A Semi-Markovian Individual Model of Users for P2P Video Streaming Applications
Grégory Bonnet, Ihsan Ullah, Guillaume Doyen, Lionel Fillatre, Dominique Gaïti, Igor Nikiforov

To cite this version:

HAL Id: hal-00952019
https://hal.archives-ouvertes.fr/hal-00952019
Submitted on 26 Feb 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A Semi-Markovian Individual Model of Users for P2P Video Streaming Applications

Grégory Bonnet\textsuperscript{1}, Ihsan Ullah\textsuperscript{2}, Guillaume Doyen\textsuperscript{2}, Lionel Fillatre\textsuperscript{3}, Dominique Gait\textsuperscript{2} and Igor Nikiforov\textsuperscript{3}

\textsuperscript{1}University of Caen Lower-Normandy (GREYC) - \textsuperscript{2}Troyes University of Technology
\textsuperscript{3}UMR STRM 6279 ICD
\textsuperscript{1}MAD Team - \textsuperscript{2}ERA Team - \textsuperscript{3}LM2S Team
FRANCE
email: \textsuperscript{1}gregory.bonnet@unicaen.fr - \textsuperscript{2}name.surname@utt.fr

Abstract—P2P-based live video streaming has become one of the major applications on the Internet. Due to the importance of the user behavior in these systems, intensive measurement studies have been performed over them in order to provide global models. However, in real networks the individual user behavior does not necessarily follow the global or average behavior. As the measurement studies face difficulties to establish the link between a trace and an individual user due to both technical and privacy issues, we propose individual models of users inspired from fictional characters who represent different kinds of consistent behaviors. In this paper, we define such models with a non-homogenous semi-markovian process and give validation results that show that our model is consistent with global models in terms of population and session duration.

I. INTRODUCTION

P2P-based live video streaming allows cooperating end-hosts to self-organize into an overlay network. Peers share their computing and upload resources by caching and relaying the video content to each other. However, these systems still suffer from performance problems such as startup and playback delays. Since peers depend upon each other, activities of users have a direct impact over the performance of these systems. Therefore, due to the importance of the user behavior in P2P streaming systems, numerous intensive measurement studies have been performed over them in order to understand it. Based on their results, these studies aim at better designing and controlling video streaming systems.

If the measurement studies provide global models, they face difficulties in establishing the link between all traces and individual users due to dynamic IP addresses and peer identifiers, presence of NAT and privacy rules. Moreover, the individual user behavior does not follow the global behavior and a global model is not suitable for dealing with users having different preferences, interests and habits. Therefore, individual models are required in several perspectives, from their integration in simulation tools and workload generation, resource allocation management, or the design of user-centric systems. In this paper, we propose such models of users inspired from fictional characters \cite{1} who represent different kinds of behaviors. We give validation results that show that our model is consistent with measurement-based global models.

Section II presents the related work about measurement campaigns and sociological studies. Section III presents the abstract individual model and its parameters inspired from the personas designed by the On-Demand Project\textsuperscript{1}. We show in Section IV that our model fits well the global behavior in the litterature and we highlight its limits. Section V draws conclusions and gives directions for the future work.

II. RELATED WORK

A. Large-scale measurement campaigns

For a few years, the strong adoption of video streaming have lead to the carry out of numerous and massive measurements campaigns spanning world wide systems and long term periods \cite{2}–\cite{7}. Table I presents an overview of the different models extracted from these measurements.

In the matter of population models, \cite{3}–\cite{5}, \cite{7} agree on the fact that the user population is cyclic and this pattern is repeated every day with a peak in the middle of the week \cite{3} and an off-peak during the week-ends \cite{3}, \cite{4}. This weekly pattern is repeated with an off-peak during holiday. Concerning arrival models, exponential laws \cite{3}, \cite{7} and Poisson laws \cite{2}, \cite{5} are the most used. As for session duration, the models are mostly based on lognormal laws \cite{2}, \cite{5}, \cite{6} and exponential laws \cite{4}, \cite{7}. Concerning the popularity models, all studies \cite{3}–\cite{5}, \cite{7} agree on the fact that the Zipf law fits well to link popularity rank and number of requests. Nevertheless, the Zipf law does not fit well with extreme values: the most and the less popular video are over and under-estimated respectively.

B. Identifying individual behaviors

Due to both technical and privacy issues, all the measurement studies only focus on the global behavior of users because crawlers and passive methods cannot observe a given user from a session to another. At the oppposite, studies related to the definition of individual models of users rely on sociological approaches \cite{8}. Despite these works, it is difficult to extract formal behaviors applicable to video streaming system. For that reason, \cite{1} use the concept of persona for the On-Demand Project which aims at improving the QoS of scalable video-on-demand networks. It consists in defining fictional characters in terms of identity, skills, habits and goals that are consistent with the sociological literature. Although this work

\textsuperscript{1}http://www.sics.se/projects/ondemand_iptv
seems useful for individual models, it only presents qualitative descriptions (see the upper part of Table II) rather than a formal model.

To conclude, on the first hand, measurement studies propose only global models which can be far from individual behaviors. On the other hand, sociological and persona-based approaches lack of formalization. Consequently, we propose to consider the personas defined in [1] and use them to define formalized individual models of users.

III. A NON-HOMOGENEOUS SEMI-MARKOVIAN MODEL

We propose to modelize each user with a non-homogeneous semi-markovian process. It means that the state of a user (online or offline for instance) at a given time not only depends on its state at the previous time as in any markovian process but also on the time it spends in this state and on the global time of the process. Such kind of process fits well with the video streaming context because previous studies show that the behavior of a user vary with respect to the time of the day (e.g. watching longer or more often in the evening than in the morning) and the time he spent online impacts on the time it will stay [9].

A. The abstract individual model

We make the following assumptions:

1) We only consider mono-channel video streaming applications. Consequently, we define two states \( \{X_1, X_2\} \) where the semantics is respectively the user’s online presence and its offline presence;
2) We assume a cyclic behavior on each day. Consequently, we consider the process global time \( t \in \mathbb{N}^+ \) as a day discretized in one-minute intervals, \( \{t_1, \ldots, t_{1440}\} \);
3) In order to be consistent with [2], [4]–[6] and as log-normal laws are commonly used to modelize slow fading phenomenon, we consider that the transition from state \( X_1 \) to state \( X_2 \) is controlled by a log-normal law. As an individual user watches longer in certain time-of-the-day, the \( \mu_i^t(d) \) and \( \sigma_i^t(d) \) depend on the persona \( i \), on the kind of content \( d \), on the global time of the process \( t \) and on \( t_{X_1} \) the time spent in the state \( X_1 \);
4) As Poisson laws are widely used to modelize arrival processes, we consider that the transition from state \( X_2 \) to state \( X_1 \) is controlled by a Poisson law which is consistent with [2], [5]. As an individual user has habits about its watching time, the parameter \( \lambda_i^t \) depends on the persona \( i \) and on the global time of the process \( t \).

The transition probability from state \( X_1 \) to state \( X_2 \) is the probability to be disconnected after a given time \( t_{X_1} \) spent in the state \( X_1 \) :

\[
P(X(t) = X_2|X_1, t_{X_1}) = \int \frac{e^{-\frac{(\log x - \mu_i^t(d))^2}{2 \sigma_i^t(d)^2}}}{x \sigma_i^t(d) \sqrt{2\pi}} \, dx
\]

where \( \mu_i^t(d) \) and \( \sigma_i^t(d) \) are the parameters of the log-normal law with respect to the persona \( i \), the global time \( t \) and the content \( d \). The transition probability from state \( X_2 \) to \( X_1 \) is the probability of being connected at a given time-of-the-day. We modelize this phenomenon with a Poisson law and, as we are interested in the arrival of a single user, \( k = 1 \). Consequently,

\[
P(X(t) = X_1|X_2) = e^{-\lambda_i^t \lambda_i^t}
\]

where \( \lambda_i^t \) is the parameter of the law with respect to the persona \( i \) and the global time \( t \).

B. Setting the parameters

In order to instanciate the parameters of the model, we base them on the personas defined by [1]. In the sequel, we denote these personas with \( \{J, E, S, A, P, L\} \), namely Johnatan, Emma, Stephan, Anna, Peter and ELlen. All the parameters and their values are summarized in Table II.

1) Arrival model: In conformity with measurements [3]–[5], [7], we define five time periods in a day, namely morning, noon, afternoon, evening and night. For each persona and each time period, we set the \( \lambda_i^t \) parameters of our model for a one-minute time step (see the second part of Table II). We choose those values such that it reflects the watching habits of each persona. For instance, Johnatan watches more often in the evening than in the morning. Moreover, the sum of these values is lesser than 1, meaning that a given user does not join the network every day. Finally, the sum of the \( \lambda_i^t \) approximate the global evolution of a population.

2) Session duration model: Each persona watches television each day for an average time as given in Table II. According to [1], \( J \) and \( A \) are very regular, \( E \) and \( L \) are regular, \( P \) is irregular and \( S \) is very irregular. This regularity is given by the variance \( \sigma^2 = \frac{x}{\bar{x}} \) where \( \bar{x} \) is the total time session per day (\( k = 1 \) for very irregular to 4 for very regular).
The third part of Table II gives the mean and the variance in minutes of the total online presence for each persona as well as the \( \mu_i \) and \( \sigma_i \) parameters of their lognormal laws.

As [1] solely give a watching duration per day for each persona, we need to define \( \mu_i^d(d) \) and \( \sigma_i^d(d) \) from \( \mu_i \) and \( \sigma_i \). Furthermore, the session time is correlated with the interest of the watcher in the content. To this end, we assume that the session time at a given time period (local session time) is a sharing of this watching duration per day multiplied by an interest coefficient defined from both the watching time habits and the socioprofessional group of each persona. The fourth part of Table II gives this sharing for each persona according to the time period. The values of these parameters are arbitrary defined but reflect the preferences of the personas. The fifth part of Table II gives the interest coefficient \( \theta_i^d \) for each persona \( i \) with respect to each content type \( d \). We choose the values of the coefficients such that their mean equals 1 which represents the fact that, in average, the expected session duration follows the previous distribution. Consequently, we defined \( \mu_i^d(d) \) to \( \sigma_i^d(d) \) as follows:

\[
\begin{align*}
\mu_i^d(d) &= \mu_i M_i^d \theta_i^d \\
\sigma_i^d(d) &= \sigma_i M_i^d \theta_i^d
\end{align*}
\]

3) Population model: As an individual model, a non-homogeneous semi-markovian process is instantiated for each user in the network. Consequently, we propose to set the proportion of the personas according to the data given by the French National Institute of Statistics and Economic Studies (INSEE). From these data, the proportion of a given persona type is given in the sixth part of Table II.

IV. VALIDATION RESULTS

To validate our model, we compare the different behaviors of a persona with respect to the different kinds of content. Then, we simulate a single persona many times throughout many days in order to highlight its mean behavior. Finally, we simulate a population of all the personas and show that the results fit with the global behavior established in the literature.

A. Impact of the content type on the behavior

In order to highlight the influence of the content on the persona’s behavior, we simulate 2000 users of a single persona type in the presence of each kind of content. Figure 1.a presents the total watching time on a day for each persona. We can notice that, if the personas may be similar for a given content as persona A and P for sports for instance, each of them exhibits its own distinctive behavior in terms of watching time on a day. Due to space constraints, we provide on Figure 1.b the CDF of such durations only for persona S. Each CDF has the same lognormal shape but different scales with respect to the kind of content. Figure 1.c shows the results for persona S in terms of session durations and population in the network.

If, as defined in the model, the content does not directly impact the population, it has an influence on it. When a user is not interested by a given content, its session durations are reduced and it is not as present in the network as it was interested. For instance, Figure 1.c shows the population of
persona $S$ with respect to a given content. For all content, $S$ is mostly present in the middle of the day, in the evening and at night. The flash crowd times are the same. However, fiction leads to very short sessions, reducing the number of online users and provoking greater oscillations in the population.

B. Global behavior of a persona

We simulate the behavior of 2000 users with the same persona on three days with a different kind of content each day and we compute the average behavior. In this paper, we only presents results for personas $J$ and $A$.

Figure 2.a presents the results for persona $J$. Its mean total online presence is 141.43 minutes with a mean session duration of 32.49. It is always consistent with its global behavior given in Table II. The global population is higher during evening and night, which is consistent with its watching habits. Figure 2.b presents the results for persona $A$. Its mean total online presence per day is 89.02 minutes with a mean session duration of 18.85. It is always consistent with its global behavior given in Table II and $A$ is the most present during afternoon and evening, which is consistent with its habits that states that it watches television with its children.

In both simulations, the session duration frequencies on Figure 2 follow a multimodal distribution. These modes represent the mean session duration for each kind of content with respect to the interest coefficient. Obviously some modes are merged; so the number of modes vary for each persona. In both cases, we observe a long tail distribution that is highlighted by the CDF and that is consistent with the literature.

C. Global behavior of a population of users

We set 10000 users with a persona in the proportion given in Table II. We simulate three days with a realistic mixed content each day and we compute the average behavior.

Figure 3.a represents the frequency of the session durations of the population. As previously, it presents a multimodal distribution where each mode highlights a different behavior with respect to a given content and a given interest coefficient. The distribution shows a high number of short sessions and a long tail of longer session which is consistent with the lognormal model of [10].

Figure 3.b compares the cumulative distribution function (CDF) of the session duration of our model to the global models proposed by [5], [6], [10]. We can notice that our model fits well with the global model of [5] and have the same shape than the others. Indeed, Jensen-Shannon divergence and the symmetric Kullback-Leibler divergence with [5] is only 0.091486 and 0.48833 respectively. However, [6], [10] present more very short sessions. An explanation is that the authors focus on reality shows and soccer where the users watch very short scenes (cues or goal kicks). Hence, they present more very short sessions.

Figure 3.c represents the population throughout the day. It shows that our model is consistent with the global model presented in [9]–[11]: there is a low population in the morning, a first peak at noon, a decrease in the afternoon, an apex in the evening and a low population at night. The sudden changes in the population represent flash crowds, a common phenomenon in P2P streaming systems.

D. Limits of our model

A first set of limits are induced by the model itself. Firstly, our model only considers mono-channel systems. Even if one model per channel is used simultaneously, the transition between each model is not captured. Secondly, our model assumes that the user does not know the kind of content before being connected because the parameter $\lambda_i^j$ is not dependent from the kind of content. Thirdly, the content duration is not taken into account. Consequently, we cannot modelize the fact that a user wants to watch a single show, and no more.

A second set of limits are induced by the values of the parameters. The arrival model presents sudden changes around fixed hours because we discretize a day in only five periods. It can be smoothed with a finer discretization but, the more a day is discretized, the more it is difficult to set the parameter $\lambda_i^j$ accurately. Another limit concerns the assumption of a session duration based on a sharing of the global watching duration.

V. Conclusion and future works

Accurately modeling users in live streaming systems opens several perspectives from its integration in simulation tools, resource allocation, or the design of user-centric systems able to provide a negotiated Quality of Experience. However, measurement studies only propose global models which can be far from individual behavior and sociological approaches lack of formalization. In this paper, we propose an individual
model of users based on fictional characters. Validation results show that our model is consistent with the global models in the literature regarding user population and session durations.

This work has to be enlarged through the integration of other kind of elements that impacts the user behavior (such as buffering level or channel popularity) and we are working on it. Concerning the use of this model, we first plan to integrate it into P2PTV-sim [12] to enhance it with individual peer behaviors, but our longer term objective consists in designing a decision system able to statistically identify the behavior of a peer and adapt the streaming system accordingly.

REFERENCES


