Modeling the Content Popularity Evolution in Video-on-Demand Systems

Attila Kőrösi[1], Balázs Székely[2], and Miklós Máté[1]

[1] Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics
H-1529 B.O. Box 91, Hungary
{korosi,mate}@tmit.bme.hu
[2] Institute of Mathematics
Budapest University of Technology and Economics
szbalazs@math.bme.hu

Abstract. The simulation and testing of Video-on-Demand (VoD) services require the generation of realistic content request patterns to emulate a virtual user base. The efficiency of these services depend on the popularity distribution of the video library, thus the traffic generators have to mimic the statistical properties of real life video requests. In this paper the connection among the content popularity descriptors of a generic VoD service is investigated. We provide an analytical model for the relationships among the most important popularity descriptors, such as the ordered long term popularity of the whole video library, the popularity evolutions and the initial popularity of the individual contents. Beyond the theoretical interest, our method provides a simple way of generating realistic request patterns for simulating or testing media servers.

Keywords: Video popularity, analytical model.

1 Introduction and Related Works

Building true Video-on-Demand (VoD) services with strong quality and availability grantees becomes feasible with the widespread adoption of broadband Internet access. The demand for VoD systems is high, as the customers are gradually turning away from scheduled broadcasts to personalized multimedia contents. VoD systems have high bandwidth requirements; therefore the effect of introducing a VoD service on the existing network must be examined through simulations before deployment, thus there is a strong need for accurate modeling of all components of a VoD. Perhaps the most important component of VoD systems are the clients, because the characteristics of the network traffic of the VoD largely depends on their content selections.

The long-term popularity distribution is the most important characteristic of a content library. The relative popularity of a content is defined with the number of requests for that content divided by the total number of requests in a (usually long) time interval. Content popularities are usually displayed in

R. Szabó et al. (Eds.): AccessNets 2010, LNICST 63, pp. 47–61, 2011.
© Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2011

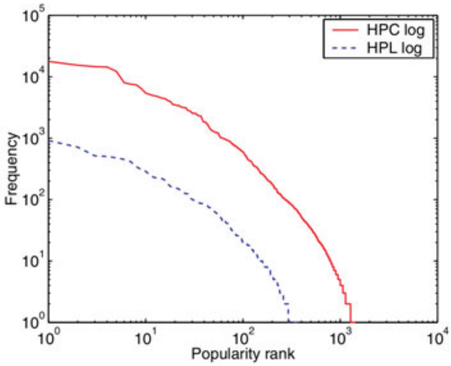


Fig. 1. Typical ordered long-term popularity distributions (source: [6])

decreasing order of popularity, and on a log-log scale, as in Figure 1. This curve is called *ordered long-term popularity*, which can be considered as a probability distribution, and it is usually modeled with a Zipf-like distribution based on empirical studies [1]. The standard Zipf distribution is linear on a log-log scale, but the real-world popularity distributions are not, thus several modifications were proposed to the Zipf distribution to fit the empirical data; such Zipf-like distributions include the Zipf-Mandelbrot law [4] and the k-transformation [6]. Recently, the use of the stretched exponential distribution has been suggested instead of a Zipf-like distribution [3]. We remark, that often, like in this paper, the absolute popularity is studied, without being divided by the total number of requests.

The daily popularity or short term popularity of a content library is also an important descriptor, because replicating frequently accessed items in a location closer to the clients is often required in order to decrease network bandwidth requirements. There are several such bandwidth-optimization schemes, ranging from simple caching to complex content delivery networks [5]. The efficiency of these solutions depends on the steepness of the popularity curve; if the majority of the requests are for a small number of contents, then a caching scheme can be very efficient.

The popularity evolution or lifespan, which is the timely change of the relative popularity of the individual contents. It is also interesting, because several caching optimizations depend on the prediction of the popularity changes. The most common one is when a content, which is expected to become popular in the near future, is inserted into the caches (precaching). Therefore, it is important to analyse the properties and reasons of the short-term popularity changes, and their connection to the long-term popularity distribution. The most commonly observed popularity evolution curve shows an increase immediately after the introduction, a short apex, and a long decrease [6], but other shapes have also been observed [7].

Contents can be classified into categories, based on the type of their popularity evolution. A quite common classification is the distinction between "news" and “movie” types. News are typically very popular for a short time after their introduction, but become obsolete very quickly. On the other hand, movies have smaller initial popularity, but remain relevant significantly longer.

In this paper we resolve the connections between the long-term popularity and the other popularity descriptors. These descriptors are the distribution of the video types, and the properties that depend on the type: the release day distribution, the popularity evolution, and the distribution of the initial popularity. As far as we know no such model is available in the literature.

The main contribution of our paper is the following. If one of the above parameters is unknown, an approximation of the missing parameter can be constructed. Our model can handle arbitrary long-term popularity distributions and the other parameters can be chosen arbitrarily – as long as they do not contradict to each other. Beyond the theoretical interest, our method provides a simple way of generating realistic request patterns for simulating or testing media servers.

The rest of the paper is organized as follows. In Section 2 we introduce the model and the notions appearing in the paper, along with the connection among the popularity descriptors. In Section 3 we show how the missing parameter can be approximated. In Section 4 we describe how our model can be used for generating user events, and compare our method to the ones found in the literature. Finally, in Section 5 we summarise our results and draw the conclusions.

2 Notations and the Popularity Model

In this section we introduce the main notations and describe the popularity model. Afterwards, we present the main relations that we use to derive the results in later sections.

2.1 Description of the Model

The observed period consists of D observation days, indexed with the set $\{1, 2, \ldots, D\}$, during which a total of N videos have been released. We assign four parameters to each video, namely:

type $\theta \in \Theta$ according to a given type distribution G on the set Θ.

The following three parameters depend on the type:

initial popularity I_θ is a positive real number valued *random variable*, which determines the number of claims for a video of type θ on the day it is released. The distribution of I_θ is denoted by F_θ.

popularity evolution function $h_\theta : \{1, 2, \ldots\} \to [0, \infty)$ is a *deterministic function*, which describes how the popularity changes for one video of type θ during its lifetime in the observed period. h_θ is an intrinsic parameter of the video, as it can be seen from the following definition. For $n \geq 1$ we define

$$h_\theta(n) := \frac{\#\text{ of claims for a video of type } \theta \text{ on day } n \text{ after its release}}{I_\theta} \qquad (1)$$

Consequently, the number of claims for a video of type θ on day n after its release is $I_\theta h_\theta(n)$.

release day d_θ from $\{1, 2, \dots, D\}$ according to a release day distribution $\{p_{\theta,d}, 1 \le d \le D\}$ depending on θ. Note that the observation days and the release days are indexed with the same set.

Remark 1. Instead of observation days we can take observation weeks. In this case the other parameters can be changed appropriately. For example the initial popularity counts the requests on the video during its first week in the system.

Based on the above definitions, video k $(1 \le k \le N)$ can be represented by its type θ_k and the starting day $d_k := d_{\theta_k}$, thus the popularity evolution function of video k is $h_k := h_{\theta_k}$, and its initial popularity is $I_k := I_{\theta_k}$.

Definition 1. *Let X_k denote the long term popularity of video k $(1 \le k \le N)$, that is, the number of claims for video k, introduced on day d_k, during the observed period of D days.*

It is easy to see that the following equation holds:

$$X_k = I_k \sum_{m=1}^{D-d_k+1} h_k(m). \tag{2}$$

Since X_k depends on the random variables (I_k, d_k, h_k) and $(I_k, d_k, h_k), k = 1, \dots, N$ is a sequence of independent and identically distributed (i.i.d.) random variables (they were generated independently) it can be seen that the $X_1, \dots, X_N$ long term popularities are also i.i.d. random variables. (h_k is also a random variable since $h_k = h_{\theta_k}$ and θ_k is a random variable.)

For further reference, we define the overall intrinsic popularity of video k over the whole observed period. First, we introduce the aggregated intrinsic popularity of video k during its first n days:

$$H_k(n) := \sum_{m=1}^{n} h_k(m). \tag{3}$$

Observe that $H_k(n)$ is an increasing function of n and $H_k(1) = h_k(1) = 1$. Using this notation we can write

$$X_k = I_k H_k(D - d_k + 1)$$

since the video k is added on day d_k so it is in the system for $D - d_k + 1$ days.

2.2 Long Term Popularity Parameters

Now we introduce two parameters that describe the global behavior of the system:

long term popularity curve $\Pi : \{1, 2, \ldots\} \to [0, \infty)$ is a decreasing deterministic function. For an appropriately long period we count the number of claims for each video and put these numbers in decreasing order, thus $\Pi(i)$ is the number of claims for the ith most popular video in that period.

long term popularity cdf Φ is defined as

$$\Phi(x) := \mathbf{P}(X_k \leq x), \quad x > 0. \tag{4}$$

for every video. It is the *cumulative distribution function* (cdf) of the number of claims arrive for a randomly chosen video during a long period. Since $X_1, X_2, \ldots, X_N$ are independent and identically distributed random variables by definition, we can omit the index k from $\Phi_k(x)$.

There exists a one-to-one correspondence between the long term parameters in the following sense. Before presenting the precise statement let us recall that the empirical distribution function of the sample $X_1, \ldots, X_N$ generated independently from the distribution Φ is defined by

$$\Phi_N(x) := \frac{1}{N} \sum_{k=1}^{N} \mathbf{I}\{X_k \leq x\}.$$

Moreover, by definition, Π is the ordered sample of $X_1, X_2, \ldots, X_N$.

Proposition 1. *Let the long term popularities $X_1, X_2, \ldots, X_N$ be a sequence of independent random variables with common distribution Φ.*

Then

$$\Phi_N(x) = \frac{N - \Pi^{-1}(x)}{N},$$

where $\Pi^{-1}(x)$ denotes the generalized inverse, $\Pi^{-1}(x) = \sup\{i : \Pi(i) \geq x\}$.

If N is large then Φ_N is close to Φ, since the empirical distribution function converges to the original distribution function as the sample size (N) increases. Thus

$$\Phi(x) \approx \frac{N - \Pi^{-1}(x)}{N}. \tag{5}$$

In the rest of the paper we will use Φ to describe the long term popularity, since Eq. (5) gives a simple relation between Φ and Π. Further, using Φ instead of Π is better for modeling purposes, since Eq. (2) describes the connection between the X_k long term popularity and the other parameters, and Φ is the distribution of X_k. This connection among the distribution functions will be discussed in Sec. 2.3 in detail.

Proof (for Proposition 1). The popularity curve shows that for the ith most popular video there have been $\Pi(i)$ claims in the long-run. The inverse of Π shows that no more than $\Pi^{-1}(x)$ videos have x or more claims. This means that less than $N - \Pi^{-1}(x)$ videos had less than x claims. Since the number of videos is N, the portion of videos that have less than x claims is $\frac{N-\Pi^{-1}(x)}{N}$. If the popularities of the N videos are equal to the N elements (unordered) sample $X_1, X_2, \ldots, X_N$ of Φ then by the definition of Φ_N one yields $\Phi_N(x) = \frac{N-\Pi^{-1}(x)}{N}$.

2.3 Long Term Popularity as a Function of the Parameters

In this subsection we express Φ using the type distribution (G), release day distribution ($\{p_{\theta,d},\ d = 1, 2, \ldots, D\}$), the initial popularity distributions (F_θ) and the popularity changes (h_θ).

Recalling the definition of Φ in equation (4) we have

$$\begin{aligned}\Phi(x) = \mathbf{P}(X \le x) &= \int_\Theta \mathbf{P}(X \le x|\theta) G(\,\mathrm{d}\theta) \\ &= \int_\Theta \sum_{d=1}^{n} \mathbf{P}(X \le x|\theta, d) p_{\theta,d} G(\,\mathrm{d}\theta).\end{aligned}$$

Using the interpretation (2), the definition of H (3) and that the distribution of I_θ is denoted by F_θ, the last term in the previous formula equals

$$\begin{aligned}\Phi(x) &= \int_\Theta \sum_{d=1}^{n} \mathbf{P}(I_\theta H_\theta(D-d+1) \le x) p_{\theta,d} G(\,\mathrm{d}\theta) \\ &= \int_\Theta \sum_{d=1}^{n} \mathbf{P}\left(I_\theta \le \frac{x}{H_\theta(D-d+1)}\right) p_{\theta,d} G(\,\mathrm{d}\theta) \\ &= \int_\Theta \sum_{d=1}^{n} F_\theta\left(\frac{x}{H_\theta(D-d+1)}\right) p_{\theta,d} G(\,\mathrm{d}\theta).\end{aligned}$$

Thus we have

$$\Phi(x) = \int_\Theta \sum_{d=1}^{n} F_\theta\left(\frac{x}{H_\theta(D-d+1)}\right) p_{\theta,d} G(\,\mathrm{d}\theta). \tag{6}$$

Definition 2. *We say that the model is well defined, if the given parameters (Φ, G, $\{F_\theta\}$, $\{h_\theta\}$ and $\{p_{\theta,d}\}$) satisfy equation (6).*

The functional equation (6) will be used in Section 3, for computing the missing parameter functions.

3 Connections among the Popularity Descriptors

In this section we assume that the number of types T is finite. We will show how the missing parameter can be approximated if the other parameters are known.

First, observe that if T is finite then the equation (6) can be written in the form

$$\Phi(x) = \sum_{\theta=1}^{T} g_\theta \sum_{d=1}^{D} p_{\theta,d} F_\theta\left(\frac{x}{H_\theta(D-d+1)}\right), \tag{7}$$

where g_θ denotes the probability that G concentrates to type θ, $\theta = 1, 2, \ldots, T$. We will solve the four implicit problems for the missing F_θ, h_θ, $p_{\theta,d}$, g_θ cases, each of them will be presented in separate subsections. The problems being implicit means that we always suppose that Φ is known.

3.1 Approximation of the Initial Popularities

In this section we determine the suitable initial popularity distributions (F_θ) in case the popularity change functions (h_θ), the type distribution (g_θ), the release day distribution (p_d) and the long term popularity, Π or Φ, are given.

We use equation (7) for obtaining numerical approximations for F_θ, $\theta \in \Theta$. We will determine the cdf of the initial distributions of several fixed points by solving a Linear Programming (LP) problem [2].

Let the set of *base points* $\{x_1, x_2, \ldots, x_L\}$ be given in increasing order. These are the points at which the quality of the approximation of Φ will be checked. The variables of the LP problem are

$$f_{\theta,i,d} = F_\theta(\tfrac{x_i}{H_\theta(d)}) \ (1 \le \theta \le T,\, 1 \le i \le L,\, 1 \le d \le D).$$

For fixed $f_{\theta,i,d}$ the approximation of Φ at points x_i is given by the following L equations:

$$(\forall i)\ \widetilde{\Phi}(x_i) = \sum_{\theta=1}^{T} g_\theta \sum_{d=1}^{D} p_{\theta,d} f_{\theta,i,d}$$

We have to ensure that the variables $f_{\theta,i,d}$ determine the distribution functions for any θ. Therefore, we assume that if $\frac{x_i}{H_\theta(d_1)} \le \frac{x_j}{H_\theta(d_2)}$ then $f_{\theta,i,d_1} \le f_{\theta,j,d_2}$ for any type θ. Thus, we can define the following LP problem:

$$\min \varepsilon$$
$$(\forall i) : -\varepsilon \le \Phi(x_i) - \widetilde{\Phi}(x_i) \le \varepsilon$$
$$\left(\forall \theta, i, j, d_1, d_2 \text{ such that } \tfrac{x_i}{H_\theta(d_1)} \le \tfrac{x_j}{H_\theta(d_2)}\right) : 0 \le f_{\theta,i,d_1} \le f_{\theta,j,d_2} \le 1$$

Solving this problem yields the best approximating initial popularity functions.

To illustrate how the method works we present the following example.

Example 1. (The figures related to this example are shown in Figure 2 and 3) Let $T = 2$, $g_\theta \in \{0.6, 0.4\}$, and $h_1(i) = \frac{(50+i)^{-3}}{51^{-3}}$, $h_2(i) = \frac{2^{-i}}{1/2}$. The evolution functions are fundamentally different: h_1 has power law decay (represents regular movies), while h_2 has exponential decay (represents news like videos). Further, let $F_\theta(x) = 1 - (1 - x)^{-\alpha_\theta}$, with $\alpha_\theta \in \{8, 2\}$. The release day distribution is uniform over the observed period for any type that consists of $D = 50$ weeks according to Remark 1. There were $N = 5000$ different movies.

We generate the long-term popularity curve Φ using these parameters and equation (7). Then we solve the problem of finding F_θ, $\theta = 1, 2$ with the obtained Φ. Then we compare the approximated initial popularity distributions with the given ones and we also compare the generated Φ based on the original functions and Φ generated from h's and the approximated F's.

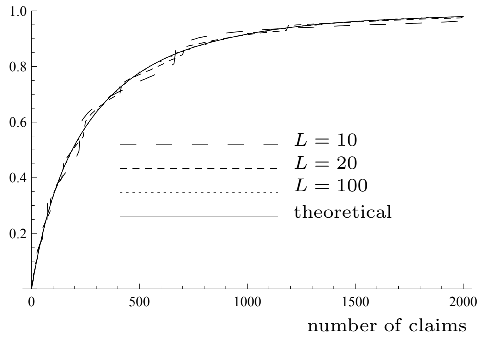


Fig. 2. The distribution function Φ and approximated Φ if $(F_\theta,\ \theta \in \Theta)$ is the missing parameter

The quality of the approximation was investigated for three different base point sets. For $L = 10, 20, 100$ the base points are $x_i = x_{min} \left(\frac{x_{max}}{x_{min}}\right)^{\frac{i-1}{L-1}}$, where $x_{min} = 0.1$ and $x_{max} = 2000$. The base points are chosen so that the increment of Φ between two neighboring x_i's is constant. Of course, one may use other base point set but in our experience this kind of set provides fairy good approximation not only for Φ but the initial popularity distributions $F_\theta, \theta \in \Theta$ as well.

In Figures 2 and 3 the original parameters and the three approximated parameters are depicted for Φ, F_1 and F_2. The difference between the original and the approximated Φ at the base points $x_1, \ldots, x_L$ is at most 10^{-9}. The difference at the other points depends on the approximation of F_θ's. We approximated F_θ's by using jump functions (and not linear approximates) that causes fairly big errors. However, the difference between the original and the approximated Φ never exceeds 0.05, 0.02, 0.004 in cases $L = 10, 20, 100$ respectively.

3.2 Approximation of the Popularity Changes

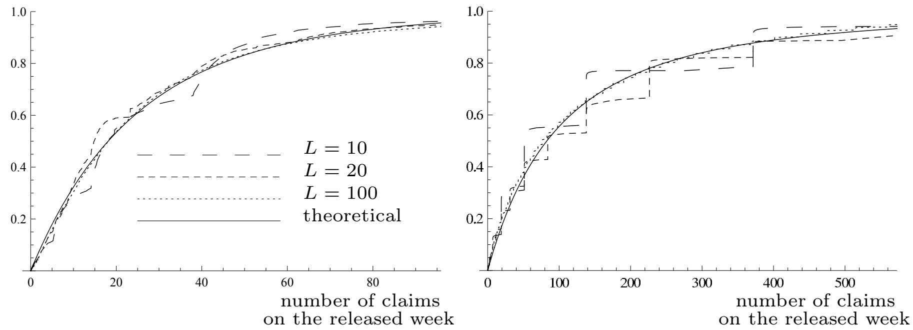
In this section we will approximate the popularity evolution functions, considering that the long-term Φ and the initial F_θ popularity distributions are given for T types. We will use equation (6) and, like in the previous subsection, the solution, $h_\theta, \theta \in \Theta$, will be the solution of an appropriate LP problem. We will minimize the error ε:

$$\varepsilon = \sup_{x \in \underline{x}} \left| \sum_{\theta=1}^{T} g_\theta \sum_{d=1}^{D} p_d F_\theta \left(\frac{x}{H_\theta(D-d+1)} \right) - \Phi(x) \right| ,$$

where $\underline{x} = \{x_1, \ldots, x_L\}$ is a set of given points. We first describe the idea of the approximation for $T = 1$, then we present the general case for $T > 1$.

Assume $T = 1$. Using the notation $p_d = p_{\theta,d}$, we have

$$\sum_{d=1}^{D} p_d F \left(\frac{x}{H(D-d+1)} \right) = \Phi(x).$$



(a) Theoretical and approximated F_1 (b) Theoretical and approximated F_2

Fig. 3. Approximation of the initial popularity distributions F_θ, $\theta \in \Theta$. Two types with different initial popularity distributions. The details are given in Example 1. The accuracy of the approximations increase as more base points are added to the LP problem.

The sum can be interpreted as a mean of the discrete distribution Q, which is concentrated on the points $\frac{1}{H(D-d+1)}$, $d = 1, \ldots, D$, and point $\frac{1}{H(D-d+1)}$ has probability p_d. Thus, for a random variable X, with distribution Q, we can write:

$$\sum_{d=1}^{D} p_d F\left(\frac{x}{H(D-d+1)}\right) = \mathbf{E}_Q F\left(Xx\right), \tag{8}$$

where $\mathbf{E}_Q$ denotes the expectation with respect to Q. We will approximate Q by a recursive sequence of distributions, such that the

$$\sup_{x \in \underline{x}} |F_Q(x) - F_{\widehat{Q}}(x)|. \tag{9}$$

distance between the distributions of Q and its $\widehat{Q}$ approximations is small. The first element of the sequence is denoted by $\widehat{Q}^1$, which is a jump function with jumps at the arbitrarily chosen $\underline{r}^1 = (r_1, r_2, \ldots, r_K)$ points. We want to assign weights $\underline{s}^1 = (s_1, s_2, \ldots, s_K)$ as $(\widehat{Q}^1(\{r_i\}) = s_i)$ to make sure that the distance

$$\sup_{x \in \underline{x}} \left| \varPhi(x) - \sum_{j=1}^{K} s_j F\left(r_j x\right) \right| \tag{10}$$

is minimal. This will be accomplished via the following LP problem:

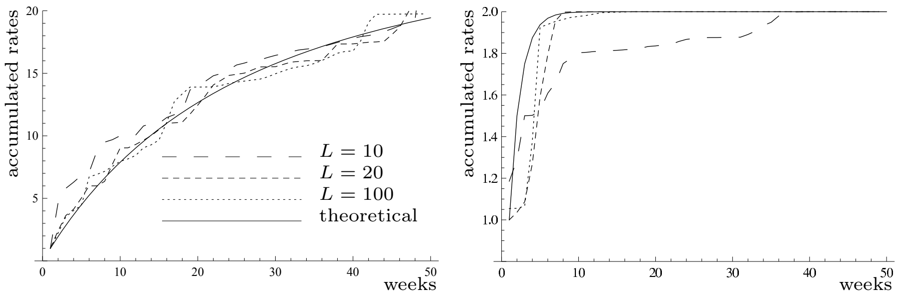
$$\begin{gathered} \min \epsilon \\ (\forall i) \quad -\epsilon \leq \textstyle\sum_{j=1}^{K} s_j F\left(r_j x_i\right) - \varPhi(x_i) \leq \epsilon \\ \textstyle\sum_j s_j = 1 \\ 0 \leq s_1, \ldots, 0 \leq s_K \end{gathered}$$

This solution for $\underline{s}^{1*}$ for given $\underline{r}^1$ minimizes (10). Now, we do the following heuristically reasonable refinement step: to get better approximation in the sense of (9) we add new jump points to $\underline{r}^1$. First, take an intermediate distribution B concentrated on $1 = z_1 > z_2 > \cdots > z_D$ and each z_n carries p_n weight. Then, we solve the LP problem that minimizes $\sup_{x\in\underline{x}} |F_B(x) - F_{\widehat{Q}^1}(x)|$. The solution $1 = z_1 > z_2^* > \cdots > z_D^*$ corresponds to the best H^* such that if $H^*(n) = 1/z_n^*$, then $\sum_{d=1}^{D} p_d F\left(\frac{x}{H^*(D-d+1)}\right)$ is the closest function of this form to $\sum_{j=1}^{K} s_j^1 F(r_j x_i)$ on the set $\{x_1, \ldots, x_L\}$. Second, we add the set $\underline{z}^* = \{z_1^*, z_2^*, \ldots, z_{D-1}^*\}$ to $\underline{r}^1$ and start the approximating procedure described above with $\underline{r}^2 = \underline{r}^1 \cup \underline{z}^*$. Similarly, we can construct $\widehat{Q}^2$. We continue this refining procedure until we obtain a satisfactory collection of h_θ such that Φ and the approximated Φ is close enough.

Next, suppose that $T > 1$. For a given $\underline{r}_\theta = (r_{\theta,1}, \ldots, r_{\theta,K})$, $\theta = 1, \ldots, T$ and given type distribution $g_\theta, 1 \leq \theta \leq T$ we want to find the best $\underline{s}_\theta = (s_{\theta,1}, \ldots, s_{\theta,K})$, $\theta = 1, \ldots, T$ in the sense that the failure of the approximation is minimal, that is,

$$\min \epsilon$$
$$-\epsilon \leq \sum_{\theta=1}^{T} g_\theta \sum_{j=1}^{K} s_{\theta,j} F_\theta(r_{\theta,j} x_i) - \Phi(x_i) \leq \epsilon,\ (\forall i)$$
$$\sum_j s_{\theta,j} = 1,\ (\forall \theta)$$
$$0 \leq s_{\theta,1}, \ldots, 0 \leq s_{\theta,K},\ (\forall \theta)$$

As in the case $T = 1$, for each θ separately we add new points $\underline{z}_\theta^*$ to $\underline{r}_\theta$ by constructing the best approximating intermediate distribution B_θ. We iterate this refinement technique until we obtain a satisfactory collection of h_θ.



(a) Theoretical and approximated H_1 (b) Theoretical and approximated H_2

Fig. 4. H_θ and approximated H_θ, $\theta = 1, 2$. The details are given in Example 1 and 2.

Since the algorithm presented above employs heuristic considerations it is worth verifying it through an example.

Example 2. (The figures related to this example are shown in Figure 4.) The theoretical parameters are the same as in Example 1, and the solution follows the same pattern. We generate the long-term popularity curve Φ using the given parameters and equation (7), then we solve the problem of finding H_θ, $\theta = 1, 2$ with the obtained Φ. Then we compare the approximated popularity change functions with the given ones and we also compare the generated Φ based on the original functions and Φ generated from the original F's and the approximated H's. Although the convergence to the theoretical popularity changes is not guaranteed, because of the heuristic approximation method, the approximated popularity distribution Φ converges to the theoretical one.

3.3 Approximation of the Release Day Distribution

In this subsection we will show how the $p_{\theta,d}$ release day distribution can be computed from known Φ, F_θ, h_θ and G. Similarly to the previous subsections, we will use equation (6):

$$\Phi(x) = \sum_\theta \sum_d p_{\theta,d} g_\theta F_\theta \left(\frac{x}{H_\theta(D - d + 1)} \right) \tag{11}$$

We solve an LP problem at points $x = x_i$, $1 \leq i \leq K$ and solved for $p_{\theta,d}$ with the bounds $0 \leq p_{\theta,d} \leq 1$ for any θ and $\sum_d p_{\theta,d} = 1$.

3.4 Approximation of the Type Distribution

In this subsection we will investigate how the g_θ type distribution can be computed from known Φ, F_θ and h_θ. The solution is quite simple. Generate the functions ϕ_θ from the functions F_θ and h_θ by using equation (6) as though there was only one type for each θ above. Using equation (6) again, we have the following equation:

$$\Phi(x) = \sum_\theta g_\theta \sum_d p_d F_\theta \left(\frac{x}{H_\theta(D - d + 1)} \right) = \sum_\theta g_\theta \Phi_\theta(x) \tag{12}$$

Now, for finding g_θ we have to solve an LP problem in some points $\{x_1, \ldots, x_L\}$ similar ways as in the previous sections.

Remark 2. On the accuracy of the approximations. In case of finding F_θ, $\theta \in \Theta$ and finding h_θ, $\theta \in \Theta$ there is no guarantee for convergence. However, for certain parameter combinations the approximations get very close tho the theoretical values. The accuracy of the approximation highly depends on the number of base points $x_1, x_2, \ldots, x_L$. In Figure 3 and 4 it can be seen that the approximations of F get closer to the theoretical ones while the approximations of H does not. In case of finding the type distribution and the release day distribution the approximations typically converge. The explanation is that if the functions Φ_θ, $\theta = 1, \ldots, T$ are not pairwise equal, then typically we can find $L = T$ points $\{x_1, \ldots, X_T\}$ such that the vectors $[\Phi_\theta(x_1), \ldots, \Phi_\theta(x_T)]^t$, $\theta = 1, \ldots, T$ are linearly independent. Consequently, the system of equations in (12) has one unique solution. The same argument can be repeated for the convergence of the release day distribution. If one can find $L = TD$ points such that the vectors $\left[F_\theta\left(\frac{x_i}{H_\theta(D-d+1)}\right), i = 1, \ldots, TD\right]^t$, $\theta = 1, \ldots, T, d = 1, \ldots, D$ are linearly independent then the system of equations in (11) has one unique solution. The condition typically holds if the function F_θ, $\theta = 1, \ldots, T$ and the functions H_θ, $\theta = 1, \ldots, T$ are pairwise different.

Remark 3. If there is exactly one more unknown parameter beyond the type distribution g_θ, a Non Linear Programming (NLP) problem can be written with linear constrains and fourth degree objective function.

Proof. Let $\varepsilon_{\theta,i}$ be the difference of the approximation of type θ from Φ in x_i:

$$\varepsilon_{\theta,i} = \Phi(x_i) - \sum_{d=1}^{D} p_{\theta,d} F_\theta\left(\frac{x_i}{H_\theta(D-d+1)}\right).$$

Then the error of the approximation in l_2 norm in $x_1, x_2, \ldots, x_L$ is

$$\sqrt{\sum_{i=1}^{L}\left(\sum_{\theta\in\Theta} g_\theta \varepsilon_{\theta,i}\right)^2}.$$

This error is minimal if its square is minimal, thus the NLP problem for finding the two unknown parameters is

$$\begin{gathered}\min \textstyle\sum_{i=1}^{L}\left(\sum_{\theta\in\Theta} g_\theta \varepsilon_{\theta,i}\right)^2 \\ \varepsilon_{\theta,i} = \Phi(x_i) - \textstyle\sum_{d=1}^{D} p_{\theta,d} F_\theta\left(\frac{x_i}{H_\theta(D-d+1)}\right) \ (\forall\theta\forall i) \\ \textstyle\sum_\theta g_\theta = 1 \\ 0 \le g_\theta \ (\forall\theta).\end{gathered}$$

4 Client Requests Generation

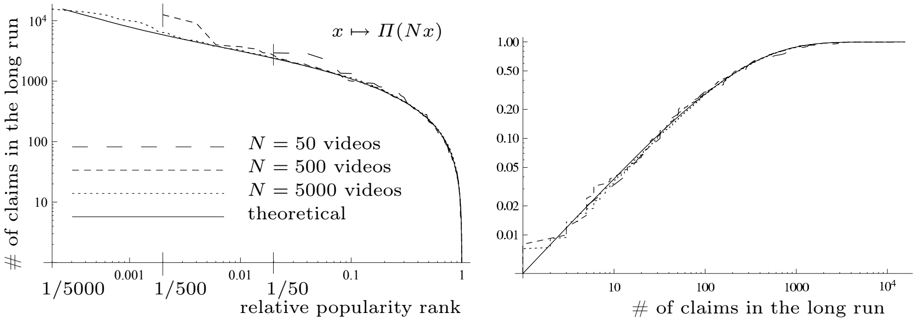
In this section we demonstrate how our model can be used to generate a series of client requests for testing or simulating a VoD system. The number of requests and their timely distribution will all follow the given distributions, because they are optimized independently.

Assume that the type distribution $\{g_\theta\}$, the release day distribution $\{p_d\}$, the initial popularities $\{F_\theta\}$ and the popularity changes $\{h_\theta\}$ are given. The generating method is simple, since the construction is designed to solve this problem easily: to each video k we generate such release day d_k, evolution type h_k and initial popularity I_k, that the number of claims on day d for video k is $\mathbf{I}\{d \ge d_k\} I_k h_k(d - d_k + 1)$.

The distribution of the requests within the observation period is fairly easy, as, according to other studies [6,7], its distribution is independent of the other popularity descriptors. After the number of requests has been calculated for a given period, their exact timing can be determined using the given intensity distribution. If the observation period is one day, then this distribution is usually called diurnal access pattern, which has usually its maximum in the evening, and its minimum during the night.Similar recurring request intensity changes can be observed over weeks as well.

To overcome the problem that the number of claims for a day can be fractional, because we do not require that h_θ is integer valued, we take either $\lfloor I_k h_k(d - d_k + 1) \rfloor$ or $\lfloor I_k h_k(d - d_k + 1) \rfloor + 1$ according some probability distribution, while ensuring that the sum of these integers is exactly $\lfloor X_k \rfloor = \lfloor I_k H_k(D_k) \rfloor$. This can be done very easily.

The long-term distribution $\varPhi$ of the simulated system (the empirical distribution) converges to the theoretical $\varPhi$ because of Proposition 1. Figure 5 shows that the empirical distribution and the theoretical distribution are close to each other and the simulated long-term popularity curves also approximate the theoretic one. The continuous line is the analytical result, the dashed curves show the cases, when the number of videos in the system is $50, 500, 5000$ in the scenario described in Example 1. The difference of $\varPhi(x)$ and the approximated $\varPhi$ at any x is not larger than 10^{-2} (50 videos), 10^{-3}(500 videos), and 10^{-4} (5000 videos).



(a) Ordered long run popularity scaled to 1 $\varPi(N\cdot)$ on log-log scale

(b) Long term popularity distribution $\varPhi = \frac{N - \varPi^{-1}}{N}$ on log-log scale

Fig. 5. Simulation results. The differences between the long run parameters ($\varPi$, $\varPhi$) of the simulated system and the theoretical parameters decrease as the number of videos increases. The relative popularity in Figure (a) means that on the x-axes the numbers x/N are depicted for $x = 1, 2, \ldots, N$ ($N = 50, 500, 5000$), where x denotes the popularity rank of the video in decreasing order.

Our method is comparable to the method of Medisyn [6]. Medisyn starts the request generation with a given long-term popularity curve $\varPi$, then, for each video in the library, it generates a random type, which can be "news" or "movie". The probability of a video being "news" depends on the popularity of the video. In their measurements the authors found that the "news" type videos tend to be more popular than the "movie" type ones, therefore they included this bias in their generator. Once the type is known for a video, it generates a release day according to the given release day distribution (the authors of Medisyn consider both the intensity and the interval between the releases). Then it selects a life span function (popularity evolution) for the video with randomly chosen parameter. This function is from an exponential family (its parameter is Pareto distributed) for the "news" type videos and from a lognormal family (the parameter is normally distributed) for the "movie" type ones. Finally, the total number of requests is distributed along the timeline according to the release day of the video and its life span function. In this way the initial popularity defined in our model is also obtained implicitly. Therefore, irrespectively of the randomly selected life span, Medisyn solves the problem presented in Section 2.3.

5 Conclusions

We provided a stochastic model for finding the relationships among the following popularity descriptors: (1) the ordered long-term popularity, (2) video type distribution, (3) release day distribution, (4) the distribution of the initial popularity of each individual video and (5) the popularity change over time for each individual video.

An important feature of our model is the possibility of constructing an approximation of any missing popularity descriptor, unless the conditions contradict to each other. The missing parameter is the solution of an appropriate LP problem in all four cases (the ordered long-term popularity does not need to be approximated), thus we have four similar, but not identical approximation schemes.

The two most important out of the four problems, from practical point of view, are finding the initial popularity distribution and the popularity evolution for the content types. As the examples have shown, the approximation works well for finding the initial popularities, the results were very close to the original distribution. Finding suitable popularity evolution functions is much harder, our procedure does not necessarily converge to the original functions. This is natural, since the popularity evolution has great degree of freedom.

Our model is designed so that one can easily generate realistic request patterns for simulating or testing media servers. We have shown that the more videos there are in the VoD system, the parameters in the simulated system get closer to the theoretical ones.

In the future we want increase the accuracy of our approximations, and try to find exact solutions for the missing parameters in special cases. We are also interested in finding a way to modify the model in order to take randomly occurring jumps in the popularity evolution into account.

Acknowledgments. The work has been supported by HSNLab, Budapest University of Technology and Economics, http://www.hsnlab.hu

References

1. Breslau, L., Cao, P., Fan, L., Phillips, G., Shenker, S.: Web caching and zipf-like distributions: evidence and implications. In: Proceedings of Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 1, pp. 126–134. IEEE, Los Alamitos (1999)
2. Dantzig, G.B., Thapa, M.N.: Linear programming 1: Introduction. Springer, Heidelberg (1997)
3. Guo, L., Tan, E., Chen, S., Xiao, Z., Zhang, X.: The stretched exponential distribution of internet media access patterns. In: Proc. of PODC 2008, Toronto, Canada (August 2008)
4. Mandelbrot, B.: Information Theory and Psycholinguistics. Penguin Books (1968)
5. Pallis, G., Vakali, A.: Insight and perspectives for content delivery networks. Communications of the ACM, 101–106 (January 2006)
6. Tang, W., Fu, Y., Cherkasova, L., Vahdat, A.: Modeling and generating realistic streaming media server workloads. Comput. Netw. 51(1), 336–356 (2007)
7. Yu, H., Zheng, D., Zhao, B.Y., Zheng, W.: Understanding user behavior in large-scale video-on-demand systems. In: Proc. of Eurosys 2006, Leuven, Belgium, pp. 333–344 (2006)