IVC was largely different among days. We also show the CCD of IVC for all 52,269 videos and the lognormal distribution, whose mean and STD were matched with those of the YouTube data set, i.e., mean of 9.018×10^4 and STD of 3.576×10^5. We confirmed that the IVC of YouTube videos can be well approximated by the lognormal distribution.

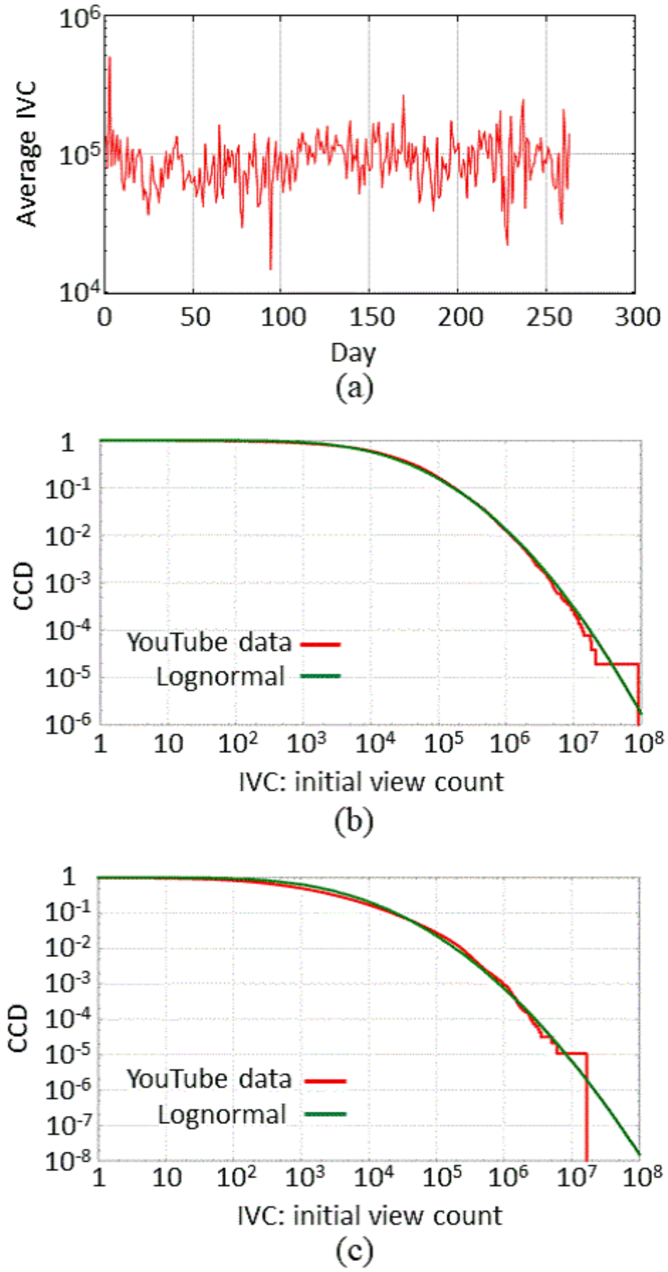


Fig. 2. (a) Average initial view count of videos uploaded on each day, (b) CCD of initial view count of YouTube videos

E. Daily View Count

Finally, we analyze various properties of the DVC of YouTube videos. Let $\tilde{x}_v(k)$ denote the DVC of video v on the k-th day from the uploaded date, and we define the normalized daily view count (NDVC) of video v on the k-th day as $\tilde{x}_v(k)$ divided by the maximum DVC of video v over the length of its life. Figure 3(a) plots the NDVC against k for each of 20 videos randomly sampled. We observed that the DVC of many YouTube videos dramatically decreased over several days just after their upload day and decreased moderately after this initial period, and this tendency of change pattern was also observed in the mean and median of NDVC of all videos. A similar tendency was also reported in existing works analyzing the dynamics of UGC popularity [6][9]. However, the change pattern of NDVC was largely different among videos.

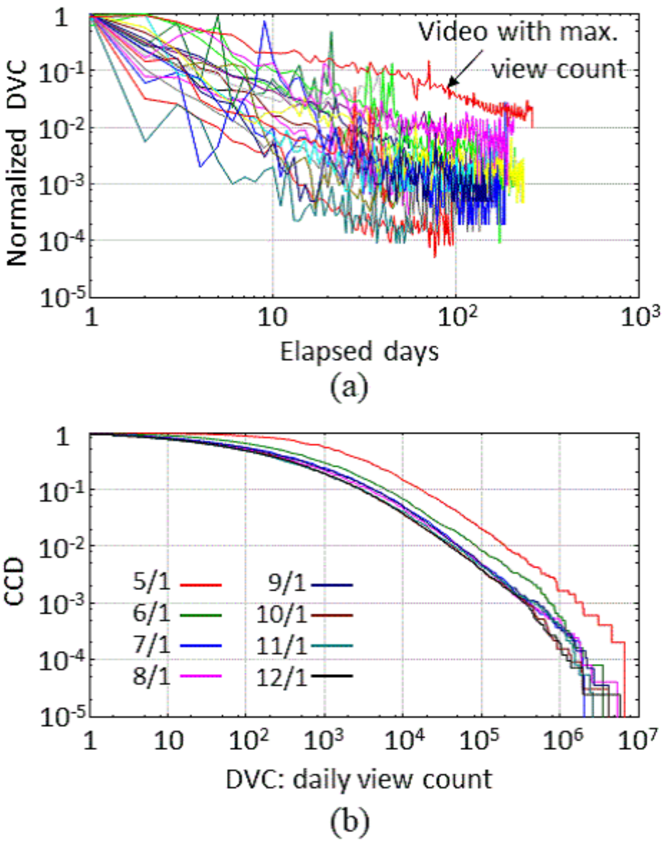


Fig. 3. (a) Dynamics of NDVC of 20 sampled videos, (b) CCD of DVC of eight sampled days

Next, we analyze the tendency of $x_v(d)$, the DVC of video v on day d. Figure 3(b) shows the CCD of the DVC for eight sampled days, i.e., the first day of each month. As mentioned in Section III-A, the data set of the YouTube DVC included only videos uploaded after April 9, 2013, so only videos with a small number of elapsed days after their upload date were included in the data set when d was small. As observed in Figure 3(a), the number of views tended to be large just after the upload date, so the sampled DVC in a small-d region concentrated on those of a large value. Therefore, on days close to the initial date of measurement, e.g., May 1 and June 1, the sampled DVC tended to concentrate on large values, so the CCD on these days shifted in the upper-right direction. However, for the other six sampled days, the CCDs of the DVC were almost identical. We confirmed that the distribution of DVC on each day became stable on days after about 100 days from the date measurement started because various videos that had different elapsed day counts after the upload date existed. Although the DVC of each video dramatically changed just after their upload date as seen in Figure 3(a), the DVC distribution of each day was stable as a result of multiplexing multiple videos with various elapsed day counts.

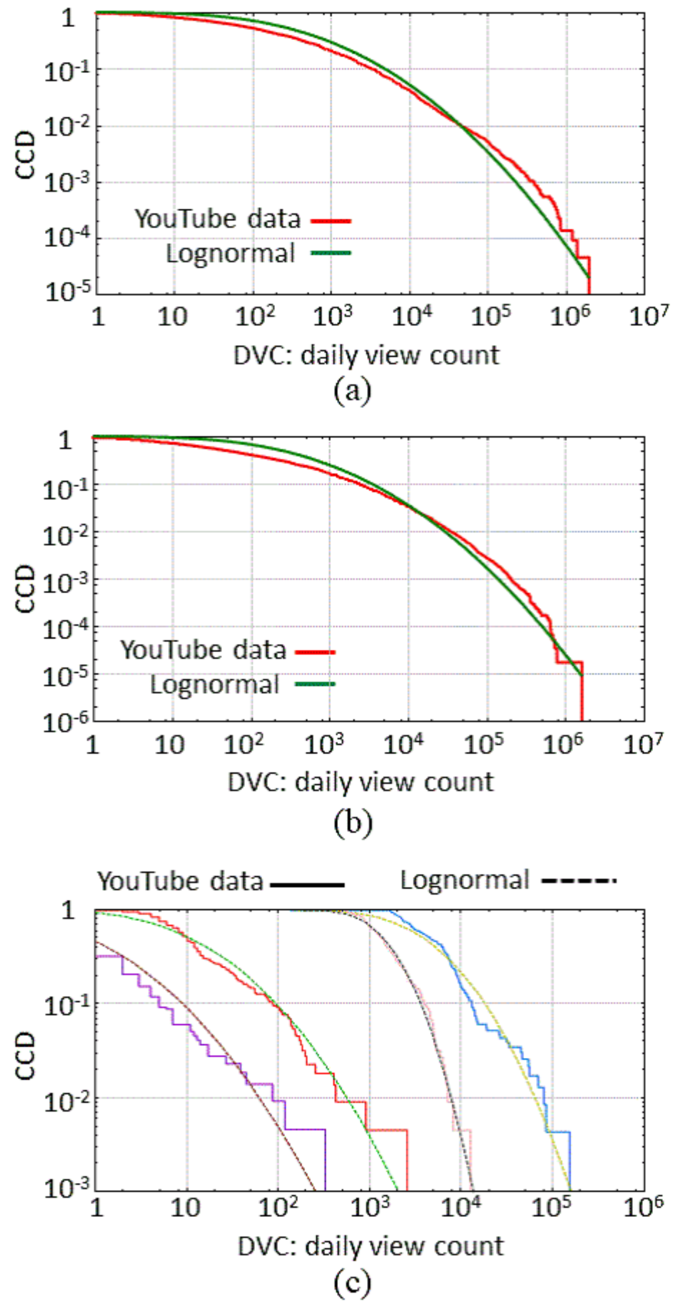


Fig. 4. (a) CCD of DVC of all videos after 100th day, (b) CCD of DVC of four sampled videos over all days

Next, we plot the CCD of the DVC of all videos for all days after day 100 in Figure 4(a). We also show the lognormal distributions whose average and STD were matched to those of the YouTube data set. The two distributions almost coincided, so we confirmed that the DVC distribution of many videos over many days can be also approximated by the lognormal distribution. Borghol et al. confirmed that the distribution of the view count of YouTube videos obeyed a lognormal distribution [7], and our finding agreed with this report. Figure 4(b) plots the CCD of the DVC of four videos randomly sampled as well as the lognormal distributions whose average and STD were matched with the average and STD of each of these four videos. We confirmed that the DVC of each video over multiple days also obeyed a lognormal distribution.

IV. Modeling Popularity Dynamics of YouTube Videos with Multiplicative Process

As observed in Figure 4(b), the DVC of each YouTube video obeyed a lognormal distribution. The multiplicative process (MPP) is widely known as a simple random process that generates this distribution [31], so we first consider applying the MPP to the time transition model of the DVC of each YouTube video in this section.

A. Multiplicative Process

When random variable X_j takes X_0 at the initial state and X_j at discrete time j, the MPP is defined as

$$X_j = F_j X_{j-1}, \tag{1}$$

where the random variable F_j, which we call *multiplicative value (MPV)*, independently obeys an identical arbitrary distribution. In other words, the MPV F_j, i.e., the magnification of X_j against the previous value X_{j-1}, is given by the identical distribution independently of j. By recursively applying this formula, ln X_j is given by

$$\ln X_j = \ln X_0 + \sum_{k=1}^{j} \ln F_k. \tag{2}$$

Therefore, when F_j independently obeys an identical distribution, ln X_j always obeys a lognormal distribution according to the central limit theorem, so X_j generated by the MPP obeys a lognormal distribution.

Next, let us consider aggregating multiple MPPs. The distribution generated by multiple MPPs depends on the distribution of the life length of each MPP [31]. For example, it was reported that aggregating multiple MPPs generates a distribution with the body of a lognormal distribution and the tail of a Pareto distribution when the life length of each MPP obeys a geometric distribution [35]. In this paper, we call the random process generated by aggregating multiple MPPs *superposed MPP (SMPP)*.

B. Applying MPP to Model DVC Dynamics of YouTube Video

By giving the distributions of initial value X_0 and MPV F_j, we can uniquely determine the MPP. Moreover, by giving the distributions of the MPP count newly added at each time step and the life length in addition to X_0 and F_j, we can uniquely determine the SMPP. We can model the DVC dynamics of YouTube videos by using the SMPP with regarding a day as the discrete time step of the MPP, the DVC as X_j, the IVC as X_0, the magnification of the DVC against that on the previous day of the same video as the MPV F_j, and the GVC as the number of MPPs newly added at each time step. However, as mentioned in Section III-B, we cannot define the life length of each video in the DVC data evaluated, so all of the MPPs that have been generated remain until the repetition process is completed. Therefore, N_j, the number of MPPs, monotonically increases as the time step proceeds, and we consider only MPPs with X_j greater than or equal to unity when evaluating the distribution of X_j generated[2].

Therefore, to apply the SMPP to reproduce the DVC distribution of YouTube videos, we need to give three distributions, (i) the GVC, the number of videos uploaded each day, (ii) the IVC, the number of views of each video on the upload date, and (iii) the MPV, the magnification of the DVC of each video against that of the precious day of the same video. As mentioned in Sections III-C and III-D, the distributions of both the GVC and the IVC are given by a lognormal distribution. Next, we investigate the distribution of the MPV by analyzing the YouTube data set as well.

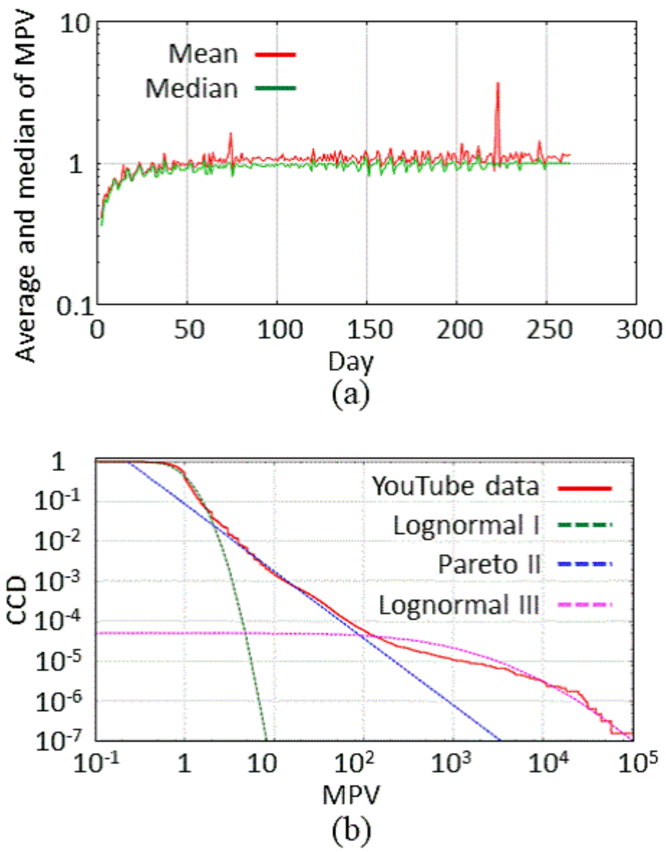


Fig. 5. (a) Mean and median of MPV on each day, (b) CCD of MPV

Figure 5(a) plots the mean and median of the MPV, $x_v(d)/x_v(d-1)$, among videos whose $x_v(d)$ and $x_v(d-1)$ were greater than zero against day d. The MPV r took the value in the range of $0 < r < \infty$, and $r = 1.0$ meant that the DVC of video v was exactly the same as that on the previous day. Moreover, $r < 1.0$ and $r > 1,0$ meant that the DVC of video v decreased or increased compared with that on the previous day, respectively. As mentioned in Section III-E, at the initial phase of the measurement, only videos whose elapsed day count from the upload date were small were included in the data set, and the DVCs of these videos were likely to decrease sharply, so the mean and median of MPV was smaller than unity. As the day progressed, the ratio of

[2] It is anticipated that there are many videos that are no longer viewed on YouTube, and we can say that the MPPs with X_j less than unity correspond to YouTube videos no longer viewed.

videos with more elapsed days from the upload date increased, and the DVCs of these videos were stable. Therefore, after about day 60, the mean and median of MPV changed stably, and the median of MPV took the value close to unity.

Figure 5(b) shows the CCD of the MPV over all videos and days, and we confirmed that the curve was convex upward and decreased rapidly in the small MPV region. However, it linearly decreased in the wide middle range of the MPV and gradually decreased in the large MPV region, so the curve had a longer tail than the power law curve. We can expect that the long-tail distribution of the MPV originated from the phenomenon that the popularity of some YouTube videos may suddenly and dramatically increase due to the diffusive effect on social networking services (SNSs), e.g., Facebook [34]. In sum, $F(r)$, the distribution of MPV r, consists of three parts: $0.01 \leq F(r) \leq 1.0$ (zone I), $5.0 \times 10^{-5} \leq F(r) < 0.01$ (zone II), and $0.0 \leq F(r) < 5.0 \times 10^{-5}$ (zone III). We compared the CCD of the MPV with the lognormal distribution in zones I and III and with the Pareto distribution in zone II.

In Figure 5(b), we also plot a lognormal distribution (Lognormal I) with an average of 1.021 and STD of 0.445, which matched those of the CCD of MPV in zone I, a Pareto distribution (Pareto II) with an average of 4.446 and STD of 7.344, which matched those of the CCD of MPV in zone II, and a lognormal distribution (Lognormal III) with an average of 3.157×10^3 and STD of 1.334×10^4, which matched those of the CCD of MPV in zone III. We confirmed that the CCD of the MPV of YouTube videos can be approximated by using these three distributions well. It is desirable to use a simple distribution for the MPV when reproducing the DVC of YouTube videos by using the SMPP to suppress the calculation time required, so we consider approximating the MPV distribution by using only Lognormal I, which dominates a large part of the MPV distribution.

In Algorithm I, we summarize the procedure of the SMPP executed at each time step k with the purpose of reproducing the DVC distribution of YouTube videos on each day. Let N_k denote the number of MPPs at the k-th time step. We also define $X_{i,k}$ as the state of MPP i, $1 \leq i \leq N_k$ at the k-th time step, and we assume that no MPP exists at the initial state, i.e., $N_0 = 0$. Moreover, Θ, Υ, and Ω are the lognormal distributions approximating the distributions of the GVC, the IVC, and the zone I of the MPV of the YouTube data set.

As mentioned in Section III-A, the DVC data set of YouTube videos started in the empty state, i.e., having no videos, on the initial date of measurement, April 9, 2013, and we continued to add the videos newly uploaded on each day to the DVC data set. We note that the above procedure of the SMPP agrees with the construction process of the DVC data set of YouTube videos.

C. Numerical Results

We regard each MPP as the random process of the DVC of each YouTube video, and we compare the distribution generated by the SMPP with the DVC distribution of YouTube videos on each day. In addition to the case denoted as *lognormal MPV*, given the MPV distribution generated by Ω, Lognormal I, mentioned in Section IV-B, we evaluated the case denoted as *actual MPV*, given the MPV distribution generated by the actual MPV distribution of the YouTube data set. To bound the $X_{i,k}$ less than the maximum value 9.056×10^7 observed in the DVC data set of YouTube video, we reselected $r_{i,k}$ when $X_{i,k}$ exceeded this maximum value for each MPP i at each time step k of the SMPP. We repeated the procedure of the SMPP 267 times, and we regarded step 1 of the SMPP as the initial date of measuring the YouTube DVC, i.e., April 9, and the final step, i.e., step 267, as the final date of measuring the YouTube DVC, i.e., December 31, 2013.

Algorithm 1 Procedure executed at k-th time step of SMPP

1: For each MPP i of $1 \leq i \leq N_k$, update $X_{i,k}$ by $X_{i,k} = r_{i,k} X_{i,k-1}$,where MPV $r_{i,k}$ is randomly selected according to Ω
2: Randomly select n_k, the number of newly added MPPs, according to Θ and update $N_{k+1} = N_k + n_k$
3: For each MPP i of newly added n_k MPPs, randomly set the initial value of $X_{i,k}$ according to Υ

Figure 6 plots the CCD of the YouTube DVC for the four selected dates, May 1, June 1, August 1, and October 1, as well as the CCD of X_j generated by the SMPP at the corresponding time steps, i.e., 19-th, 50-th, 111-th, and 172-th steps. We repeated the SMPP ten times with different random seeds and showed the CCD considering all the sampled obtained in the ten trials at each corresponding time steps. We observed a large gap between the distribution generated by the SMPP and that of the actual DVC of the YouTube data set for all the four sampled dates. For all four of the sampled dates, the distribution generated by the SMPP was shifted toward the upper and right direction compared with the actual DVC distribution, so we confirmed that the SMPP produced larger values compared with the DVC of YouTube videos. As seen in Figure 5(b), the actual distribution of the MPV had a long tail, so the possibility of applying a large value to MPV was higher when using the actual MPV distribution than that when using the Lognormal I as the MPV distribution. Therefore, the deviation of the distribution generated by the SMPP was larger when using the actual MPV distribution than that when using the Lognormal I as the MPV distribution.

In summary, we cannot reproduce the DVC distribution of YouTube videos on any day by using SMPP. Therefore, in the next section, we try to improve the accuracy of reproducing the DVC distribution of YouTube videos by extending the MPP.

V. Extension of Multiplicative Process for Modeling YouTube Popularity Dynamics

A. Applying MPV Distribution for each DVC Group

In Section IV-B, we simply applied the magnification of DVC against that of the previous day of the same video to the MPV without distinguishing the actual value of the DVC and video types. However, as mentioned in Section III-E, the DVC of many YouTube videos tends to be large and rapidly decrease on days close to the upload date, whereas it tends to be small and gradually decrease after days elapse. Therefore,

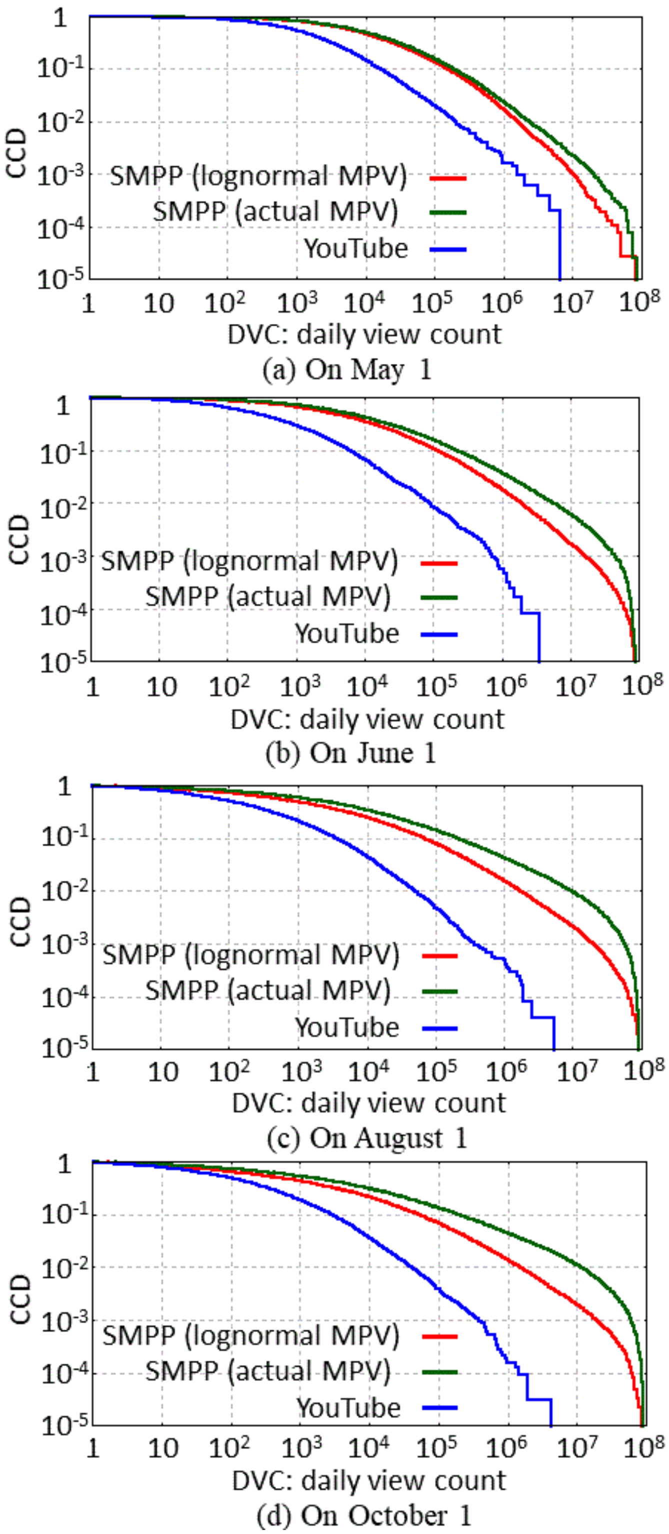


Fig. 6. Comparison of CCD of YouTube DVC for four sampled days and CCD of values generated by SMPP at corresponding time steps

we can expect that the change ratio of the DVC on the next day strongly depends on the magnitude of the DVC. We define $r_v(d)$ as the magnification of the DVC of video v on day d against that of the previous day, i.e., $r_v(d) \equiv x_v(d)/x_v(d-1)$, and Figure 7(a) shows $r_v(d)$ against $x_v(d-1)$. Although $r_v(d)$ was distributed in a wide range even when $x_v(d-1)$ was a similar value, $r_v(d)$ tended to take a large value when $x_v(d-1)$ was small. Therefore, when classifying the DVC, $x_v(d-1)$, into multiple groups on the basis of the magnitude of $x_v(d-1)$, we can expect that the distribution of MPV, $r_v(d)$, will be different among the DVC groups.

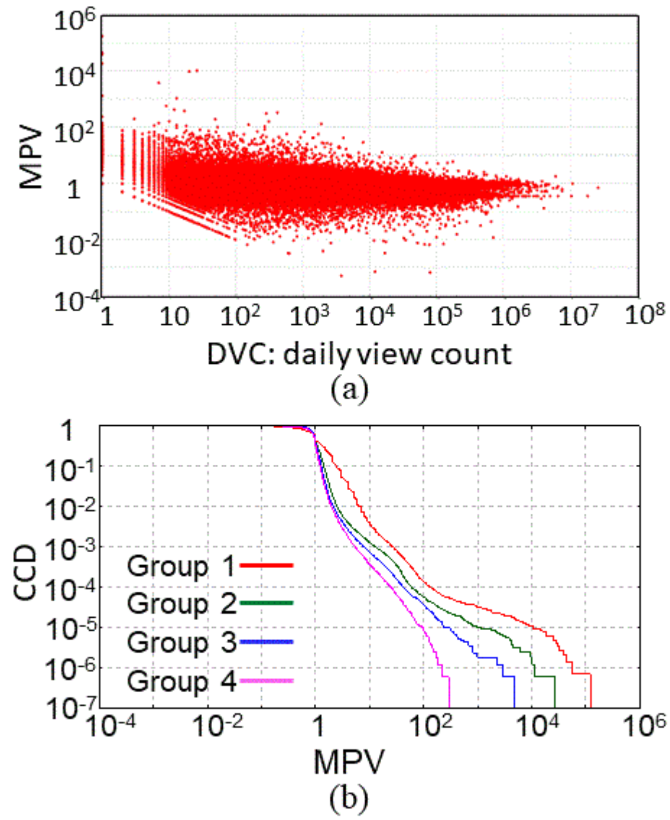


Fig. 7. (a) Scattergram between DVC and MPV of next day of each video on each day, (b) CCD of DVC of four DVC groups

TABLE IV
BOUNDARIES, MEAN, MEDIAN, AND STD OF EACH DVC GROUP

Group	Lower	Upper	Mean	Median	STD
G1	1	15	1.962	1.000	200.3
G2	16	108	1.140	1.000	37.42
G3	109	703	1.033	0.993	7.780
G4	704	∞	0.953	0.955	0.955

To confirm this, we classified the MPV samples into four groups by setting three boundaries on the DVC value so that almost the same number of MPV samples was classified into each DVC group[3]. Table IV summarizes the mean, median, and STD of the MPV samples classified as well as the lower and upper boundaries of each of the four DVC groups. We assigned the DVC group ID in ascending order of the magnitude of DVC. As seen in Figure 4(a), many DVC samples

[3]When dividing the MPV samples into various number of groups, e.g., 2, 8, 16, and 32, we also confirmed the same tendencies mentioned in the later part of this section.

concentrated on the small-value range, so the interval between the lower and upper boundaries was smaller in the DVC group of smaller magnitude. Moreover, all of the means, medians, and STDs of the MPV were smaller in the DVC group with a larger magnitude. As seen in Figure 3(a), the DVC of many YouTube videos was large and rapidly decreased on days close to the upload date, whereas it gradually decreased on average with fluctuation within a small range. Therefore, in videos with a large DVC, the DVC on the next day is more likely to decrease greatly, and the MPV tends to be small.

Figure 7(b) plots the CCD of the MPV samples in each of the four DVC groups. Although the CCD of MPV still formed a distribution with a longer tail than the power-law distribution like the curves shown in Figure 6(b) even when MPV samples were classified into the four DVC groups, the CCD of MPV was largely different among the DVC groups, and the MPV tended to take a larger value in a wider range in the DVC group with a smaller magnitude.

We can expect that many MPV samples classified into DVC group 4 were videos with an extremely large DVC just after their upload date, so the mean and median of the MPV of this group were less than 1.0 because the DVC on the next day largely decreased with high probability. In comparison, many MPV samples classified into DVC groups 1 or 2 were expected to be videos that had been existed for many days after being uploaded, so the median of MPV was unity because the variation of DVC was small. However, just a few times, the popularity of specific YouTube videos suddenly and extremely increased because of the word-of-mouth effect on SNSs [34], for example, and an extremely large MPV was observed[4]. Therefore, the mean of the MPV of DVC group 1 was about 2.0, which was much larger than the median of MPV of this group.

The MPV distribution was different among the DVC groups, so we considered applying the MPV according to the MPV distribution of the DVC group in which the current state X_j is included in the MPP. In this paper, we call this extended MPP *grouped MPP (gMPP)*. Figure 8 plots the CCD of the MPV obtained from the YouTube data set in each of the four DVC groups. We also show a lognormal distribution (Lognormal I) with the means and STDs matching those in the smallest 99% of the MPV samples, a Pareto distribution (Pareto II) with the means and STDs matching those in the largest 1% of the MPV samples, and a lognormal distribution (Lognormal III) with the means and STDs matching those in the largest 0.005% of the MPV samples. We confirmed that the MPV distribution of each of the four DVC groups can be accurately approximated by using the combinations of the lognormal distributions and Pareto distribution in two or three zones.

It is also desirable to approximate the MPV distribution of each DVC group by using a single distribution to minimize the computational overhead. Because almost all of the MPV samples of the YouTube data set existed in the zone that can be approximated by Lognormal I, in the gMPP, we apply the MPV distribution of each of the G DVC groups approximated by Lognormal I. We propose reproducing the DVC distribution

[4] The maximum MPV sample was 1.707×10^7.

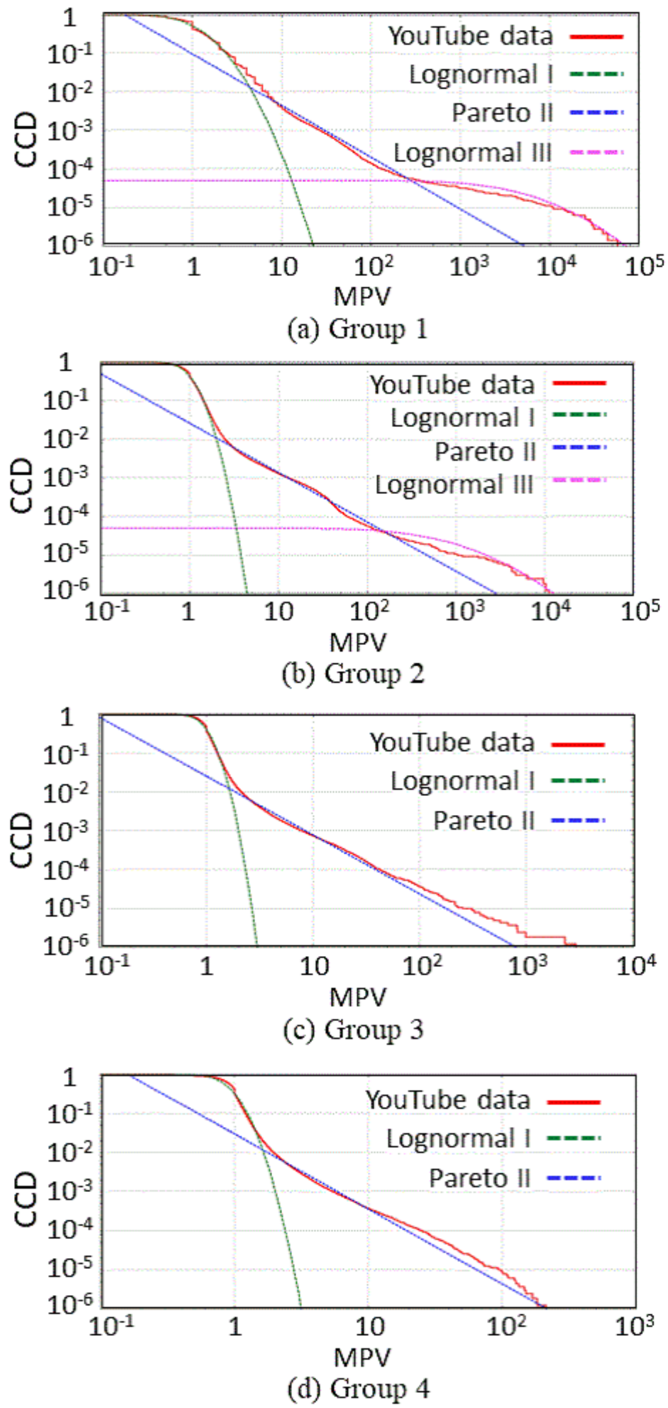


Fig. 8. Fitting of CCD curves of DVC in each of four DVC groups

of YouTube videos for each day by using the *superposed gMPP (SgMPP)* aggregating multiple gMPPs. In Algorithm II, we summarize the procedure executed at the k-th time step of the SgMPP. When we define b_g as the lower boundary of the DVC of DVC group g and let $g(x)$ denote the DVC group to which DVC x is classified, $b_{g(x)} \le x < b_{g(x)+1}$ is satisfied for each g of $1 \le g \le G$. We denote the lognormal distribution (Lognormal I) with the mean and STD of the MPV samples of the smallest 99% as Ω_g.

Algorithm 2 Procedure executed at k-th time step of SgMPP

1: For each gMPP i of $1 \leq i \leq N_k$, update $X_{i,k}$ by $X_{i,k} = r_{i,k} X_{i,k-1}$, where MPV $r_{i,k}$ is randomly selected according to $\Omega_{g(X_{i,k-1})}$
2: Randomly select n_k, the number of newly added gMPPs, according to Θ and update $N_{k+1} = N_k + n_k$
3: For each gMPP i of newly added n_k gMPPs, randomly set the initial value of $X_{i,k}$ according to Υ

B. Numerical Results

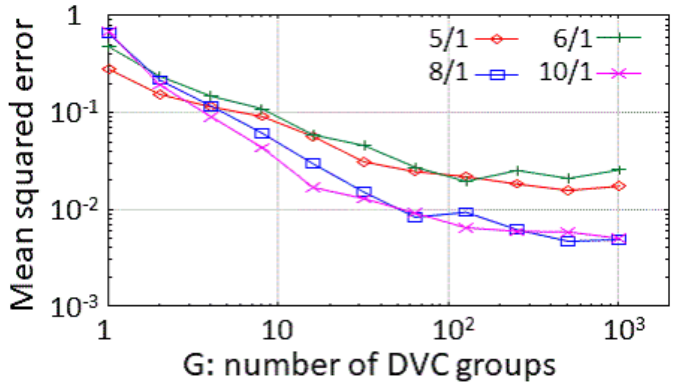
We denote the SgMPP with G DVC groups as SgMPP-G, and we evaluate the accuracy of SgMPP-G in reproducing the DVC distribution of YouTube videos with the mean squared error (MSE) [28]. Let x_s denote the boundary values of DVC when dividing the range between its minimum (1.0) and the maximum (9.056×10^7) into 100 intervals with identical length on the logarithm scale, i.e., $x_s = \exp(\log(x_{max}/100) \cdot s)$, $s = 1, 2, \cdots, 100$. Using $\hat{z}(x_s)$ and $z(x_s)$, the value of CCD generated by the SgMPP-G, and the DVC distribution of YouTube videos at $x = x_s$, we define the MSE as

$$\mathrm{MSE} = \frac{\sum_{s=1}^{100} \left\{ \hat{z}(x_s) - z(x_s) \right\}^2}{100}. \quad (3)$$

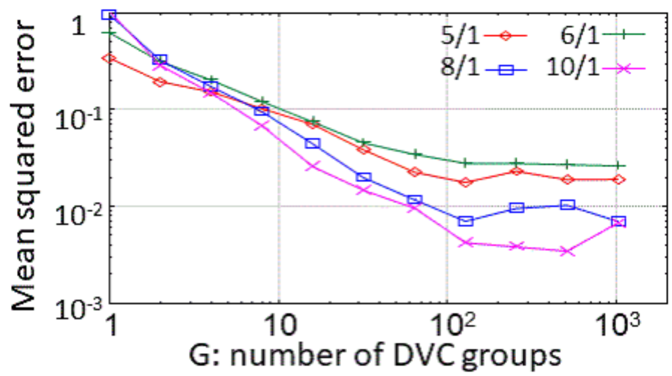
Figure 9(a) plots the MSE of the SgMPP-G at the time steps corresponding to the four sampled dates, May 1, June 1, August 1, and October 1, against G when using Ω_g as the MPV distribution of each DVC group g. We set the boundaries of G DVC groups so that an identical number of MPV samples were classified into each DVC group, and we show the average results of ten repetitions with different random seeds. We note that SgMPP-1 corresponds to the SMPP mentioned in Section IV-B. Moreover, we show the same results when using the actual MPV distribution of each DVC group g in the YouTube data set in Figure 9(b).

In the wide range of G, the SgMPP-G using Lognormal I as the MPV distribution of each DVC group achieved a similar accuracy in reproducing the DVC distribution of YouTube videos with the case using the actual MPV distribution. By using Lognormal I as the MPV distribution, we can dramatically reduce the computational time required in executing the SgMPP-G process, so using the Lognormal I distribution is desirable. In the small-G region, the MSE decreased as G increased for all the four sampled days, and the accuracy of reproducing the DVC distribution of YouTube videos improved, whereas the MSE was almost constant when increasing G in the region of G greater than about 50. We need to calculate the Lognormal I distribution of MPV for each DVD group, so a smaller G is desirable to reduce the computational time in constructing the SgMPP-G model. Therefore, it is desirable to set G in the range between about 40 and 70.

Finally, we evaluate the accuracy of the SgMPP in reproducing the DVC distribution of YouTube videos when setting $G = 64$. Figure 10 plots the CCD of the DVC of YouTube data set on the four sampled dates, May 1, June 1, August 1, and October 1, as well as the CCD of X_j generated by the SgMPP-64 at the corresponding time steps. We also repeated the SgMPP-64 for ten times with different random seeds. For all the four sampled days, especially on August 1 and October 1 after reaching the steady state, we confirmed that we can accurately reproduce the DVC distribution of YouTube videos by using the SgMPP-64.



(a) Using MPV of Lognormal I in each DVC group



(b) Using actual MPV distribution in each DVC group

Fig. 9. Mean squared error between distribution generated by SgMPP-G and that of YouTube DVC

VI. Application of Proposed Method

Here, we briefly describe an application example of the proposed method of reproducing the DVC distribution of UGCs. As a cache replacement policy for selecting content items to be removed when the storage capacity of caches is fully utilized, least recently used (LRU), which removes content with the longest elapsed time from the last access, and least frequently used (LFU), which removes content with the smallest access ratio, are most widely used [33]. Although LRU and LFU are simple policies and do not require the demand estimation of each content item, it is known that they can achieve a cache hit ratio almost equal to that obtained by optimally placing content on the basis of the access demand because popular content items remain in caches as a result of replacing content with LRU and LFU [36].

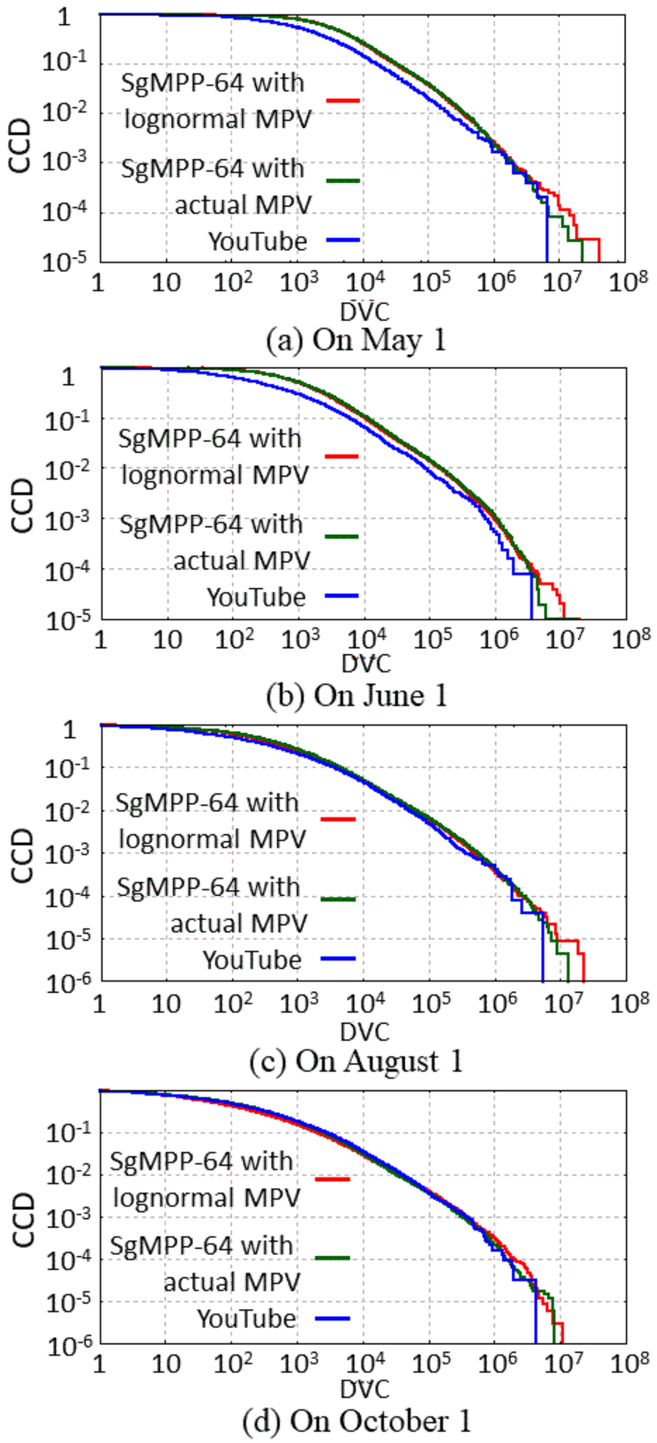


Fig. 10. Comparison of CCD of YouTube DVC on four sampled days and CCD of values generated by SgMPP-64 at corresponding steps

To optimally design the capacity of caches to satisfy the target cache hit ratio in LRU and LFU, we still need to estimate the cache hit ratio achieved for a given cache size. For example, by using the Che′s equation, we can easily derive the cache hit ratio only if the demand distribution of content items can be estimated [12]. However, the demand distribution of UGC depends on the catalogue set, so it is desirable to easily estimate the demand distribution of UGC by computer simulation when various conditions, e.g., the total user count and video count generated on each day, change. To achieve this goal in a practical time frame, we need to construct a simple time-series model that captures the transition of UGC popularity.

Moreover, dynamically constructing CDNs using virtual machines on cloud datacenters has gathered wide attention recently [23][30]. In virtual CDNs, content location can be dynamically configured based on the estimation of demand distribution [24]. The proposed SgMPP can accurately reproduce the DVC distribution of UGCs at any time point in future from the given lognormal distributions of the GVC, the IVC, and the MPV of each DVC group, and we can generate these input distributions from a small dataset obtained by monitoring the demand of UGCs in a sampled area within a limited time period. By applying the proposed SgMPP with the sampled input distributions, we can estimate the DVC distribution of UGCs in various areas at various time instances, so we can effectively design the capacity of cache servers for UGCs in existing CDNs and the content location on virtual CDNs.

VII. Conclusion

In this paper, we proposed a simple time-series model, SgMPP (superposed grouped MPP), based on the multiplicative process (MPP) that represents the dynamics of the daily view count (DVC) of YouTube videos to accurately reproduce the DVC distribution of YouTube videos, and we numerically showed that the proposed SgMPP can accurately reproduce the DVC distribution of YouTube videos. The calculation time required to reproduce the DVC distribution with SgMPP was small, so we can expect to apply the SgMPP to various designs and controls that require the demand distribution of UGC, e.g., the capacity design of cache servers for large-scale UGC services. In the future, we will theoretically reveal the principles that aggregating multiple gMPPs with infinite life length produces the lognormal distribution. Moreover, we will analyze the case when the life length of YouTube videos can be modeled by finite distribution through measuring the DVC data of YouTube over a year, and we will also investigate a time-series model for reproducing the DVC distribution of YouTube videos in consideration of locality.

References

[1] E. Abdelkrim, M. Salahuddin, H. Elbiaze, and R. Glitho, A Hybrid Regression Model for Video Popularity-based Cache Replacement in Content Delivery Networks, IEEE GLOBECOM 2016.

[2] S. Acharya, B. Smith, and P. Parnes, Characterizing User Access To Videos On The World Wide Web, MMCN 2000.

[3] L. Adamic and B. huberman, The Nature of Markets in the World Wide Web, Quarterly Journal of Economic Commerce 1, 2000.

[4] B. Ager, W. Muhlbauer, G. Smaragdakis, and S. Uhlig, Web Content Cartography, ACM IMC 2011.
[5] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, A Survey of Information-Centric Networking, IEEE Commun. Mag., vol.50, no.7, pp.26-36, July 2012.
[6] A. Arvidsson, M. Du, A. Aurelius, and M. Kihl., Analysis of User Demand Patterns and Locality for YouTube Traffic, ITC 25.
[7] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, Characterizing and Modeling Popularity of User-generated Videos, Performance Evaluation, 2011.
[8] A. Brodersen, S. Scellato, and M. Wattenhofer, YouTube Around the World: Geographic Popularity of Videos, WWW 2012.
[9] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer, Catching a viral video, Springer J. Intell. Inf. Sys., 2011.
[10] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems, IEEE/ACM trans. Networking, Vol. 17, No. 5, pp.1357-1370, Oct. 2009.
[11] G. Chatzopoulou, C. Sheng, and M. Faloutsos, A first step towards understanding popularity in YouTube, IEEE Global Internet 2010.
[12] H. Che, Y. Tung, and Z. Wang, Hierarchical Web Caching Systems: Modeling, Design and Experimental Results, IEEE JSAC, Vol. 20, No. 7, pp.1305-1314, Sep. 2002.
[13] X. Cheng, C. Dale, and J. Liu, Statistics and Social Network of YouTube Videos, IEEE IWQoS 2008.
[14] J. Choi, J. Han, E. Cho, T. Kwon, and Y. Choi, A Survey on Content-Oriented Networking for Efficient Content Delivery, IEEE Commun. Mag., vol.49, no.3, pp.121-127, Mar. 2011.
[15] S. Dernbach, N. Taft, J. Kurose, U. Weinsberg, C. Diot, and A. Ashkan, Cache Content-Selection Policies for Streaming Video Services, IEEE INFOCOM 2016.
[16] F. Duarte, F. Benevenuto, V. Almeida, and J. Almeida, Geographical Characterization of YouTube: a Latin American View, Latin American Web Congress 2007.
[17] F. Figueiredo, D. Benevenuto, J. Almeida, The Tube over Time: Characterizing Popularity Growth of YouTube Videos, ACM WSDM 2011.
[18] C. Fricker, P. Robert, and J. Roberts, A Versatile and Accurate Approximation for LRU Cache Performance, ITC 24.
[19] M. Garetto, E. Leonardi, and S. Traverso, Efficient analysis of caching strategies under dynamic content popularity, IEEE INFOCOM 2015.
[20] J. Ghimire, M. Mani, and N. Crespi, Modeling Content Hotness Dynamics in Networks, SPECTS 2010.
[21] Google Developers YouTube Data API, https://developers.google.com/youtube/v3/
[22] G. Gursun, M. Crovella, and I. Matta, Describing and Forecasting Video Access Patterns, INFOCOM 2011 Mini-conference.
[23] N. Herbaut, D. Negru, Y. Chen, P. Frangoudis, and A. Ksentini, Content Delivery Networks as a Virtual Network Function: a Win-Win ISP-CDN Collaboration, IEEE GLOBECOM 2016.
[24] M. Hu, J. Luo, Y. Wang, and B. Veeravalli, Practical Resource Provisioning and Caching with Dynamic Resilience for Cloud-Based Content Distribution Networks, IEEE Trans. Parallel and Distributed Systems, 25 (8), Aug. 2014.
[25] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard, Networking Named Content, ACM CoNEXT 2009.
[26] N. Kamiyama, M. Murata, Reproducing Popularity Dynamics of YouTube Videos, to be presented at CNSM 2018 (Mini-Conference).
[27] J. Lee, S. Moon, and K. Salamatian, An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors, IEEE/WIC/ACM WI-IAT 2010.
[28] E. L. Lehmann and G. Casella, Theory of Point Estimation, New York, Springer, 1998.
[29] K. Lerman and T. Hogg, Using a Model of Social Dynamics to Predict Popularity of News, WWW 2010.
[30] P. Marchetta, J. Llorca, A. Tulino, and A. Pescape, MC3: A Cloud Caching Strategy for Next Generation Virtual Content Distribution Networks, IFIP Networking 2016.
[31] M. Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions, Internet Mathematics, Vol. 1, No. 2, 2003.
[32] J. Ott, M. Sanchez, J. Rula, F. Bustamante, Content Delivery and the Natural Evolution of DNS, ACM IMC 2012.
[33] S. Podlipnig and L. Boszormenyi, A Survey of Web Cache Replacement Strategies, ACM Computing Surveys, Vol. 35, No. 4, pp. 374398, Dec. 2003.
[34] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, Characterizing and modeling the dynamics of online popularity, Physical Review Letters, Vol. 105, No. 15, Oct. 2010.
[35] W. J. Reed, The Pareto Law of Incomes - An Explanation and an Extension, Physica A 319, pp. 469-485, 2003.
[36] A. Sharma, A. Venkataramani, R. Sitaraman, Distributing Content Simplifies ISP Traffic Engineering, SIGMETRICS 2013.
[37] D. Soysa, D. Chen, O. Au, and A. Bermak, Predicting YouTube Content Popularity via Facebook Data: A Network Spread Model for Optimizing Multimedia Delivery, IEEE CIDM 2013.
[38] A. Su, D. Choffnes, A. Kuzmanovic, and F. Bustamante, Drafting Behind Akamai: Inferring Network Conditions Based on CDN Redirections, ACM Trans. Networking, 17(6), pp.1752-1765, 2009.
[39] G. Szabo and B. Huberman, Predicting the Popularity of Online Content, ACM Communications, 2010.
[40] J. Tirado, D. Higuero, F. Isaila, and J. Carretero, Multi-model prediction for enhancing content locality in elastic server infrastructures, IEEE HiPC 2011.
[41] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, Temporal Locality in Today’s Content Caching: Why it Matters and How to Model it, ACM CCR, 2013.
[42] Y. Wang, Z. Li, G. Tyson, S. Uhlig, and G. Xie, Optimal Cache Allocation for Content-Centric Networking, IEEE ICNP 2013.
[43] J. Xu, M. Schaar, J. Liu, and H. Li, Timely Video Popularity Forecasting based on Social Networks, IEEE INFOCOM 2015.
[44] H. Yu, D. Zheng, B. Zhao, and W. Zheng, Understanding User Behavior in Large-Scale Video-on-Demand Systems, ACM EuroSys 2006.
[45] M. Zink, K. Suh, Y. Gu, and J. Kurose, Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications, Electronic Imaging 2008.

PLACE PHOTO HERE

Noriaki Kamiyama received his M.E. and Ph.D. degrees in communications engineering from Osaka University in 1994 and 1996, respectively. From 1996 to 1997, he was with the University of Southern California as a visiting researcher. He joined NTT Multimedia Network Laboratories in 1997, and he has been at NTT Network Technology Laboratories by 2016. He was also with the Osaka University as an invited associate professor from 2013 to 2014 and an invited professor in 2015. From 2017, he is a professor of Fukuoka University. He has been engaged in research concerning content distribution systems, network design, network economics, traffic measurement and analysis, traffic engineering, and optical networks. He received the best paper award at the IFIP/IEEE IM 2013. He is a member of IEEE, the Association for Computing Machinery (ACM), and the Institute of Electronics, Information and Communication Engineers (IEICE).

PLACE PHOTO HERE

Masayuki Murata received M.E. and Ph.D. degrees in information science and technology from Osaka University in 1984 and 1988. In April 1984, he joined the Tokyo Research Laboratory IBM Japan as a researcher. From September 1987 to January 1989, he was an assistant professor with the Computation Center, Osaka University. In February 1989, he moved to the Department of Information and Computer Sciences, Faculty of Engineering Science, Osaka University. From 1992 to 1999, he was an associate professor with the Graduate School of Engineering Science, Osaka University, and since April 1999, he has been a professor. He moved to the Graduate School of Information Science and Technology, Osaka University in April 2004. He has published more than 300 papers in international and domestic journals and conferences. His research interests include computer communication networks, performance modeling, and evaluation. He is a fellow of IEICE and a member of IEEE, the Association for Computing Machinery (ACM), The Internet Society, and IPSJ.