



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Impact of the Correlation between Flow Rates and Durations on the Large-Scale Properties of Aggregate Network Traffic

Patrick Loiseau — Paulo Gonçalves — Pascale Vicat-Blanc Primet

N° 7100

November 2009

Domaine 3

 *Rapport
de recherche*

Impact of the Correlation between Flow Rates and Durations on the Large-Scale Properties of Aggregate Network Traffic

Patrick Loiseau*, Paulo Gonçalves* , Pascale Vicat-Blanc Primet*

Domaine : Réseaux, systèmes et services, calcul distribué
Équipe-Projet RESO

Rapport de recherche n° 7100 — November 2009 — 16 pages

Abstract: Since the discovery of long-range dependence in network traffic in 1993, many models have appeared to reproduce this property, based on heavy-tailed distributions of some flow-scale properties of the traffic. However, none of these models consider the correlation existing between flow rates and flow durations. In this work, we extend previously proposed models to include this correlation. Based on a planar Poisson process setting, which describes the flow-scale traffic structure, we analytically compute the auto-covariance function of the aggregate traffic's bandwidth and show that it exhibits long-range dependence with a different Hurst parameter. In uncorrelated case, the model that we propose is consistent with existing models, and predict the same Hurst parameter. We also prove that pseudo long-range dependence with a different index can arise from highly variable flow rates. The pertinence of our model choices is validated on real web traffic traces.

Key-words: Network traffic, Long-range dependence, Heavy-tailed distributions, Correlation

* INRIA/ENS Lyon, Université de Lyon

Impact de la corrélation entre les durées et les débits des flux sur les propriétés statistiques du trafic agrégé dans les réseaux

Résumé : Depuis la découverte, en 1993, de la longue mémoire dans les traces de trafic réseaux, de nombreux modèles, basés sur les distributions à queue lourde de certaines propriétés du trafic à l'échelle flux, ont été proposés pour reproduire et interpréter cette propriété. Cependant, aucun de ces modèles ne tient compte de la corrélation existant entre les débits et les durées des flux. Dans ce travail, nous étendons les modèles existants pour inclure cette corrélation. Le modèle que nous proposons est basé sur un processus de Poisson dans le plan décrivant la structure du trafic à l'échelle flux. Il nous permet de calculer analytiquement la fonction d'auto-covariance du trafic agrégé, et de montrer que celui-ci possède la propriété de longue mémoire, avec un paramètre de Hurst différent. Dans le cas décorrélé, notre modèle est cohérent avec les modèles existants, et prédit le même paramètre de Hurst. Nous montrons également qu'une pseudo longue dépendance peut apparaître avec un autre paramètre de Hurst dans le cas de débits de flux très variables. Nos hypothèses de modélisation sont enfin favorablement confrontées à des traces de trafic web réel.

Mots-clés : Trafic réseau, Dépendance à longue portée, Distributions à queue lourde, Corrélation

Contents

1	Motivation	3
2	Related work	4
3	A model accounting for the correlation between flow rates and durations	6
3.1	Definitions and notations	6
3.2	Computation of the instantaneous bandwidth correlation	8
4	Results and discussion	13
4.1	Trace description	13
4.2	Confrontation of the model with the trace	14
5	Conclusion	15

1 Motivation

Deep understanding of network-traffic properties is essential for Internet Service Providers to optimally control the traffic in order to offer users the best Quality of Service. In this context, a lot of research has been recently focusing on mathematical modeling of network traffic, especially from a statistical viewpoint. However, comprehensive modeling of all the characteristics of the traffic is a very arduous problem for it encompasses several difficulties of different natures, such as transport protocols, control mechanisms, and complexity due to the users behavior. The design of simple models, yet rich enough to account for the most important characteristics observed in the traffic, is thus a major issue for many industrial applications such as network dimensioning, optimization, control, and prediction.

A major breakthrough in this direction has been the discovery in 1993 of the self-similar nature of aggregate time series at large time scales [?, ?]. Following up, several models was proposed that posit the heavy-tailed nature of the activity periods, or ON periods, as a plausible explanation for the origin of the self-similar property; and clarify the relation between the tail exponent α_{ON} and the Hurst parameter H :

$$H = \frac{3 - \alpha_{\text{H}}}{2}, \quad (1a)$$

$$\text{where } \alpha_{\text{H}} = \min(\alpha_{\text{ON}}, 2), \text{ or } \min(\alpha_{\text{ON}}, \alpha_{\text{OFF}}, 2), \quad (1b)$$

depending on whether the model include some heavy-tailed OFF periods of tail index α_{OFF} or not. Since then, the self-similarity property has been shown to have a major impact on QoS, and such models, based on a flow-level characterization of the sources, are now commonly used to generate realistic traffic traces.

However, all of the models proposed up to now that lead to relation (1) entail a common simplified feature: the tail indices of the flow-size distribution α_{SI} , and of the flow-duration distribution α_{ON} are identical, so that both can interchangeably be used in relation (1). Even in models randomly drawing

the flow rates, this equality holds due to the assumption that flow rates are independent of flow sizes (or flow durations). Such an assumption might not hold in real internet traffic, and the equality between α_{ON} and α_{SI} would then fall down. This has been observed since 1997 [?]. Moreover, the correlation between flow rates and durations is likely to become more prominent in the future Internet, due to the emergence of new mechanisms such as the FTTH increasing the rates spectrum, and flow-aware control procedures which might strongly correlate the achieved flow rates to flow properties such as its duration or size. In this case, relation (1) is not ensured to be reliably usable to predict the Hurst parameter of the aggregate traffic. In this context, the development of a model including the correlation between flow rates and durations is a major challenge to accurately predict the Hurst parameter of the aggregate traffic. To the best of our knowledge, no such model has been proposed yet.

In this paper, we propose an extension of existing models, which takes into account the correlation between flow rates and durations (responsible for the different tail indices α_{ON} and α_{SI}). Based on a representation of the flows as a planar Poisson process, we analytically calculate the auto-covariance function of the aggregate traffic's instantaneous bandwidth and deduce the resulting Hurst parameter. We first briefly review existing models that reproduce the long-range dependence property in Section 2. Then we extend in Section 3 these models to include the correlation between flow rates and durations and show how to predict the resulting long-range dependence of the aggregate traffic. In Section 4, we experimentally demonstrate the ability of our model to accurately predict the Hurst parameter, based on a real traffic trace.

2 Related work

There exists a large number of models able at reproducing the long-range dependence property [?, ?]. We describe here only those which are explicitly relying on the notion of *flows*, particularly meaningful in network-traffic applications. Following up Mandelbrot's idea, these models rely on the introduction of a heavy-tailed distribution of infinite variance. We distinguish two categories of such models, which mainly differ in the flow arrival process: the *infinite source Poisson models* and the *renewal models*.

Infinite source Poisson models. In the simplest version of the infinite source Poisson models, flows arrive at the link as a Poisson process of rate λ , and transmit data at a fixed rate of 1 during a heavy-tailed random time of tail index α_{ON} . It is also known as the M/G/ ∞ model originally considered in [?]. Then, two different limiting regimes have to be considered when studying the cumulative bandwidth at large time scales T : if λ goes to infinity first, then T , and a proper rescaling is applied, the resulting process is a fractional Brownian motion of Hurst parameter as in equation (1) (see *e.g.* surveys in [?, ?]). In the opposite limiting regime where the time scale T goes first to infinity, the resulting process is a Lévy motion. In the case where both λ and T goes to infinity in the same time, [?] provides conditions on the ratio between the growth rate of both quantities that ensure one, or the other limiting regime (see also [?] where an intermediate case between these two conditions is studied). There exist many variants of this simple M/G/ ∞ model, which all rely on the

same mechanism of a Poissonian arrival of some heavy-tailed flows and mainly differ on the way data is transmitted within a flow (see also a survey of infinite Poisson models in [?], and the references in [?]). In [?], the authors consider an infinite source Poisson model with a general form of the “workload function” inside a flow and establish (for the first time) the Gaussian limit result. A similar model is considered in [?] (where the “workload function” is called “transmission schedule”) and the other limiting result is shown (the Lévy motion). In [?], the Poisson shot-noise model is developed. This model is basically very similar to the two models previously mentioned. The “workload function” or “transmission schedule” is now called “shot”, but “shots” still arrive according to a Poisson process. An application of this model is proposed in [?] where the shot shape is representative of the AIMD mechanism with Poisson losses. In [?], the authors consider a model where the rate within a flow is constant, but the rate of a flow is a random variable. The flow sizes are drawn at random according to a heavy-tailed distribution of tail index α_{SI} ; and the rates are drawn at random independently of the sizes, according to any finite-mean distribution. Flow duration (flow size divided by flow rate) is then heavy-tailed with $\alpha_{ON} = \alpha_{SI}$ and they find a long-range correlation structure of Hurst parameter H as in relation (1). A similar model where the durations and rates are drawn independently is proposed in [?]. The last model we mention in this infinite Poisson model category is the Cluster Point Processes (CPP) model proposed in [?]. In this model, the discrete approach of point processes is used. Flows are “clusters” of points and the flow arrival time is the time of the first packet of the cluster. The flow-arrival process is Poisson. The number of points in the cluster (the flow size) is heavy-tailed of tail index α_{SI} , and points (packets) within a cluster (flows) follow a renewal process with some inter-renewal distribution determining the mean flow rate. Based on results on point processes (see [?, ?]), the authors calculate the aggregate traffic’s spectrum and again find long-range dependence of Hurst parameter H as in relation (1a) where $\alpha_H = \alpha_{SI}$. As in the model of [?], the flow duration in this model is heavy-tailed of tail index $\alpha_{ON} = \alpha_{SI}$. As already observed in [?], this might not hold true in real Internet traffic. For example, slow-start effects might imply higher rates for larger flows. In this case, flow rate is correlated to flow size and the tail indices of the flow-duration and the flow-size distributions are different. In [?], the authors suggest as future work to introduce multi-class CPP, for example with a class for small flows (mice) having small rates, and another class for large flows (elephants) having high rates. Such a work, however, has not been done and the following question, that we address in this work, remained open: What happens if flow-duration and flow-size distributions have different tail indices? Which of these tail indices, if any, governs the long-range dependence of the aggregate traffic?

Renewal models. Another, closely related, class of models generating long-range dependence is the class of renewal models. They are based on the same general setting firstly introduced by Mandelbrot in 1969 [?] in an economical context. Each source, emitting only one flow at a time, is modeled as a renewal reward process, where the inter-renewal time (*i.e.* the interval between two consecutive flows) is heavy-tailed; and the reward is the rate of the flow. Many variants of such models have been considered, mainly differing in the distribution of the reward [?, ?, ?] (see also [?]). We present here only briefly

the particular variant where the reward strictly alternate between 0 and 1: the ON/OFF model [?]. This model allows including the notion of idle time between the transmission of two flows through the OFF periods. Moreover, ON- and OFF-periods distributions can have different tail indices α_{ON} and α_{OFF} . Then, it is shown in [?] that the same two limiting regimes as for the infinite source Poisson models yield the same limit processes. In particular, in the Gaussian limit, a fractional Brownian motion is found, whose Hurst parameter satisfies equation (1). The difference of this model, as compared to infinite source Poisson models is that it can account for long-range dependence induced by heavy-tailed idle times. The flow arrival process is also no longer, in general a Poisson process. However, when $\alpha_{\text{ON}} > \alpha_{\text{OFF}}$, which is a common case in network traffic (see *e.g.*[?]), the results in terms of predicted Hurst parameter are the same for both models.

Planar point process setting. In our model, we choose to use the infinite source Poisson model, and we represent the flows as a planar Poisson process. Such a setting has been introduced in the context of multifractal analysis in [?] to extend binomial multiplicative cascades. It is also used in [?] also in the context of multiplicative processes. To the best of our knowledge, this setting has been used in the context of additive processes to study long-range dependence only in the previously mentioned papers [?, ?].

3 A model accounting for the correlation between flow rates and durations

3.1 Definitions and notations

Figure 1 depicts our setting. The point process (T_i, D_i) (representing the arrival time and the duration of the flows) is assumed to be a planar Poisson process of measure Λ [?, ?]. This means in particular that the sequence $(D_i)_{i \geq 0}$ is independent. The measure Λ determines the mean number of points in a part of the plane. For example, $\Lambda(C(t_1))$ is the mean number of points in the cone $C(t_1)$ (see Figure 1). In the case where the measure has a constant density (or intensity), the point process is called homogeneous. Here, the measure contains the information of the flow duration distribution. To account for a heavy-tailed flow-size distribution, we use the following form for the measure Λ of an elementary square of size (dt, dd) centered on (t, d) :

$$\Lambda(dt, dd) = \frac{C dt dd}{d^{\alpha_{\text{ON}}+1}}, \text{ where } C > 0, \text{ and } \alpha_{\text{ON}} > 1. \quad (2)$$

Due to the dependence of Λ in the variable d , the point process is non-homogeneous (the density of points is higher for smaller values of d). However, due to the independence of Λ in the variable t , the arrival time of a flow T_i is independent of its duration D_i . In our case where the x -axis represents the time, this specific form of the measure Λ ensures the stationarity of the resulting traffic. To avoid integrability problems for d around zero, we also set a minimal flow duration ε , under which the distribution is 0. This threshold will not play any role in our study because we concentrate on long-range dependence. A similar setting is used for example in [?], where the behavior of the measure around $d = 0$

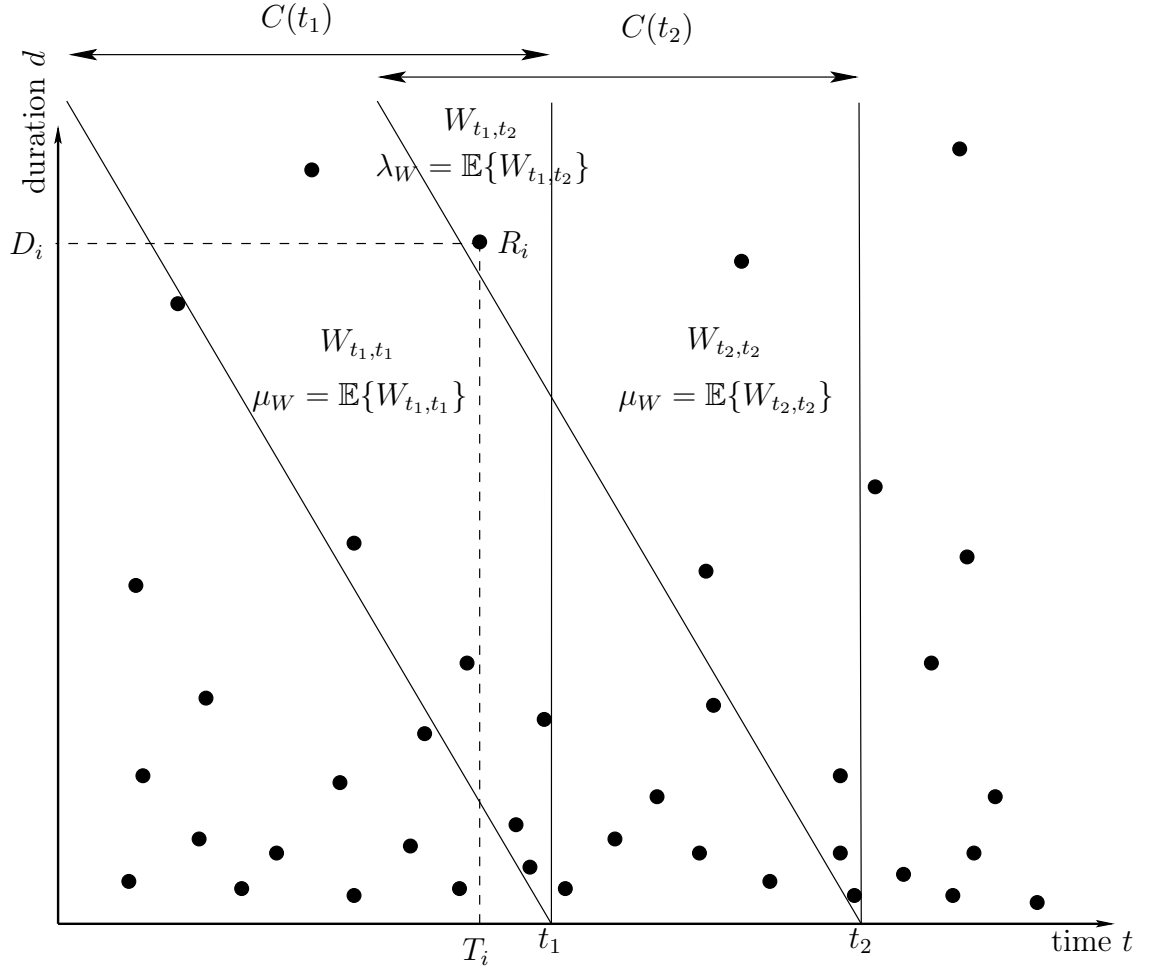


Figure 1: Setting of the model. Each point represents a flow. The x -coordinate represents the start time T_i and the y -coordinate the duration D_i . R_i is the rate of the flow. At time t_1 , the active flows are those lying in cone $C(t_1)$ (the left border of the cones has a slope of -1).

has a great role because the authors focus on small-scale multifractality (note however, that in this paper the authors study a multiplicative process which differs from our additive process considered here).

Each flow emits data at a constant rate R_i during time D_i . The rates are random variables such that sequence $(R_i)_{i \geq 0}$ is independent. However, contrarily to previously proposed models (see in particular [?, ?]), we do not assume that R_i is independent from D_i . Instead, we permit some correlation between the flow rates and durations that we will precise later in this section, when it becomes necessary to pursue the calculations. The flow size, that we denote S_i is then determined by $S_i = D_i R_i$.

We consider two time instants $t_1 < t_2$ and introduce the following notations (see Figure 1):

$$W_{t_1, t_1} = \sum_{(T_i, D_i) \in C(t_1) \setminus C(t_2)} R_i, \quad (3)$$

$$W_{t_1, t_2} = \sum_{(T_i, D_i) \in C(t_1) \cap C(t_2)} R_i, \quad (4)$$

$$W_{t_1} = W_{t_1, t_1} + W_{t_1, t_2} = \sum_{(T_i, D_i) \in C(t_1)} R_i. \quad (5)$$

The variables W_{t_2, t_2} and W_{t_2} are defined similarly. The random variable W_{t_1} is directly the instantaneous throughput at time t_1 (the sum of the rates of the flows active at time t_1). The random variables W_{t_1, t_1} and W_{t_1, t_2} are intermediate random variables, useful for the calculations developed in the following. We can interpret these random variables as follows: W_{t_1, t_1} corresponds to the traffic of the flows that are active at time t_1 but not any more at time t_2 , whereas W_{t_1, t_2} corresponds to the traffic of the flows active at time t_1 and still active at time t_2 . For clarity purposes, we introduce the following notations for the mean:

$$\lambda_W = \mathbb{E}\{W_{t_1, t_2}\} = \Lambda(C(t_1) \cap C(t_2)), \quad (6)$$

$$\mu_W = \mathbb{E}\{W_{t_1, t_1}\} = \mathbb{E}\{W_{t_2, t_2}\} = \Lambda(C(t_1) \setminus C(t_2)) = \Lambda(C(t_2) \setminus C(t_1)), \quad (7)$$

where simple calculations show that the last equality $\Lambda(C(t_1) \setminus C(t_2)) = \Lambda(C(t_2) \setminus C(t_1))$ holds due to the uniform distribution of the measure Λ in time. Finally, note that with these notations, we have:

$$\mathbb{E}\{W_{t_1}\} = \mathbb{E}\{W_{t_2}\} = \lambda + \mu. \quad (8)$$

The setting described here allows us to compute the autocovariance of the instantaneous bandwidth W_t , in order to evaluate the Hurst parameter, as we shall see now (recall that a stationary process is called long-range dependent of Hurst parameter H if its auto-covariance function decrease like τ^{2-2H} when τ goes to infinity, where $1/2 < H < 1$).

3.2 Computation of the instantaneous bandwidth correlation

We are interested in the computation of the autocovariance function of the instantaneous bandwidth, that is $\mathbb{E}\{W_{t_1} W_{t_2}\} - \mathbb{E}\{W_{t_1}\} \mathbb{E}\{W_{t_2}\}$.

Proposition 3.1. *If the planar point process $\{(T_i, D_i)\}$ is Poisson, then*

$$\mathbb{E}\{W_{t_1} W_{t_2}\} - \mathbb{E}\{W_{t_1}\} \mathbb{E}\{W_{t_2}\} = \mathbb{E}\{W_{t_1, t_2}^2\} - \mathbb{E}\{W_{t_1, t_2}\}^2 = \text{Var}\{W_{t_1, t_2}\} \quad (9)$$

Proof. Using the notations introduced above, we have:

$$\mathbb{E}\{W_{t_1} W_{t_2}\} = \mathbb{E}\{W_{t_1, t_1} W_{t_2, t_2}\} + \mathbb{E}\{W_{t_1, t_1} W_{t_1, t_2}\} + \mathbb{E}\{W_{t_1, t_2} W_{t_2, t_2}\} + \mathbb{E}\{W_{t_1, t_2}^2\}.$$

Due to the Poisson assumption, the random variables W_{t_1, t_1} , W_{t_2, t_2} , W_{t_1, t_2} are mutually independent, so that

$$\begin{aligned} \mathbb{E}\{W_{t_1} W_{t_2}\} &= \mathbb{E}\{W_{t_1, t_1}\} \mathbb{E}\{W_{t_2, t_2}\} + \mathbb{E}\{W_{t_1, t_1}\} \mathbb{E}\{W_{t_1, t_2}\} + \mathbb{E}\{W_{t_1, t_2}\} \mathbb{E}\{W_{t_2, t_2}\} + \mathbb{E}\{W_{t_1, t_2}^2\} \\ &= \mu_W^2 + 2\lambda_W \mu_W + \mathbb{E}\{W_{t_1, t_2}^2\} \\ &= (\lambda_W + \mu_W)^2 - \lambda_W^2 + \mathbb{E}\{W_{t_1, t_2}^2\} \end{aligned}$$

directly giving equation (9) in view of the definition of λ_W (equation (6)) and of equation (8). \square

Proposition 3.1 shows that the autocovariance depends only on the variance of the traffic due to the flows in the intersection of the cones $C(t_1)$ and $C(t_2)$. This had been noticed already in [?, ?], though the authors were focusing on small-scale properties. To complete the computation of the autocovariance function, we then need to compute $\text{Var}\{W_{t_1, t_2}\}$.

If the flow rates were constant and equal to 1, then W_{t_1, t_2} would simply be the number of points in $C(t_1) \cap C(t_2)$. Since the point process is Poisson, the variance $\text{Var}\{W_{t_1, t_2}\}$ would then simply be λ_W and the autocovariance would be determined by the value of λ_W (recall that the variance of the count process associated with a Poisson process is equal to its mean). Before proceeding with the calculations in a more general case, we introduce additional useful notations. We denote by N the random variable corresponding to the number of points in $C(t_1) \cap C(t_2)$. We denote its mean by

$$\lambda_N = \mathbb{E}\{N\}. \quad (10)$$

Note that in the case where the rate is always 1, we simply have $\lambda_W = \lambda_N$. The next proposition gives the general form of the autocovariance function, without specifying the correlation between R_i and D_i yet.

Proposition 3.2.

$$\mathbb{E}\{W_{t_1} W_{t_2}\} - \mathbb{E}\{W_{t_1}\}\mathbb{E}\{W_{t_2}\} = \lambda_N \mathbb{E}\{R_i^2\}, \quad (11)$$

where it is implicitly understood that the expectation is computed in $C(t_1) \cap C(t_2)$.

Proof. To evaluate the value of $\text{Var}\{W_{t_1, t_2}\}$, we successively compute the values of $\mathbb{E}\{W_{t_1, t_2}\}$ and $\mathbb{E}\{W_{t_1, t_2}^2\}$.

$$\begin{aligned} \mathbb{E}\{W_{t_1, t_2}\} &= \mathbb{E}\{\mathbb{E}\{W_{t_1, t_2} | N\}\} \\ &= \sum_{k=1}^{\infty} \mathbb{E}\left\{\sum_{i=1}^k R_i | N = k\right\} \mathbb{P}(N = k) \\ &= \sum_{k=1}^{\infty} \sum_{i=1}^k \mathbb{E}\{R_i | N = k\} \mathbb{P}(N = k) \\ &= \sum_{k=1}^{\infty} k \mathbb{E}\{R_i\} \mathbb{P}(N = k) \\ &= \lambda_N \mathbb{E}\{R_i\}. \end{aligned}$$

$$\begin{aligned} \mathbb{E}\{W_{t_1, t_2}^2\} &= \mathbb{E}\{\mathbb{E}\{W_{t_1, t_2}^2 | N\}\} \\ &= \sum_{k=1}^{\infty} \mathbb{E}\left\{\left(\sum_{i=1}^k R_i\right)^2 | N = k\right\} \mathbb{P}(N = k) \end{aligned}$$

By independence of the sequence $(R_i)_i$

$$\mathbb{E}\left\{\left(\sum_{i=1}^k R_i\right)^2 | N = k\right\} = \mathbb{E}\left\{\left(\sum_{i=1}^k R_i\right)^2\right\} = k \mathbb{E}\{R_i^2\} + (k^2 - k) \mathbb{E}\{R_i\}^2,$$

so that

$$\begin{aligned}
\mathbb{E}\{W_{t_1, t_2}^2\} &= \sum_{k=1}^{\infty} (k\mathbb{E}\{R_i^2\} + (k^2 - k)\mathbb{E}\{R_i\}^2)\mathbb{P}(N = k) \\
&= \sum_{k=1}^{\infty} k\mathbb{E}\{R_i^2\}\mathbb{P}(N = k) + \sum_{k=1}^{\infty} k^2\mathbb{E}\{R_i\}^2\mathbb{P}(N = k) - \sum_{k=1}^{\infty} k\mathbb{E}\{R_i\}^2\mathbb{P}(N = k) \\
&= \lambda_N\mathbb{E}\{R_i^2\} + (\lambda_N + \lambda_N^2)\mathbb{E}\{R_i\}^2 - \lambda_N\mathbb{E}\{R_i\}^2 \\
&= \lambda_N\mathbb{E}\{R_i^2\} + \lambda_N^2\mathbb{E}\{R_i\}^2
\end{aligned}$$

□

Until that point, we still have not used the precise form of the measure Λ (equation (2)), and the precise form of the correlation between R_i and D_i . The result of Proposition 3.2 depends only on the Poisson and independence of the sequence $(R_i)_i$ assumptions. It shows that the autocovariance function is the product of two terms: λ_N , the mean number of points in $C(t_1) \cap C(t_2)$, and $\mathbb{E}\{R_i^2\}$. The first term (λ_N) depends only on the measure Λ and is easily obtained via a simple integration (Proposition 3.3). To compute the second term ($\mathbb{E}\{R_i^2\}$), we need in addition to precise the correlation between R_i and D_i (Proposition 3.4).

Proposition 3.3. *If the measure Λ has the form of equation (2), then*

$$\lambda_N = C \frac{1}{\alpha_{ON}(\alpha_{ON} - 1)} (t_2 - t_1)^{-\alpha_{ON}+1} \quad (12)$$

Proof.

$$\begin{aligned}
\lambda_N &= \int_{d=t_2-t_1}^{\infty} \int_{t=t_1-(d-(t_2-t_1))}^{t_1} \Lambda(du, dv) \\
&= C \int_{d=t_2-t_1}^{\infty} (d - (t_2 - t_1)) \frac{1}{d^{\alpha_{ON}+1}} dd \\
&= C \int_{d=t_2-t_1}^{\infty} \frac{1}{d^{\alpha_{ON}}} dd - C(t_2 - t_1) \int_{d=t_2-t_1}^{\infty} \frac{1}{d^{\alpha_{ON}+1}} dd \\
&= C \frac{1}{\alpha_{ON} - 1} (t_2 - t_1)^{-\alpha_{ON}+1} - C \frac{1}{\alpha_{ON}} (t_2 - t_1)^{-\alpha_{ON}+1} \\
&= C \frac{1}{\alpha_{ON}(\alpha_{ON} - 1)} (t_2 - t_1)^{-\alpha_{ON}+1}
\end{aligned}$$

□

Finally, we now specify the correlation between R_i and D_i to calculate the term $\mathbb{E}\{R_i^2\}$. Our goal here is to specify a correlation which indeed leads to the different tail indices of the flow-size and flow-duration distributions. We have already mentioned that taking R_i as a random variable independent of D_i would lead to the same tail indices, independently of the distribution of R_i , provided that its mean is finite (see for instance [?, ?]). The different tail indices α_{SI} and α_{ON} can then only come from correlation between R_i and D_i . The simplest choice would be to deterministically take for each flow: $R_i = D_i^{\beta-1}$,

where $\beta = \frac{\alpha_{\text{ON}}}{\alpha_{\text{SI}}}$. In this case, we would have $S_i = D_i^\beta$ or equivalently $D_i = S_i^{1/\beta}$. This would effectively lead to the different tail indices α_{ON} and α_{SI} for the flow duration and size distributions. However, this assumption of a deterministic rate for a flow of a given duration is not realistic. Instead we choose a model where the conditional expectation and variance of the rate given the duration are specified. This model is given in the next proposition.

Proposition 3.4. *Assume that $\mathbb{E}\{R_i|D_i\} = KD_i^{\beta-1}$, where K is a constant and $\text{Var}\{R_i|D_i\} = V$, where V is a constant. We denote*

$$\alpha' = \alpha_{\text{ON}} - 2(\beta - 1). \quad (13)$$

If $\alpha' > 1$, then

$$\mathbb{E}\{R_i^2\} = \frac{1}{\lambda_N} CK^2 \frac{1}{\alpha'(\alpha' - 1)} (t_2 - t_1)^{-\alpha'+1} + V. \quad (14)$$

Proof.

$$\mathbb{E}\{R_i^2\} = \mathbb{E}\{\mathbb{E}\{R_i^2|D_i\}\} = K^2 \mathbb{E}\{D_i^{2(\beta-1)}\} + \mathbb{E}\{\text{Var}\{R_i|D_i\}\}$$

The same kind of integration as for the proof of Proposition 12 give

$$\mathbb{E}\{D_i^{2(\beta-1)}\} = \frac{1}{\lambda_N} C \frac{1}{\alpha'(\alpha' - 1)} (t_2 - t_1)^{-\alpha'+1},$$

while we clearly have $\mathbb{E}\{\text{Var}\{R_i|D_i\}\} = V$, completing the proof. \square

We are now able to state the final proposition establishing the decrease of the autocovariance function with $t_2 - t_1$, and then the long-range dependence of the process $(W_t)_t$.

Proposition 3.5 (Autocorrelation function and long-range dependence of the process $(W_t)_t$). *If $\mathbb{E}\{R_i|D_i\} = KD_i^{\beta-1}$ and $\text{Var}\{R_i|D_i\} = V$, where K, V are constants and $\alpha' = \alpha_{\text{ON}} - 2(\beta - 1) > 1$, then*

$$\begin{aligned} \mathbb{E}\{W_{t_1}W_{t_2}\} - \mathbb{E}\{W_{t_1}\}\mathbb{E}\{W_{t_2}\} &= CK^2 \frac{1}{\alpha'(\alpha'-1)} (t_2 - t_1)^{-\alpha'+1} \\ &+ CV \frac{1}{\alpha_{\text{ON}}(\alpha_{\text{ON}}-1)} (t_2 - t_1)^{-\alpha_{\text{ON}}+1}. \end{aligned} \quad (15)$$

The process $(W_t)_t$ is then (asymptotically) long-range dependent with Hurst parameter H following equation (1a) ($H = \frac{3-\alpha_{\text{H}}}{2}$), where

$$\alpha_{\text{H}} = \min(\alpha', \alpha_{\text{ON}}). \quad (16)$$

Proof. The first part is immediate from Propositions 3.2, 3.3 and 3.4. For the Hurst parameter, recall that a process is long-range dependent of Hurst parameter H if its autocovariance function algebraically decreases like $(t_2 - t_1)^{2H-2}$. \square

Proposition 3.5 is our main result for this section. It establishes the long-range dependence of the instantaneous bandwidth and gives the Hurst parameter. Before going back to the in2p3 trace and verifying the consistence of our model choice with the data, we make a few general remarks on the result of Proposition 3.5, and the choice of the model.

Let us first comment on specific values of $\beta = \frac{\alpha_{\text{ON}}}{\alpha_{\text{SI}}}$.

If $\beta = 1$ ($\alpha_{\text{SI}} = \alpha_{\text{ON}}$): This case corresponds to the classical case of Taqqu's model and other infinite source Poisson models, where the rate is not correlated to the duration and we have the same tail indices for the flow-size and flow-duration distributions. In this case, the result of Proposition 3.5 reduces to Taqqu's relation (1a), where the tail index controlling the long-range dependence α_{H} is the unique tail index $\alpha_{\text{SI}} = \alpha_{\text{ON}}$ (bear in mind that there are no heavy-tailed OFF times here because we used a Poisson flow arrival).

If $\beta > 1$ ($\alpha_{\text{SI}} < \alpha_{\text{ON}}$): This is the case of the in2p3 trace, where the rate increases in average with the duration of the flows. In this case the tail index controlling the long-range dependence is $\alpha_{\text{H}} = \alpha' < \alpha_{\text{ON}}$. The long-term correlations are then stronger than in the case of constant rates. Note that, depending on the value of β , we can have either $\alpha_{\text{H}} < \alpha_{\text{SI}}$, $\alpha_{\text{H}} > \alpha_{\text{SI}}$ or $\alpha_{\text{H}} = \alpha_{\text{SI}}$. It means that the tail index controlling the long-range dependence does not necessarily lie between α_{SI} and α_{ON} , but can be smaller than both these tail indices.

If $\beta < 1$ ($\alpha_{\text{ON}} < \alpha_{\text{SI}}$): This case corresponds to a situation where the rate decreases in average with the duration of the flows. This could happen for instance with some scheduling policy giving priority to small flows. In this case the tail index controlling the long-range dependence is $\alpha_{\text{H}} = \alpha_{\text{ON}} < \alpha'$.

In the case where $V = 0$ (this corresponds to the deterministic case $R_i = D_i^{\beta-1}$), the tail index controlling the long-range dependence is always $\alpha_{\text{H}} = \alpha'$, even if $\beta < 1$. Note that depending on the values of β , we can have in this case surprising situations similar to the situations discussed for $\beta > 1$ where the tail index controlling the long-range dependence α_{H} is greater than the two tail indices α_{SI} and α_{ON} . Note that the calculation of the autocovariance function of equation (15) could easily be performed with some other (non-constant) forms of the conditional variance $\text{Var}\{R_i|D_i\}$. For example, an algebraically increasing variance ($\text{Var}\{R_i|D_i\} = D_i^\gamma$) would lead to another index $\alpha'' = \alpha_{\text{ON}} - \gamma$, possibly controlling the long-range dependence for some set of parameters.

The long-range dependence that we have discussed until now is asymptotic long-range dependence. However, depending on the values of the constants K and V , we can observe situations where within some intermediate range of scale, the term of the autocovariance function (15) with the larger exponent dominates. For example, in the case $\beta > 1$, the first term of the autocovariance asymptotically dominates, but if the value of V is very large and the value of K is not too large, the second term can dominate within some scale range, thus leading to *pseudo long-range dependence* controlled by the tail index α_{ON} . Note that the two constants K and V do not have the same units and cannot be compared directly to one another. Instead, the whole terms of equation (15) have to be compared.

Let us finally mention that to "specify enough" the correlation between the flow rates and durations, in order to compute the autocovariance function of equation (15), we had to specify only the first two conditional moments $\mathbb{E}\{R_i|D_i\}$ and $\text{Var}\{R_i|D_i\}$. This is not surprising since the autocovariance function is a second-order moment quantity, and we found only second-order self-similarity (*i.e.*, long-range dependence). To investigate finer statistical properties of the instantaneous bandwidth, we might need to specify the entire conditional probability $\mathbb{P}(R_i = r|D_i = d)$.

4 Results and discussion

4.1 Trace description

In this section, we use a real traffic trace to validate the ability of our model to predict the aggregate-traffic Hurst parameter. The trace is acquired at the output link of the in2p3 research center (Lyon, France), with the capture tool *MetroFlux* [?]. The traffic is captured from the VLAN corresponding to RENATER¹ web traffic. This VLAN is encapsulated in the 10 Gbps output link of in2p3. Although we captured more than one day of traffic, we restrict our trace to a 30 minutes stationary trace, corresponding to the incoming traffic between 3pm and 3:30pm on January 18, 2009. The mean throughput in this period is 127.3 Mbps. This traffic mainly corresponds to web traffic, but has the particularity of including a larger number of elephants than usual traffic. Since the in2p3 is a nuclear-physics research center, these large transfers are likely to correspond to the transfer of experiment results from experiment centers like CERN.

Figure 2 shows the flow-size and flow-duration distributions. While both distributions clearly appear to be heavy-tailed, they exhibit sharply different tail indices: $\alpha_{SI} = 0.8544$ and $\alpha_{ON} = 1.1994$ respectively (the tail indices are estimated with the recent wavelet method proposed in [?]). Several explanations could be posited to interpret why longer flows achieve higher rates, thus explaining these different tail indices. A plausible explanation might lie in transient effects at the beginning of each flow. Another explanation could lie in the variable locations of the downloaded files, if for example users tend to download large flows from closer locations (thus achieving higher rates more quickly because of smaller RTTs). The largest files might also be intentionally stored on servers with the highest capacities. These are only hypotheses and we do not elaborate further here on the origin of the different tail indices for the flow-size and flow-duration distributions, which is out of our scope.

To investigate the long-range dependence of the aggregate traffic, we use the wavelet method described in [?]. This method provides a robust estimation of the Hurst parameter, with an estimation of the confidence interval in the Gaussian case. Here, we observed a fairly Gaussian traffic, with a kurtosis of 3.15 at scale 10 ms. Figure 3 displays the log diagram of the aggregate traffic bandwidth in time windows of size $\Delta = 10$ ms. We clearly observe a linear behavior in the coarse scales, characteristic of a scaling behavior, with an estimated Hurst parameter of $\hat{H} = 0.901 \pm 0.045$. The estimation is performed in the scale range $[0.64 - 40.96]$ s (lying far beyond the mean ON time of 0.12 s) but the same scaling behavior appears to extend to the scale 100 s. This Hurst parameter shows a good agreement with relation (1) of Taqqu's ON/OFF model, where $\alpha_H = \alpha_{ON}$ would be the tail index governing the long-range dependence. However, as already mentioned, the trace studied here is not well modeled by the ON/OFF model with constant rates, and we now show that the model developed in previous section match well the real trace and allows us to correctly predict this Hurst parameter. Note finally that we observed on the trace a coarse-scale parameter of 0.65 for the flow arrival process. Such a value is unlikely to explain the long-range dependence of the aggregate traffic. Although not strictly satisfied, the Poisson flow arrival assumption made in our model should

¹National research and education network

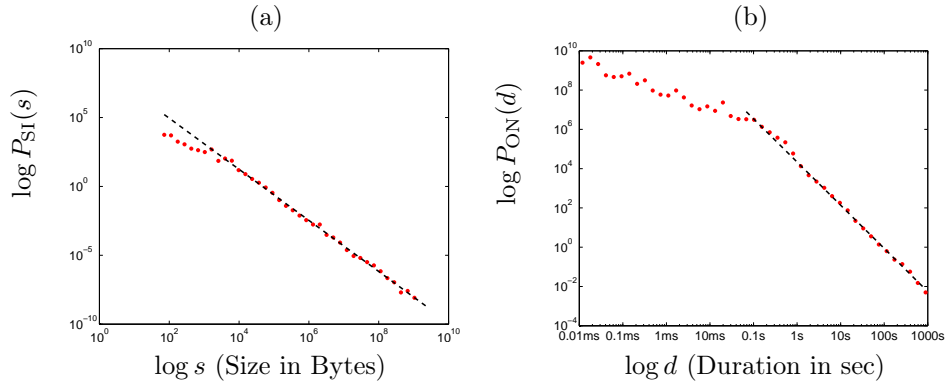


Figure 2: (a) Flow-size distribution – (b) Flow-duration distribution. The tail indices estimated with the wavelet method are respectively: $\alpha_{SI} = 0.8544$ and $\alpha_{ON} = 1.1994$. Straight lines materialize these slopes and have been vertically adjusted to the data. The mean ON time is 0.12 s.

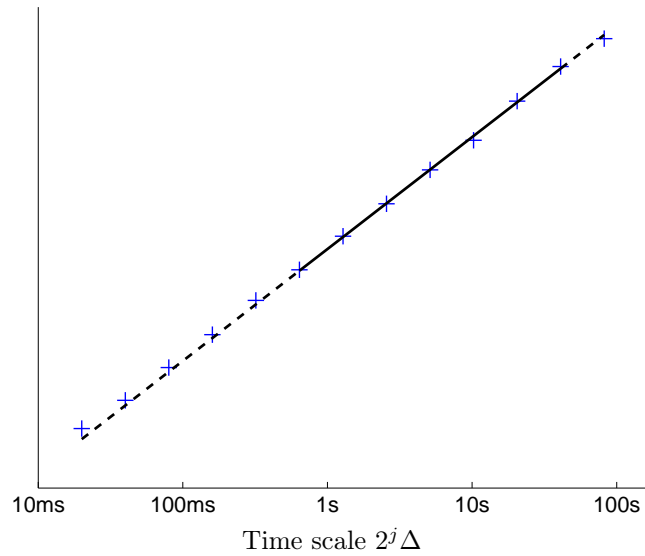


Figure 3: Log diagram of the aggregate traffic bandwidth in time windows of size $\Delta = 10$ ms. The estimated value of the Hurst parameter is $\hat{H} = 0.901 \pm 0.045$ (scale range of estimation: $[0.64 - 40.96]$ s)

not hamper its use to predict the Hurst parameter, most likely related to the heavy-tailed flow size/durations.

4.2 Confrontation of the model with the trace

Our goal here is to confront the model we proposed in previous section to the real data, especially regarding the assumptions we made on the correlation between flow rates and durations in Proposition 3.5. From the in2p3 trace, we have

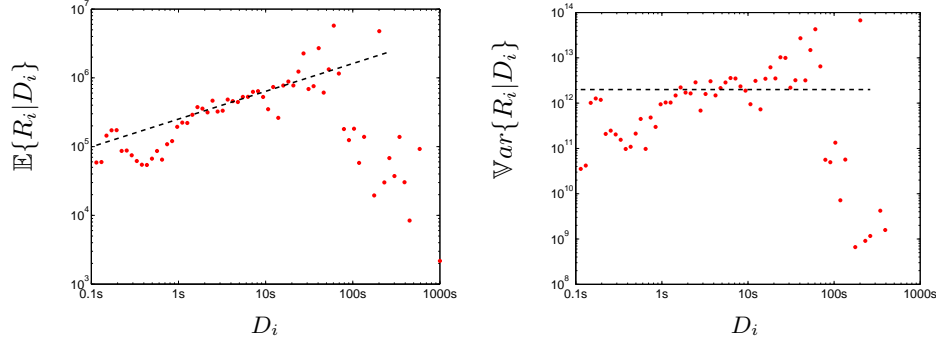


Figure 4: First two moments of the conditional probability of the flow rates given the flow durations. The red dots are empirical histograms estimated on the in2p3 trace. The black lines are not estimated from the empirical histograms. They have slope (left) $\beta - 1 = 0.4037$ where β is estimated from the tail indices α_{SI} and α_{ON} – (right) slope 0. They have been vertically adjusted to fit the data in the range $[1 - 100]$ s

obtained the estimations $\alpha_{SI} = 0.8544$ and $\alpha_{ON} = 1.1994$, which gives $\beta = 1.4037$, $\alpha' = 0.392$. Note that $\alpha' < 1$, so that the autocovariance function calculated in equation (15) theoretically diverges. As we shall see, the long-range dependence observed in Section 4.1 is in fact *pseudo long-range dependence*, *i.e.*, we are interested in the autocovariance function decrease in a finite-scale range, so that this divergence is not a problem (let us recall here that in Section 4.1 we have estimated the values $\hat{H} = 0.901 \pm 0.045$ in the scale range: $[0.64 - 40.96]$ s).

Figure 4 shows the conditional moments $\mathbb{E}\{R_i|D_i\}$ and $\text{Var}\{R_i|D_i\}$, as functions of D_i , estimated from the in2p3 trace. We plotted lines corresponding to the model of Proposition 3.5: $\mathbb{E}\{R_i|D_i\} = K D_i^{\beta-1}$ and $\text{Var}\{R_i|D_i\} = V$, where K, V are constants, with the estimated value of β . These lines are vertically adjusted and show a good match in the range $[1 - 100]$ s, with the values $K = 10^{5.2}$, $V = 10^{12.4}$. This confirms the relevance of our model in this scale range, which roughly corresponds to the scale range in which we estimated the Hurst parameter in Section 4.1. With these values of K and V , we can compare the two terms of the autocovariance function of equation (15). We find that in the scale range of estimation of H ($[0.64 - 40.96]$ s), the first term of equation (15) is always five times smaller than the second term. We conclude that the long-range dependence observed on Figure 3 is pseudo long-range dependence controlled by the second term of equation (15), *i.e.*, by the tail index of the flow-duration distribution α_{ON} . The estimated value of the Hurst parameter shows a very good agreement with relation (1a) with $\alpha_H = \alpha_{ON}$, confirming the accuracy of the model in Proposition 3.5 in predicting the Hurst parameter's value provided that the terms of equation (15) are appropriately compared in the scale range of interest.

5 Conclusion

In this work, we developed a model taking into account the correlation existing between flow rates and durations responsible for the different tail indices α_{SI}

and α_{ON} . Based only on specifications of the first two order moments of the conditional probability of the rates given the duration, we showed that the instantaneous bandwidth exhibits long-range dependence with a Hurst parameter as in relation (1a), where the controlling tail index α_H can be either α_{ON} or a combination of α_{ON} and α_{SI} , whichever is the smaller. We also showed that, in certain circumstances, pseudo long-range dependence with a different Hurst parameter can be observed in a finite scale range. In the case where the correlation between flow rates and durations vanishes, our results coincide with the usual results of classical models.

Finally, we validated the ability of our model to predict the correct Hurst parameter on a real web traffic trace. However, the non-stationarity inherent to real Internet data at large time scales (around one hour) constrained us to restrain our study to 30 minutes of traces. Consequently, we observed pseudo long-range dependence, and longer stationary traces would be required to illustrate the ability of our model to predict asymptotic long-range dependence, which, as we saw, can exhibit a different Hurst parameter.

The model proposed in this work is the first model, to the best of our knowledge, that includes the correlation between flow rates and durations. This correlation is very important. It has been observed for over a decade [?], and is likely to become even more important with the appearance of FTTH and flow-aware approaches. Further theoretical developments would be needed to give a more rigorous mathematical statement of our results, in particular concerning the Gaussian limit, and a potential other limit that could arise similarly to classical models. Also, we assumed a Poisson flow arrival, which is a useful simplification to conduct the calculations, but might not always hold. Considering less restrictive flow-arrival processes would also be an interesting improvement of our results, be it only to clarify in which situations a correlated flow-arrival process can affect the aggregate traffic self-similarity.



Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
<http://www.inria.fr>
ISSN 0249-6399