A Semi-Markovian Individual Model of Users for P2P Video Streaming Applications

Grégory Bonnet[1], Ihsan Ullah[2], Guillaume Doyen[2], Lionel Fillatre[3], Dominique Gaïti[2] and Igor Nikiforov[3]
[1]University of Caen Lower-Normandy (GREYC) - [23]Troyes University of Technology
[1]UMR 6072 GREYC - [23]UMR STRM 6279 ICD
[1]MAD Team - [2]ERA Team - [3]LM2S Team
FRANCE
email: [1]gregory.bonnet@unicaen.fr - [23]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 controling 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 [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[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 [2]–[7]. Table I presents an overview of the different models extracted from these measurements.

In the matter of population models, [3]–[5], [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 [3] and an off-peak during the week-ends [3], [4]. This weekly pattern is repeated with an off-peak during holiday. Concerning arrival models, exponential laws [3], [7] and Poisson laws [2], [5] are the most used. As for session duration, the models are mostly based on lognormal laws [2], [5], [6] and exponential laws [4], [7]. Concerning the popularity models, all studies [3]–[5], [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 [8]. Despite these works, it is difficult to extract formal behaviors applicable to video streaming system. For that reason, [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 fictionnal characters in terms of identity, skills, habits and goals that are consistent with the sociological literature. Although this work

[1]http://www.sics.se/projects/ondemand_iptv

Ref.	User arrival	Session duration	Popularity
[2]	Poisson law ($\lambda = 0.68$)	Log-normal law	Not measured
[3]	Exponential law	Exponential law	Zipf law ($\alpha = 0.27$)
[4]	Not measured	Two log-normal laws ($\mu = (0.16, 0.2)$, $\sigma = (0.06, 0.27)$)	Zipf law ($\alpha = 0.667$)
[5]	Pseudo-Poisson law ($\lambda = 17$, $N = 27$)	Log-normal law ($\mu = 2.2$, $\sigma = 27$)	Zipf law
[6]	Not measured	Log-normal law ($\mu = 4.835$, $\sigma = 1.704$)	Normal law ($\mu = 33.2$, $\sigma = 17.1$)
[7]	Combination of exponential laws	Exponential law	Zipf law

TABLE I
OVERVIEW OF THE GLOBAL MODELS IN THE LITERATURE

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, $\{t_1 \dots t_{1440}\}$;
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 $\mu_t^i(d)$ and $\sigma_t^i(d)$ 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 λ_t^i 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 :

$$P(X(t) = X_2 | X_1, t_{X_1}) = \int_0^{t_{X_1}} \frac{e^{\frac{-(\log x - \mu_t^i(d))^2}{2.(\sigma_t^i(d))^2}}}{x.\sigma_t^i(d).\sqrt{2\pi}}.dx$$

where $\mu_t^i(d)$ and $\sigma_t^i(d)$ 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,

$$P(X(t) = X_1 | X_2) = e^{-\lambda_t^i} \lambda_t^i$$

where λ_t^i 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 **J**ohnatan, **E**mma, **S**tephan, **A**nna, **P**eter and E**L**len. 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 λ_t^i 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 λ_t^i 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 unregular and S is very unregular. This regularity is given by the variance $\sigma^2 = \frac{\bar{x}}{k}$ where $\bar{x}$ is the total time session per day ($k = 1$ for very irregular to 4 for very regular).

Parameter \ Persona	**Johnatan** (J)	**Emma** (E)	**Stephan** (S)	**Anna** (A)	**Peter** (P)	**Ellen** (L)
Features defined by [1]						
Age	17	25	33	46	58	69
Interests	Sports and serials	No matters	News and sports	Serials	News and reports	Talk-shows and reports
Watching time per day	2 - 3 h.	1.5 h.	1.5 h. (high Δ)	1.5 h.	1.5 h.	2 h.
Watching time habits	Evening, night	Almost at night	Noon, after work	Afternoon, evening	Noon, evening	No habits
Socioprofessional	Student	Store clerk	Executive	Housewife	Full professor	Retired
λ_t^i values on one-minute intervals with respect to persona i and global time t						
4 a.m. to 9:59 a.m.	0.00125	0.00125	0.00125	0.00375	0.00125	0.00375
10 a.m. to 3:59 p.m.	0.00375	0.00375	0.0075	0.005	0.005	0.005
4 p.m. to 6:59 p.m.	0.00375	0.00125	0.00125	0.00625	0.00375	0.00375
7 p.m. to 10:59 p.m.	0.0075	0.005	0.00375	0.00625	0.0075	0.00375
11 p.m. to 3:59 a.m.	0.00375	0.00625	0.00375	0.00125	0.00125	0.00125
μ_i and σ_i values on a full day with respect to persona i						
Watching duration per day	150	90	90	90	90	120
Variance	37.5	30	90	22.5	45	40
μ_i	5.0098	4.498	4.4943	4.4984	4.497	4.7861
σ_i	0.0408	0.0608	0.1051	0.0527	0.0744	0.0527
Sharing M_t^i of the total online time with respect to persona i and global time t						
4 a.m. to 9:59 a.m.	0.05	0.05	0.05	0.05	0.05	0.25
10 a.m. to 3:59 p.m.	0.1	0.05	0.2	0.2	0.3	0.25
4 p.m. to 6:59 p.m.	0.1	0.05	0.1	0.35	0.2	0.15
7 p.m. to 10:59 p.m.	0.4	0.2	0.4	0.2	0.35	0.25
11 p.m. to 3:59 a.m.	0.35	0.65	0.25	0.2	0.1	0.1
Values of the interest coefficients θ_d^i for persona i and content type d						
Fiction	1	1	0.3	2.4	0.3	1
Reality	0.5	1	1.7	0.3	2.4	1.5
Sports	1.5	1	1	0.3	0.3	0.5
Proportion of each persona in the network						
	0.182	0.168	0.177	0.186	0.174	0.113

TABLE II
PARAMETERS OF THE MODEL

The third part of Table II gives the mean and the variance in minutes of the total online presence for each persona i as well as the μ_i and σ_i parameters of their lognormal laws.

As [1] solely give a watching duration per day for each persona, we need to define $\mu_t^i(d)$ and $\sigma_t^i(d)$ from μ_i and σ_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 θ_d^i for each persona i with respect to each content type d. We choose the values of the coefficients such that their mean egals 1 which represents the fact that, in average, the expected session duration follows the previous distribution. Consequently, we defined $\mu_t^i(d)$ to $\sigma_t^i(d)$ as follows:

$$\mu_t^i(d) = \mu_i.M_i^t.\theta_d^i$$
$$\sigma_t^i(d) = \sigma_i.M_i^t.\theta_d^i$$

3) Population model: As an individual model, a non-homogeneous semi-markovian process is instanciated 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 litterature.

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

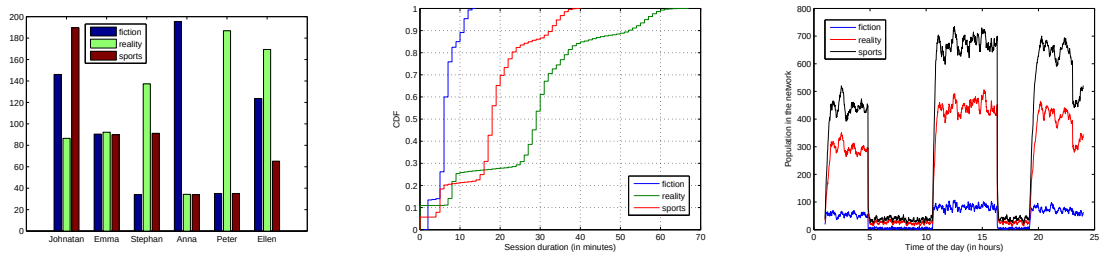

Fig. 1. Behavior with respect to content: (a) total watching duration for each persona ; (b) CDF of Stephan's session durations ; (c) Population of Stephan

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 represents 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 litterature.

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 log-normal 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 λ_t^i 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 λ_t^i 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 negociated 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

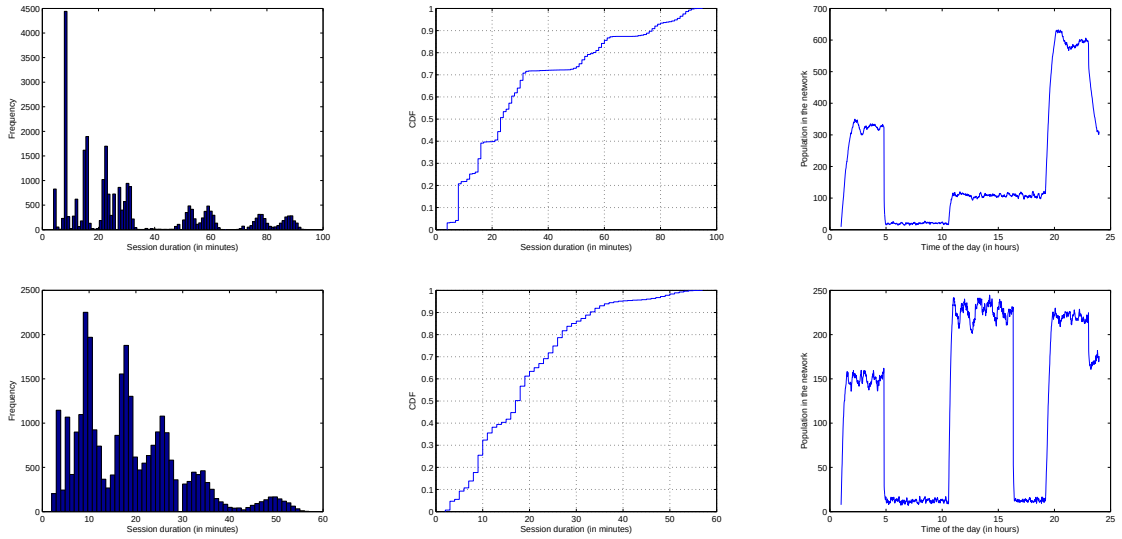

Fig. 2. Global behavior: (a) Johnatan on upperside ; (b) Anna on underside

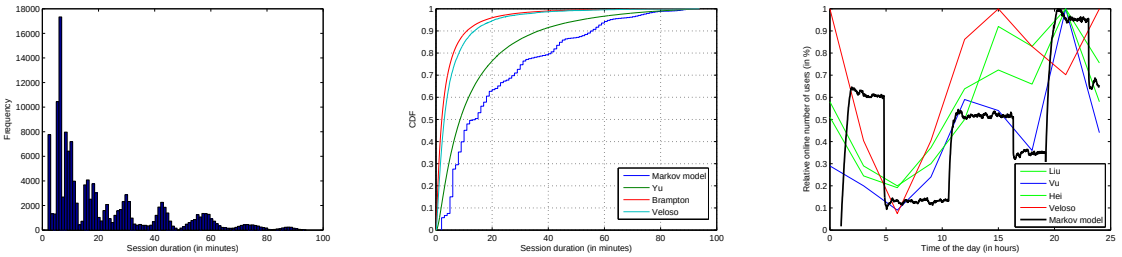

Fig. 3. Global simulations: (a) frequency of session durations ; (b) CDF of session durations; (c) population evolution

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 satistically identify the behavior of a peer and adapt the streaming system accordingly.

REFERENCES

[1] A. Rudstrom and M. Sjolinder, "Capturing TV user behaviour in fictional character descriptions," Swedish Institute of Computer Science, Tech. Rep., October 2008.

[2] P. Branch, G. Egan, and B. Tonkin, "Modeling interactive behaviour of a video based multimedia system," in *Proceedings of the IEEE International Conference on Communications*, 1999, pp. 978–982.

[3] S. Acharya and B. Smith, "Characterizing user access to videos on the world wide web," *Lecture Notes in Computer Science*, vol. Vol. 2720, pp. 375–384, 2004.

[4] M. Vilas, X.-G. Paneda, R. Garcia, D. Melendi, and V.-G. Garcia, "User behavior analysis of a video-on-demand service with a wide variety of subjects and lengths," in *Proceedings of the 31st EUROMICRO Conference on Software Engineering and Advanced Applications*, 2005, pp. 330–337.

[5] H. Yu, D. Zheng, B.-Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in *Proceedings of the 1st European Conference on Computer Systems*, 2006.

[6] A. Brampton, A. MacQuire, I. Rai, J.-P. Nicholas, L. Mathy, and M. Fry, "Characterising user interactivity for sports video-on-demand," in *Proceedings of the International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2007.

[7] B. Chang, L. Dai, Y. Cui, and Y. Xue, "On feasibility of P2P on-demand streaming via empirical VoD user behavior analysis," in *Proceedings of the 28th International Conference on Distributed Computing Systems*, 2008, pp. 7–11.

[8] A.-M. Rubin, "The interactions of television uses and gratifications," in *Proceedings of the Annual Meeting of the Association for Eduction in Journalism*, 1981, pp. 1–25.

[9] Z. Liu, C. Wu, B. Li, and S. Zhao, "Distilling superior peers in large-scale P2P streaming systems," in *Proceedings of INFOCOM*, 2009, pp. 82–90.

[10] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin, "A hierarchical characterization of a live streaming media workload," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment*, 2002, pp. 117 – 130.

[11] L. Vu, I. Gupta, J. Liang, and K. Nahrstedt, "Measurement and modeling of a large-scale overlay for multimedia streaming," in *Proceedings of the International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, 2007, pp. 1–7.

[12] A. Couto da Silva, E. Leonardi, M. Mellia, and M. Meo, "A bandwidth-aware scheduling strategy for P2P-TV systems," in *Proceedings of the 8th International Conference on Peer-to-Peer Computing*, 2008.