Proof of a Fundamental Result in Self-Similar Traffic Modeling

Murad S. Taqqu, Walter Willinger and Robert Sherman *

Abstract

We state and prove the following key mathematical result in self-similar traffic modeling: the superposition of many *ON/OFF* sources (also known as *packet trains*) with strictly alternating *ON*- and *OFF*-periods and whose *ON*-periods or *OFF*-periods exhibit the *Noah Effect* (i.e., have high variability or infinite variance) can produce aggregate network traffic that exhibits the *Joseph Effect* (i.e., is self-similar or long-range dependent). There is, moreover, a simple relation between the parameters describing the intensities of the Noah Effect (high variability) and the Joseph Effect (self-similarity). This provides a simple physical explanation for the presence of self-similar traffic patterns in modern high-speed network traffic that is consistent with traffic measurements at the source level. We illustrate how this mathematical result can be combined with modern high-performance computing capabilities to yield a simple and efficient linear-time algorithm for generating self-similar traffic traces.

We also show how to obtain in the limit a Lévy stable motion, that is, a process with stationary and independent increments but with infinite variance marginals. While we have presently no empirical evidence that such a limit is consistent with measured network traffic, the result might prove relevant for some future networking scenarios.

1 Introduction

In our recent paper [22], we stated a mathematical result that allows for a simple and plausible physical explanation for the observed self-similarity of measured Ethernet LAN traffic (e.g., see [12, 13]), involving the traffic generated by the individual sources or source-destination pairs that make up the aggregate packet stream. Developing an approach originally suggested by Mandelbrot [15] and brought to the attention of probabilists by Taqqu and Levy [19], we presented (without proof) a result that states that the superposition of many strictly alternating independent and identically distributed *ON/OFF* sources (also known as "packet trains"; e.g., see [8]), each of which exhibits a phenomenon called the "Noah Effect", results in self-similar aggregate traffic. Intuitively, the Noah Effect for an individual *ON/OFF* source model results in *ON*- and *OFF*-periods, i.e., "train lengths" and "intertrain distances" that can be very large with non-negligible probability; that is, each *ON/OFF* source individually exhibits characteristics that cover a wide range of time scales. The Noah Effect is synonymous with the *infinite variance syndrome*, and as the mathematical vehicle for modeling such phenomena we use *heavy-tailed* distributions with infinite variance (e.g., Pareto or truncated stable distributions). The *ON*- and *OFF*-periods are not required to have the same distribution.

In this paper, we provide the proof of this fundamental result in self-similar traffic modeling as stated in [22]. The proof does not follow from the work of Mandelbrot [15] or Taqqu and Levy [19]; it is more delicate and requires a different approach and new methodologies. By presenting the mathematical results in the well-known framework of the popular *ON/OFF* sources or packet train models, we are able to identify the Noah Effect as the essential point of departure from traditional to self-similar traffic modeling. Moreover, the parameter α describing the "intensity"

*M.S. Taqqu is with the Department of Mathematics, Boston University, Boston, MA 02215. W. Willinger is with AT&T Labs - Research, Murray Hill, NJ 07974. R. Sherman is with the California Institute of Technology, Pasadena, CA 91125.

of the Noah Effect (or equivalently, the "heaviness" of the tail of the corresponding infinite variance distribution) of the *ON*- and/or *OFF*-periods of a "typical" source is related to the Hurst parameter H, where the latter has been suggested in [12] as a measure of the degree of self-similarity (or equivalently, of the "Joseph Effect") of the aggregate traffic stream. Our results apply in practice when there is a large enough number of *ON/OFF* sources, and for time scales that are within natural cut-offs. The lower cut-off for the time scales reflects the fact that we essentially ignore (at the source level as well as at the aggregate level) the fine-structure that is concerned with how individual packets are sent over the raw media (in our case, a coaxial cable) and is determined by the media-access protocols (in our case, the Ethernet's CSMA/CD protocol). Thus, our results also provide evidence that for a sufficiently high levels of aggregation (large M) and for time scales away from the lower cut-off, the nature of aggregate Ethernet LAN traffic does not depend on the fine structure of the underlying media-access mechanisms The upper cut-off region for the time scales arises naturally from the presence of daily cycles that can be found in most working packet networks.

As an application, we combine this mathematical result with modern high-performance computing and communication capabilities, in order to obtain a highly efficient, linear-time parallel algorithm for synthetically generating self-similar network traffic at the packet level. Such fast algorithms for generating long traces of realistic network traffic have hitherto been scarce but are invaluable for a systematic study of many of the currently unsolved issues related to network engineering and traffic management of modern high-speed networks.

We also show in Section 2.3 how it is possible to obtain in the limit, the Lévy stable motion, an infinite variance process whose increments are stationary and independent. Although the practical relevance of this particular result to networking is not clear at this stage, it is related to Theorem 1, and we provide here a precise statement and a rigorous proof.

2 Self-Similarity via Infinite Variance Phenomena

The setting considered in this section is the same as the one presented in [22]; to ensure that our presentation below is self-contained, we repeat here the basic mathematical framework and the corresponding notation. Recall that in [21], we presented an idealized *ON/OFF* source model which allows for long packet trains ("*ON*" periods, i.e., periods during which packets arrive at regular intervals) and long inter-train distances ("*OFF*" periods, i.e., periods with no packet arrivals). In that model, however, the *ON*- and *OFF*-periods did not strictly alternate: they were i.i.d. and hence an *ON*-period could be followed by other *ON*-periods, and an *OFF*-period by other *OFF*-periods. The model was a relatively straightforward extension of the one first introduced by Mandelbrot [15] and Taqqu and Levy [19]. The setting considered here and in [22] differs from our earlier work in the sense that the processes of interest have strictly alternating *ON*- and *OFF*-periods and agree therefore with the *ON/OFF* source models commonly considered in the communications literature. The *ON*- and *OFF*-periods, moreover, may have different distributions, either with infinite or finite variance (a partial treatment of the finite variance case can be found in [11]). Although our main result is essentially the same as in [21], namely, that the superposition of many such packet trains exhibits, on large time scales, the self-similar behavior that has been observed in the Ethernet LAN traffic data and WAN traces (see [12]), the case of strictly alternating *ON/OFF* sources is much more delicate, and we provide here a rigorous proof. For related work, we refer to [14] and [6].

2.1 The Case of Homogeneous Sources

2.1.1 Notation and Assumptions

Suppose first that there is only one source and focus on the stationary binary time series $\{W(t),\ t \geq 0\}$ it generates. $W(t) = 1$ means that there is a packet at time t and $W(t) = 0$ means that there is no packet. Viewing $W(t)$ as the reward at time t, we have a reward of 1 throughout an *ON*-period, then a reward of 0 throughout the following *OFF*-period, then 1 again, and so on. The length of the *ON*-periods are i.i.d., those of the *OFF*-periods are i.i.d., and the lengths of *ON*- and *OFF*-periods are independent. The *ON*- and *OFF*-period lengths may have different distributions. An *OFF*-period always follows an *ON*-period, and it is the pair of *ON*- and *OFF*-periods that defines an interrenewal period.

Suppose now that there are M i.i.d. sources. Since each source sends its own sequence of packet trains it has its own reward sequence $\{W^{(m)}(t),\ t \geq 0\}$. The superposition or cumulative packet count at time t is $\sum_{m=1}^{M} W^{(m)}(t)$. Rescaling time by a factor T, consider

$$W_M^*(Tt) = \int_0^{Tt} \left(\sum_{m=1}^{M} W^{(m)}(u) \right) du,$$

the aggregated cumulative packet counts in the interval $[0, Tt]$. We are interested in the statistical behavior of the stochastic process $\{W_M^*(Tt),\ t \geq 0\}$ for large M and T. This behavior depends on the distributions on the *ON*- and *OFF*-periods, the only elements we have not yet specified. Motivated by the empirically derived fractional Brownian motion model for aggregate cumulative packet traffic in [21], or equivalently, by its increment process, the so-called fractional Gaussian noise model for aggregate traffic (i.e., number of packets per time unit), we want to choose these distributions in such a way that, as $M \to \infty$ and $T \to \infty$, $\{W_M^*(Tt),\ t \geq 0\}$ adequately normalized is $\{\sigma_{\lim} B_H(t),\ t \geq 0\}$, where $\sigma_{\lim}$ is a finite positive constant and B_H is *fractional Brownian motion*, the only Gaussian process with stationary increments that is self-similar. By self-similar, we mean that the finite-dimensional distributions of $\{T^{-H} B_H(Tt),\ t \geq 0\}$ do not depend on the chosen time scale T. The parameter $1/2 \leq H < 1$ is called the *Hurst parameter* or the *index of self-similarity.* Fractional Brownian motion is a Gaussian process with mean zero, stationary increments and covariance function $EB_H(s)B_H(t) = (1/2)\{s^{2H} + t^{2H} - |s-t|^{2H}\}$. Its increments $G_j = B_H(j) - B_H(j-1),\ j = 1, 2, \ldots$ are called *fractional Gaussian noise.* They are strongly correlated:

$$EG_H(j)G_H(j+k) \sim H(2H-1)k^{2H-2} \quad \text{as} \quad k \to \infty,$$

where $a_k \sim b_k$ means $a_k/b_k \to 1$ as $k \to \infty$. The power law decay of the covariance characterizes long-range dependence. The higher the H the slower the decay. For more information about fractional Brownian motion and fractional Gaussian noise, refer for example to Chapter 7 of Samorodnitsky and Taqqu [16].

To specify the distributions of the *ON/OFF*-periods, let

$$f_1(x), \quad F_1(x) = \int_0^x f_1(u)du, \quad F_{1c}(x) = 1 - F_1(x),$$

$$\mu_1 = \int_0^\infty x f_1(x) dx, \quad \sigma_1^2 = \int_0^\infty (x - \mu_1)^2 f_1(x) dx$$

denote the probability density function, cumulative distribution function, complementary (or tail) distribution, mean length and variance of an *ON*-period, and let $f_2, F_2, F_{2c}, \mu_2, \sigma_2^2$ correspond to an *OFF*-period. Assume as $x \to \infty$,

$$\text{either} \quad F_{1c}(x) \sim \ell_1 x^{-\alpha_1} L_1(x) \text{ with } 1 < \alpha_1 < 2 \quad \text{or} \quad \sigma_1^2 < \infty,$$

and

$$\text{either} \quad F_{2c}(x) \sim \ell_2 x^{-\alpha_2} L_2(x) \text{ with } 1 < \alpha_2 < 2 \quad \text{or} \quad \sigma_2^2 < \infty,$$

where $\ell_j > 0$ is a constant and $L_j > 0$ is a slowly varying function at infinity, that is $\lim_{x\to\infty} L_j(tx)/L_j(x) = 1$ for any $t > 0$. For example, $L_j(x)$ could be asymptotic to a constant, to $\log x$, to $(\log x)^{-1}$, etc. Since the function L_j will be used as normalization in (1) below, it is preferable not to absorb the constant ℓ_j into it. (We also assume that either probability densities exist or that $F_j(0) = 0$ and F_j is non-arithmetic, where F_j is called arithmetic if it is concentrated on a set of points of the form $0, \pm\lambda, \pm 2\lambda, \cdots$.) Note that the mean μ_j is always finite but the variance σ_j^2 is infinite when $\alpha_j < 2$. For example, F_j could be Pareto, i.e. $F_{jc}(x) = K^{\alpha_j} x^{-\alpha_j}$ for $x \geq K > 0$, $1 < \alpha_j < 2$ and equal 0 for $x < K$, or it could be exponential. Observe that the distributions F_1 and F_2 of the *ON*- and *OFF*-periods are allowed to be different. One distribution, for example, can have a finite variance, the other an infinite variance.

2.1.2 A Fundamental Result in Self-Similar Traffic Modeling

In order to state the main result, the following notation will be convenient. When $1 < \alpha_j < 2$, set $a_j = \ell_j(\Gamma(2-\alpha_j))/(\alpha_j - 1)$. When $\sigma_j^2 < \infty$, set $\alpha_j = 2, L_j \equiv 1$ and $a_j = \sigma_j^2/2$. The normalization factors and the limiting constants in the theorem below depend on whether

$$b = \lim_{t\to\infty} t^{\alpha_2-\alpha_1} \frac{L_1(t)}{L_2(t)}$$

is finite, 0, or infinite. If $0 < b < \infty$ (implying $\alpha_1 = \alpha_2$ and $b = \lim_{t\to\infty} L_1(t)/L_2(t)$), set $\alpha_{\min} = \alpha_1 = \alpha_2$,

$$\sigma_{\lim}^2 = \frac{2(\mu_2^2 a_1 b + \mu_1^2 a_2)}{(\mu_1+\mu_2)^3 \Gamma(4-\alpha_{\min})}, \quad \text{and} \quad L = L_2;$$

if, on the other hand, $b = 0$ or $b = \infty$, set

$$\sigma_{\lim}^2 = \frac{2\mu_{\max}^2 a_{\min}}{(\mu_1+\mu_2)^3 \Gamma(4-\alpha_{\min})}, \quad \text{and} \quad L = L_{\min},$$

where min is the index 1 if $b = \infty$ (e.g. if $\alpha_1 < \alpha_2$) and is the index 2 if $b = 0$, max denoting the other index. We claim that under the conditions stated above the following holds:

Theorem 1. *For large M and T, the aggregate cumulative packet process $\{W_M^*(Tt),\ t \geq 0\}$ behaves statistically like*

$$TM \frac{\mu_1}{\mu_1+\mu_2} t + T^H \sqrt{L(T)M}\, \sigma_{\lim} B_H(t)$$

where $H = (3-\alpha_{\min})/2$ and $\sigma_{\lim}$ is as above. More precisely,

$$\mathcal{L}\lim_{T\to\infty} \mathcal{L}\lim_{M\to\infty} \frac{\left(W_M^*(Tt) - TM\frac{\mu_1}{\mu_1+\mu_2}t\right)}{T^H L^{1/2}(T) M^{1/2}} = \sigma_{\lim} B_H(t), \tag{1}$$

where $\mathcal{L}\lim$ means convergence in the sense of the finite-dimensional distributions.

Heuristically, Theorem 1 states that the mean level given by $TM(\mu_1/(\mu_1+\mu_2))t$ provides the main contribution for large M and T. Fluctuations from that level are given by the fractional Brownian motion $\sigma_{\lim} B_H(t)$ scaled by a lower order factor $T^H L(T)^{1/2} M^{1/2}$. As in [19], it is essential that the limits be performed in the order indicated. We will consider the case where the limits are performed in reverse order in more detail in Section 2.3 below. Also note that $1 < \alpha_{\min} < 2$ implies $1/2 < H < 1$, i.e., long-range dependence. Thus, the main ingredient that is needed to obtain an $H > 1/2$ is the heavy-tailed property

$$F_{jc}(x) \sim \ell_j x^{-\alpha_j} L_j(x), \qquad \text{as } x \to \infty,\ 1 < \alpha_j < 2 \tag{2}$$

for the *ON*- or *OFF*-period; that is, a hyperbolic tail (or power law decay) for the distributions of the *ON*- or *OFF*-periods with an α between 1 and 2. A similar result obtains if $W_M^*(Tt)$ is replaced by the cumulative number of bytes in $[0, Tt]$.

2.1.3 Some Special Cases

(a) Suppose $\alpha_1 = \alpha_2 = 2$, that is, the *ON*- and *OFF*-periods both have finite variance. Then (1) holds with $H = 1/2$ and $L = 1$, and the limit is

$$\left[\frac{\mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2}{(\mu_1+\mu_2)^3}\right]^{1/2} B(t)$$

where $B(t)$ is Brownian motion.

(b) Suppose $F_1 \equiv F_2$, that is, the *ON*- and *OFF*-periods have the same distribution. Then $\sigma_{\lim}^2 = a/(2\mu\Gamma(4-\alpha))$ where α and μ are, respectively, the common index and mean. In particular, if $\alpha < 2$, (1) holds with $H = (3-\alpha)/2$, $L = L_1 = L_2$ and the limit is

$$\left[\frac{l}{2\mu(\alpha-1)(2-\alpha)(3-\alpha)}\right]^{1/2} B_H(t).$$

(c) Suppose $\alpha_{\min} < \alpha_{\max}$, that is, either the *ON*- or *OFF*-period has infinite variance and one of them has a heavier probability tail than the other. Then (1) holds with $H = (3-\alpha_{\min})/2$, $L = L_{\min}$ and the limit is

$$\left[\frac{2\mu_{\max}^2 \ell_{\min}}{(\mu_1+\mu_2)^3(\alpha_{\min}-1)(2-\alpha_{\min})(3-\alpha_{\min})}\right]^{1/2} B_H(t).$$

2.1.4 A Weak Convergence Result

Proceeding as in [18] or establishing the necessary tightness conditions directly (see Section 3), one can easily verify that $\mathcal{L}\lim_{T\to\infty}$, in Theorem 1, can be replaced by the convergence, as $T \to \infty$, of probability measures in the space $C[0,\infty)$ of continuous functions on $[0,\infty)$; this mode of convergence is also known as *weak convergence* (e.g., see [1]). Weak convergence implies, for example, convergence of maxima and minima, and is particularly useful in queuing applications.

2.2 Heterogeneous Sources

Weakening our hypothesis, suppose now that not all sources are identical. Assume R types of sources and that there is a proportion $M^{(r)}/M$ of sources of type $r = 1, \ldots, R$, with $M^{(r)}/M$ not converging to 0, as $M \to \infty$ $(r = 1, \ldots, R)$. For $j = 1, 2$, let $F_j^{(r)}, \alpha^{(r)}, \sigma^{(r)}, L^{(r)}$ be the "characteristics" of source r.

Theorem 2. *For large $M^{(r)}$, $r = 1, \ldots, R$ and large T, the aggregated cumulative packet traffic $\{W_M^*(Tt),\ t \geq 0\}$ behaves statistically like*

$$T \left(\sum_{r=1}^{R} M^{(r)} \frac{\mu_1^{(r)}}{\mu_1^{(r)} + \mu_2^{(r)}} \right) t + \sum_{r=1}^{R} T^{H^{(r)}} \sqrt{L^{(r)}(T) M^{(r)}}\ \sigma_{\lim}^{(r)}\ B_{H^{(r)}}(t)$$

where $H^{(r)} = (3 - \alpha_{\min}^{(r)})/2$ and the $B_{H^{(r)}}$ are independent fractional Brownian motions.

Theorem 2 states that the limit is a superposition of independent fractional Brownian motions with different $H^{(r)}$'s. As far as the fluctuations are concerned, however, the term with the highest $H^{(r)}$ (or equivalently, the term with the smallest $\alpha^{(r)}$) ultimately dominates as $T \to \infty$. Note that Theorem 2 allows for the possibility that for some source types r, the distributions of the corresponding *ON*- and *OFF*-periods have finite variance. In this case, the contribution of type r to the limit will consist of an ordinary Brownian motion component. Theorem 2 is an easy consequence of Theorem 1.

2.3 Taking the Limit in Reverse Order

We have seen that to obtain a fractional Brownian motion limit (see (1)), it was essential to first let $M \to \infty$ and then $T \to \infty$. This limit regime was shown in [22] to be consistent with measured aggregate traffic from Ethernet LAN environments; that is, under non-trivial overall loads on a Ethernet LAN, there are typically a few hundred active source-destination pairs (large M) and for time scales T that range through several orders of magnitude, the measurements exhibit approximately Gausssian marginals and long-range dependence.

In this section, we examine what happens when one first lets $T \to \infty$ and then $M \to \infty$. While at this stage, we do not have empirical evidence for this limit regime in the traffic measurements that are presently available to us, the result stated below is of independent interest and is given here for completeness. It may prove relevant in some future networking scenarios.

A partial treatment involving only heavy-tailed *ON*-periods can be found in Konstantopoulos and Lin [10]. We describe here what happens in the general case. The situation, in fact, is very similar to that considered in Taqqu and Levy [19]. The limit, when first $T \to \infty$, is not fractional Brownian motion but the Lévy stable motion, a process with stationary independent increments.

The Lévy stable motion $\{\Lambda_{\alpha,\sigma,\beta}(t),\ t \geq 0\}$ is self-similar with index $1/\alpha$. $\sigma > 0$ is a scale parameter and $-1 \leq \beta \leq 1$ is the skewness parameter. When $\alpha = 2$, it has continuous paths and is Brownian motion, but when $1 < \alpha < 2$, it has discontinuous paths and infinite variance. Because it has stationary independent increments, it is sufficient to provide the distribution (or its Fourier transform, the characteristic function) at time $t = 1$. The characteristic function of $\Lambda_{\alpha,\sigma,\beta}(1)$ belongs to the "stable" family and is given by

$$Ee^{i\theta\Lambda_{\alpha,\sigma,\beta}(1)} = \exp\{-\sigma^\alpha |\theta|^\alpha (1 - i\beta(\text{sign}\theta)) \tan\frac{\pi\alpha}{2}\},\ -\infty < \theta < \infty, \tag{3}$$

when $\alpha < 2$. In particular, the complementary distribution function satisfies

$$P[\Lambda_{\alpha,\sigma,\beta}(1) > x] \sim C_\alpha \sigma^\alpha \frac{1+\beta}{2} x^{-\alpha} \quad \text{as} \quad x \to \infty$$

where

$$C_\alpha = \left| \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} \right|, \tag{4}$$

hence $\Lambda_{\alpha,\sigma,\beta}(1)$ has heavy tails. By self-similarity,

$$P[\Lambda_{\alpha,\sigma,\beta}(t) > x] = P[t^{1/\alpha}\Lambda_{\alpha,\sigma,\beta}(1) > x] \sim C_\alpha \sigma^\alpha \frac{1+\beta}{2} t x^{-\alpha} \quad \text{as} \quad x \to \infty.$$

For more details, see Samorodnitsky and Taqqu [16]. The characteristic function (3) can also be used when $\alpha = 2$. It reduces then to $\exp\{-\sigma^2\theta^2\}$, which is the characteristic function of a Gaussian random variable.

For simplicity of presentation, we will assume that the slowly varying functions L_1 and L_2 in the distributions F_1 and F_2 of the *ON* and *OFF* periods are equal to 1. The next theorem shows that the Lévy stable motion appears as $T \to \infty$ even when one considers only a single source ($W(u) = W^{(m)}(u)$ for a fixed source m).

Theorem 3. *Let* $\alpha_{\min} = \min(\alpha_1, \alpha_2)$. *Then*

$$\mathcal{L} \lim_{T\to\infty} \frac{1}{T^{1/\alpha_{\min}}} \int_0^{Tt} (W(u) - EW(u))du$$

equals

$$\frac{\mu_2 \ell_1^{1/\alpha_1}}{(\mu_1+\mu_2)^{1+1/\alpha_1}} \Lambda_{\alpha_1,\sigma,1}(t) \quad \text{if} \quad \alpha_1 < \alpha_2,\ \alpha_1 < 2, \tag{5}$$

$$\frac{\mu_1 \ell_2^{1/\alpha_2}}{(\mu_1+\mu_2)^{1+1/\alpha_2}} \Lambda_{\alpha_2,\sigma,-1}(t), \quad \text{if} \quad \alpha_2 < \alpha_1,\ \alpha_1 < 2, \tag{6}$$

$$\frac{[\mu_2^\alpha \ell_1 + \mu_1^\alpha \ell_2]^{1/\alpha}}{(\mu_1+\mu_2)^{1+1/\alpha}} \Lambda_{\alpha,\sigma,\beta}(t) \quad \text{if} \quad \alpha = \alpha_1 = \alpha_2 < 2, \tag{7}$$

where $\sigma = C_\alpha^{-1/\alpha}$ *and, in (7),*

$$\beta = \frac{\mu_2^\alpha \ell_1 - \mu_1^\alpha \ell_2}{\mu_2^\alpha \ell_1 + \mu_1^\alpha \ell_2}. \tag{8}$$

In the case $\alpha_1 = \alpha_2 = 2$, *the limit is*

$$\frac{[(\mu_2\sigma_1)^2 + (\mu_1\sigma_2)^2]^{1/2}}{(\mu_1+\mu_2)^{3/2}} B(t),$$

where σ_1^2 *and* σ_2^2 *are the variances of the ON and OFF period respectively, and where* $B(t)$ *is standard Brownian motion.*

Remarks: (1) When $\alpha_1 = \alpha_2$, the limiting process is symmetric if $\mu_2\ell_1 = \mu_1\ell_2$, for example, if the *ON* and *OFF* periods are identically distributed.

(2) When $\alpha_1 = \alpha_2 = 2$, the limit is the same as in the case $T \to \infty$, $M \to \infty$.

(3) If the slowly varying functions L_1 and L_2 are not asymptotically equal to 1, then $T^{1/\alpha_{\min}}$ should be replaced by $T^{1/\alpha_{\min}}\widetilde{L}(T)$, where $\widetilde{L}(T)$ is a slowly varying function. For example, if $\alpha_{\min} = \alpha_1$, then $\widetilde{L}(T)$ is such that $\lim_{T\to\infty} \widetilde{L}(T)^{-\alpha_1} L_1(T^{1/\alpha_1}\widetilde{L}(T)x) = 1$ for all $x > 0$ (see [7], relation (2.6.4)).

(4) The result of Theorem 3 holds also in the sense of weak convergence (in the Skorohod topology J_1 on $D[0,1]$, the space of right-continuous functions on $[0,1]$ with left limits), that is, it holds as convergence of random processes in t.

If, after taking the limit $T \to \infty$, we accumulate a number M of sources, and/or let $M \to \infty$, the limiting processes do not change.

Theorem 4. *The conclusion of Theorem 3 holds also for*

$$\mathcal{L} \lim_{T\to\infty} \frac{1}{(MT)^{1/\alpha_{\min}}} \int_0^{Tt} \sum_{m=1}^{M} (W^{(m)}(u) - EW^{(m)}(u))du, \quad M \geq 1,$$

and for

$$\mathcal{L} \lim_{M\to\infty} \mathcal{L} \lim_{T\to\infty} \frac{1}{(MT)^{1/\alpha_{\min}}} \int_0^{Tt} \sum_{m=1}^{M} (W^{(m)}(u) - EW^{(m)}(u))du.$$

This is an immediate consequence of Theorem 3 because each m yields an independent copy of the Lévy stable motion.

When $\alpha_{\min} < 2$ one then gets different limits, depending upon whether $T \to \infty$, $M \to \infty$ or $M \to \infty, T \to \infty$. The limit in the first case is the Gaussian fractional Brownian motion. The limit in the second case is the infinite variance stable Lévy motion. In practice, the behavior depends on the relative sizes of M and T. Observe that this does not contradict the behavior of $V(T) =$Var$(\int_0^T W(u)du)$. The relation $V(T) \approx CT^{2H}$ holds for large but finite T. If the limiting distribution is Gaussian, then T^H is the right normalization for the process and $V(T)$ will converge to a finite limit. If the limiting distribution is the Lévy stable motion, then the correct normalization for the process is $T^{1/\alpha_{\min}}$. Since

$$H = \frac{3-\alpha_{\min}}{2} > \frac{1}{\alpha_{\min}} \quad \text{for} \quad 1 < \alpha_{\min} < 2,$$

T^H is greater than $T^{1/\alpha_{\min}}$. Hence the normalization $T^{1/\alpha_{\min}}$ makes Var$(T^{-1/\alpha_{\min}} \int_0^T W(u)du) = T^{-2/\alpha_{\min}} V(T) \approx CT^{2(H-1/\alpha_{\min})}$ tend to infinity as $T \to \infty$, which is consistent with the fact that the Lévy stable motion has infinite variance for $\alpha_{\min} < 2$.

3 Proof of Theorem 1

We shall prove Theorem 1; Theorem 2 is an easy extension. To this end, note that the aggregated reward $\int_0^t W(u)du$ by time t has variance

$$V(t) = \text{Var}\left(\int_0^t W(u)du\right) = 2\int_0^t \left(\int_0^v \gamma(u)du\right) dv \tag{9}$$

where $\gamma(u) = EW(u)W(0) - (EW(0))^2$ denotes the covariance function of W. We claim that it is sufficient to prove

$$V(t) \sim \sigma_{\lim}^2 t^{2H} L(t) \quad \text{as} \quad t \to \infty. \tag{10}$$

Indeed, suppose that this last relation holds. Then

$$\mathcal{L} \lim_{M\to\infty} M^{-1/2} \sum_{m=1}^{M} (W^{(m)}(t) - EW^{(m)}(t)) = G(t), \quad t \geq 0,$$

by the usual Central Limit Theorem. Moreover, the process $\{G(t),\ t \geq 0\}$ is Gaussian and stationary (since the $W_m(t)$'s are stationary) and has mean zero and covariance function $\{\gamma(t),\ t \geq 0\}$. Now, (10) implies

$$\mathcal{L} \lim_{b\to\infty} (T^{2H} L(T))^{-1/2} \int_0^{Tt} G(u)du = \sigma_{\lim} B_H(t),\ t \geq 0. \tag{11}$$

Indeed, the limit in (11) must be Gaussian with mean zero and have stationary increments since the integral of G has these properties. Moreover, by (10), its variance must be $\sigma_{\lim}^2 t^{2H}$ for fixed t (e.g., see [5, Lemma 1, p. 275]). Since these properties characterize the fractional Brownian motion $\sigma_{\lim} B_H(t)$, (see for example, [16, Corollary 7.2.3, p. 320]), Relation (11) follows, proving Theorem 1.

It is therefore sufficient to establish (10). Since the reward process W is stationary, its mean is $EW(t) = P(t \text{ is } on) = \mu_1/(\mu_1+\mu_2)$ for all t. (By "t is *on*", we mean "t is in an *ON*-period".) The main difficulty is to evaluate its covariance function $\gamma(t)$. If $\pi_{11}(t) = P(\text{time } t \text{ is } on \mid \text{time } 0 \text{ is } on)$, then $EW(0)W(t) = P(\text{times } 0 \text{ and } t \text{ are } on)$ $= \pi_{11}(t)\mu_1/\ (\mu_1+\mu_2)$ and hence

$$\gamma(t) = \frac{\mu_1}{\mu_1+\mu_2}\left[\pi_{11}(t) - \frac{\mu_1}{\mu_1+\mu_2}\right]. \tag{12}$$

Since as $t\to\infty$, $\pi_{11}(t) \to P(\text{time } t \text{ is } on) = \mu_1/(\mu_1+\mu_2)$, we have $\lim_{t\to\infty}\gamma(t) = 0$. We need, however, to determine the asymptotic behavior of $\gamma(t)$ or, at least, that of its double integral $V(t)$ given in (9).

Following [4], we shall evaluate the Laplace transform of π_{11}, in order to get that of γ and V. The renewal equation for $\pi_{11}(t)$ is

$$\pi_{11}(t) = G_{1c}(t) + \int_0^t F_{1c}(t-u)dH_{12}(u), \tag{13}$$

with $G_{1c}(t) = P(\text{remaining life of the first } on \text{ interval} > t \mid \text{time } 0 \text{ is } on)$, and where $H_{12}(u)$ is the renewal function corresponding to the inter-renewal distribution $F_1 * F_2$. ($H_{12} = \sum_{k=1}^{\infty}(F_1 * F_2)^{*k}$. Its density $h_{12}(u)$, when defined, is the probability density that the end of an *OFF*-period occurs at time u given that time 0 is *on*). To understand (13) note that if 0 is *on*, then t can be *on* either if it belongs to the same *ON*-period as time 0 or, if there is a subsequent *OFF*-period, that there is an *ON/OFF* transition at some time u and t belongs to this subsequent *ON*-period. The corresponding Laplace transform is

$$\widehat{\pi}_{11}(s) = \widehat{G}_{1c}(s) + \widehat{h}_{12}(s)(1-\widehat{f}_1(s))/s.$$

(The Laplace transform of a function $A(t)$ is denoted $\widehat{A}(s) = \int_0^\infty e^{-st}A(t)dt,\ s>0$.)[1]

Since we start in a stationary regime, we have $G_{1c}(t) = \mu_1^{-1}\int_t^\infty F_{1c}(u)du$, and thus

$$\widehat{G}_{1c}(s) = \frac{1}{s} - \frac{1-\widehat{f}_1(s)}{\mu_1 s^2}, \tag{14}$$

and since (we have a delayed renewal process here)

$$\widehat{h}_{12}(s) = \frac{(1-\widehat{f}_1(s))\widehat{f}_2(s)}{\mu_1 s[1-\widehat{f}_1(s)\widehat{f}_2(s)]},$$

we can derive $\widehat{\pi}_{11}(s)$. Then, using (12), we get, as in [4],

$$\widehat{\gamma}(s) = \frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2 s} - \frac{(1-\widehat{f}_1(s))(1-\widehat{f}_2(s))}{(\mu_1+\mu_2)s^2[1-\widehat{f}_1(s)\widehat{f}_2(s)]}. \tag{15}$$

Since by (14), $(1-\widehat{f}_j)/\mu_j s = 1 - s\widehat{G}_{jc}$ and $\widehat{f}_j = 1 - \mu_j s + \mu_j s^2 \widehat{G}_{jc},\ j=1,2,$ $\widehat{\gamma}$ becomes

$$\begin{aligned}\widehat{\gamma}(s) \quad = \quad & \frac{\mu_1\mu_2}{\mu_1+\mu_2}\ \frac{1}{(\mu_1+\mu_2)s}\ - \\ & \frac{\mu_1\mu_2}{\mu_1+\mu_2}\ \frac{(1-s\widehat{G}_{1c}(s))(1-s\widehat{G}_{2c}(s))}{1-\prod_{j=1}^{2}(1-\mu_j s+\mu_j s^2\widehat{G}_{jc}(s))}.\end{aligned} \tag{16}$$

[1]Although convenient, it is not necessary to assume that densities exist. (In this case, it is sufficient to suppose that F_j is non-arithmetic and $F_j(0)=0,\ j=1,2$.) The functions $\widehat{f}_j$ and $\widehat{h}_{12}$ are then defined as $\widehat{f}_j = s\widehat{F}_j$ and $\widehat{h}_{12} = s\widehat{H}_{12}$.

The denominator of the second fraction behaves like $(\mu_1+\mu_2)s - s^2(\mu_1\widehat{G}_{1c}+\mu_2\widehat{G}_{2c}+\mu_1\mu_2)$ as $s\to 0$. (It is important not to drop the s^2 term.) Substituting this in (16), simplifying and then using (9), yields

$$\begin{aligned}\widehat{V}(s) &= 2\frac{1}{s^2}\widehat{\gamma}(s)\\ &\sim \frac{2\left(\mu_2^2\mu_1\widehat{G}_{1c}(s)+\mu_1^2\mu_2\widehat{G}_{2c}(s)-\mu_1^2\mu_2^2\right)}{(\mu_1+\mu_2)^3 s^2},\end{aligned} \tag{17}$$

as $s\to 0$. Now consider two cases:

(a) If $F_{jc}(u)\sim \ell_j u^{-\alpha_j}L_j(u)$ as $u\to\infty$, $1<\alpha_j<2$, then as $t\to\infty$,

$$G_{jc}(t)=\frac{1}{\mu_j}\int_t^\infty F_{jc}(u)du \sim \frac{\ell_j}{\mu_j(\alpha_j-1)}t^{-\alpha_j+1}L_j(t), \tag{18}$$

and consequently (see [2, Theorem 1.7.6]),

$$\mu_j\widehat{G}_{jc}(s)\sim a_j s^{\alpha_j-2}L_j(\frac{1}{s}) \quad \text{as} \quad s\to 0, \tag{19}$$

with $a_j=\ell_j(\Gamma(2-\alpha_j))/(\alpha_j-1)$

(b) If F_j has finite variance σ_j^2, then $\widehat{f}_j(s)\sim 1-\mu_j s+(1/2)(\sigma_j^2+\mu_j^2)s^2$ as $s\to 0$, and hence, by (14),

$$\mu_j\widehat{G}_{jc}(s)\sim(\sigma_j^2+\mu_j^2)/2 \quad \text{as} \quad s\to 0.$$

Recall, that in the case (b), we set $a_j=\sigma_j^2/2$, $\alpha_j=2$ and $L_j\equiv 1$.

Substituting these expressions for $\widehat{G}_{jc}$, $j=1,2$ in (17) we get, in all cases $1<\alpha_1,\alpha_2\le 2$,

$$\begin{aligned}\widehat{V}(s) &\sim \frac{2\left(\mu_2^2 a_1 s^{\alpha_1-4}L_1(\frac{1}{s})+\mu_1^2 a_2 s^{\alpha_2-4}L_2(\frac{1}{s})\right)}{(\mu_1+\mu_2)^3}\\ &= \sigma_{\text{lim}}^2\Gamma(4-\alpha_{\min})s^{\alpha_{\min}-4}L(\frac{1}{s}),\end{aligned} \tag{20}$$

as $s\to 0$, using the notation introduced before the statement of Theorem 1. We want to conclude that

$$V(t)\sim\sigma_{\text{lim}}^2 t^{3-\alpha_{\min}}L(t) \quad \text{as} \quad t\to\infty, \tag{21}$$

which is Relation (10) with $2H=3-\alpha_{\min}$.

Trying to invert the Laplace transform $\widehat{V}(s)$ through integration in the complex plane (viewing s as complex-valued) would involve making assumptions of analyticity that cannot be verified. To get (21) it is sufficient to show that (20) holds and

$$\lim_{\lambda\downarrow 1}\liminf_{x\to\infty}\inf_{t\in[1,\lambda]}\frac{V(tx)-V(x)}{x^{3-\alpha_{\min}}L(x)}\ge 0 \tag{22}$$

(with $L\equiv 1$ when $\alpha_{\min}=2$) by the extended form of Karamata's Tauberian Theorem (e.g., see [2, Theorem 1.7.6]).

It remains to verify (22) and, in order to do so, we will work in the time domain. Recall equation (9) and observe that the contribution of additive terms in V that are non-decreasing (i.e. with non-negative derivative) can be ignored since they will automatically satisfy (22). The key is to express γ in terms of

$$A(t)=H_{12}(t)-\frac{t}{\mu_1+\mu_2}.$$

Recall that $H_{12}(t)$ is the renewal function of an alternating renewal process whose interrenewal periods have distribution $F_1 * F_2$ and mean $\mu_1+\mu_2$. Consequently, the function $A(t)$ is positive for large t and the function $B(t) = A(t)+1$ is always non-negative ([5, Theorem 1, p. 366]). [2] Using the expressions of γ and π_{11} given in (9) and (13) respectively, we get

$$\gamma(t) = \frac{\mu_1}{\mu_1+\mu_2}\left[\frac{\mu_2}{\mu_1+\mu_2}G_{1c}(t) + \int_0^t F_{1c}(t-u)dA(u)\right].$$

Since G_{1c} is non-negative, we will focus on

$$Q(t) = \int_0^t F_{1c}(t-u)dA(u) = A(t) - \int_0^t F_1(t-u)dA(u),$$

since $A(0) = 0$. Integrating the last term by parts and using $F_1(0) = 0$ gives

$$\begin{aligned} Q(t) &= A(t) + \int_0^t A(u)dF_1(t-u) \\ &= A(t) + \int_0^\infty A(u)dF_1(t-u). \end{aligned}$$

(If f_1 exists, $Q(t) = A(t) - \int_0^\infty A(u)f_1(t-u)du$.) Therefore the integral of Q, $Q^*(\tau) := \int_0^\tau Q(t)dt$, equals

$$\begin{aligned} Q^*(\tau) &= \int_0^\tau A(t)dt - \int_0^\infty A(u)F_1(\tau-u)du \\ &= \int_0^\tau A(u)F_{1c}(\tau-u)du. \end{aligned}$$

Expressing Q^* in terms of $B(u) = A(u)+1 \geq 0$, we get

$$\begin{aligned} Q^*(\tau) &= \int_0^\tau B(u)F_{1c}(\tau-u)du - \int_0^\tau F_{1c}(\tau-u)du \\ &= \int_0^\tau B(u)F_{1c}(\tau-u)du - \mu_1 G_1(\tau). \end{aligned}$$

Since the first integral is non-negative, it is sufficient to focus on $-\mu_1 G_1(\tau)$ and to show that Relation (22) is satisfied with $V(t)$ replaced by $-\int_0^t G_1(\tau)d\tau$. That relation holds because $-\int_x^{tx} G_1(\tau)d\tau \geq -tx + x = x(1-t)$ and $\inf_{t\in[1,\lambda]} x(1-t) = x(1-\lambda)$, concluding the proof.

Proof of weak convergence. Having shown that the finite-dimensional distributions converge, in order to establish weak convergence in the space of continuous functions, it is sufficient to prove that tightness holds, for example, that for all $t_1, t_2 > 0$ and large enough T,

$$E|X(Tt_2) - X(Tt_1)|^\gamma \leq C|t_2 - t_1|^\delta, \tag{23}$$

for some $\gamma > 0$ and $\delta > 1$ (see e.g. [1], p. 95). Here $X(t)$ is the limit in (1) after letting $M \to \infty$. Focusing on the case $H > 1/2$, we shall choose $\gamma = 2$ and show that Relation (23) holds with some $\delta > 1$. Since $X(t)$ is Gaussian with variance $V(t)$ given in (9) and has stationary increments,

$$A \equiv E|X(Tt_2) - X(Tt_1)|^2 = T^{-2H}L^{-1}(T)V(T(t_2-t_1)).$$

[2] If $F_1 * F_2$ has finite variance $\sigma_{12}^2 = \sigma_1^2 + \sigma_2^2$, then $A(t) \to (\sigma_{12}^2 - \mu_{12}^2)/2\mu_{12}^2$ where $\mu_{12} = \mu_1 + \mu_2$ (see also [9, p. 195]). $B(t) \to (\sigma_{12}^2 + \mu_{12}^2)/2\mu_{12}^2$. A and B tend to infinity if $F_1 * F_2$ has finite mean but infinite variance. In the latter case, by [20, Theorem 4], if $F_1 * F_2$ is regularly varying at infinity with exponent $-\alpha$, $1 < \alpha < 2$, then A (and B) is regularly varying with exponent $2-\alpha$.

Since $V(t)$ satisfies (10), there is, for any $\epsilon > 0$, a fixed number T_ϵ such that for all $T > T_\epsilon$,

$$A \leq (1+\epsilon)\sigma^2_{\lim}(t_2 - t_1)^{2H}\frac{L(T(t_2-t_1))}{L(T)}.$$

Now, $L(Tu)/L(T)$ tends to 1 as $T \to \infty$ for all $u > 0$ and is uniformly bounded by $C_0 u^{-\epsilon_0}$, for any $\epsilon_0 > 0$. Choosing ϵ_0 so that $\delta = 2H - \epsilon_0 > 1$, yields $A \leq C(t_2 - t_1)^\delta$, which establishes tightness and hence the weak convergence.

4 Proof of Theorem 3

Since we want to evaluate $\int_0^{Tt}(W(u) - EW(u))du$, we shall always consider rewards with their mean $EW(u) = P[\text{time is } on] = \mu_1/(\mu_1+\mu_2)$ subtracted. Let $U_1(1), U_2(1), \ldots$ denote the i.i.d. *ON* periods with distribution F_1 and mean μ_1 and let $U_2(1), U_2(2), \ldots$ denote the i.i.d. *OFF* periods with distribution F_2 and mean μ_2. The 0^{th} *ON* and *OFF* periods $U_1(0)$ and $U_2(0)$, respectively, may have a different distribution in order to ensure stationarity. A renewal interval will include both an *ON* and an *OFF* period.

Since in the k^{th} renewal interval, $k \geq 1$, the reward $1 - \mu_1/(\mu_1+\mu_2)$ lasts for $U_1(k)$ units of time and the reward $0 - \mu_1/(\mu_1+\mu_2)$ lasts for $U_2(k)$ units of time, the cumulative reward (with its mean subtracted) in the k^{th} renewal interval, is

$$\begin{aligned} J(k) &= U_1(k)\left[1 - \frac{\mu_1}{\mu_1+\mu_2}\right] + U_2(k)\left[0 - \frac{\mu_1}{\mu_1+\mu_2}\right] \\ &= \frac{\mu_2}{\mu_1+\mu_2}[U_1(k) - EU_1(k)] - \frac{\mu_1}{\mu_1+\mu_2}[U_2(k) - EU_2(k)] \end{aligned} \tag{24}$$

because $EU_1(k) = \mu_1$, $EU_2(k) = \mu_2$. Denote the renewal epochs by $S(k) = \sum_{j=0}^k (U_1(j) + U_2(j))$ and the number of renewal intervals through time T by $1 + K(T)$; the "1" here corresponds to the 0^{th} renewal interval (which starts at time 0 and ends at time $S(0)$) and the $K(T)^{th}$ renewal includes time T.

We can now express the cumulative reward (with its mean subtracted) from time 0 to time Tt as

$$\int_0^{Tt}(W(u) - EW(u))du = \left[O_p(1) + \sum_{k=1}^{K(Tt)} J(k) + O_p(1)\right] I(Tt > S(0)) + O_p(1)TtI(Tt \leq S(0)). \tag{25}$$

In order to understand this expression, let us first introduce the notation $O_p(1)$.

Any random variable X can be expressed as $X = O_p(1)$, since it satisfies $P[X < \infty] = 1$ and hence is bounded in probability. The notation $O_p(1)$ is useful when it is not necessary to specify X fully, for example, if all one wants is to use the property that, for $\epsilon > 0$, $T^{-\epsilon}O_p(1) = O_p(T^{-\epsilon})$ is a random variable that tends to zero in probability as $T \to \infty$. Note that $O_p(1) + O_p(1) = O_p(1)$ where each $O_p(1)$ possibly denotes a different random variable. The indicator function of a set A is denoted $I(A)$.

Let us now explain (25). Consider the case where $Tt > S(0)$. The first $O_p(1)$ in the bracket in (25) corresponds to the cumulative reward in the 0^{th} renewal interval. The sum $\sum_{k=1}^{K(Tt)} J(k)$ includes the rest, but since Tt falls in the $K(Tt)^{th}$ renewal interval, it is necessary to subtract the excess reward (this is done by the last $O_p(1)$ term in the bracket).

Focus now on the case $Tt \le S(0)$, that is when Tt is included in the 0^{th} renewal interval. The cumulative reward (with its mean subtracted) is cumbersome to express because it depends on whether one starts with an *ON* or *OFF* period and upon what type of period Tt falls. All we need, however, is that its absolute value is bounded by $O_p(1)Tt$ (here $O_p(1) \le 1$). The contribution of this term will be negligible. Indeed, by Markov's inequality, for any $\epsilon > 0$,

$$P\left[O_p(1)Tt\ I(Tt \le S(0)) \ge \epsilon\right] \le \frac{Tt}{\epsilon} P(S(0) \ge Tt)$$

since here $O_p(1) \le 1$. Moreover, $P(S(0) < T) = O(T^{-\alpha_{\min}+1})$ as $T \to \infty$ if $\alpha_{\min} = \min(\alpha_1, \alpha_2) < 2$ (see (18) and $P(S(0) > T) = o(1)$ as $T \to \infty$ if $\alpha_{\min} = 2$ ($EU^2 < \infty \Rightarrow P(U > u) = o(u^{-2})$). Since

$$T^{-1/\alpha_{\min}} O_p(1) Tt I(Tt \le S(0)) = O_p(T^{-(1/\alpha_{\min})-\alpha_{\min}+2}) \to 0 \quad \text{as} \quad T \to \infty,$$

this term will not contribute to the limit. Therefore,

$$\begin{aligned}
\mathcal{L}\lim_{T\to\infty} T^{-1/\alpha_{\min}} \int_0^{Tt} (W(u) - EW(u))du &= \mathcal{L}\lim_{T\to\infty} T^{-1/\alpha_{\min}} \sum_{k=1}^{K(Tt)} J(k) \\
&= \mathcal{L}\lim_{T\to\infty} \left(\frac{K(T)}{T}\right)^{1/\alpha_{\min}} (K(T))^{-1/\alpha_{\min}} \sum_{k=1}^{K(Tt)} J(k) \\
&= \mathcal{L}\lim_{T\to\infty} \left(\frac{1}{\mu_1 + \mu_2}\right)^{1/\alpha_{\min}} T^{-1/\alpha_{\min}} \sum_{k=1}^{[Tt]} J(k)
\end{aligned}$$

since $K(T)/T \to 1/\mu_1 + \mu_2$ almost surely and $K(Tt) \sim K(T)t \sim [K(T)t] \to \infty$ as $T \to \infty$, where [ ] denotes the integer part (see [3], Theorem 7.3.2). The summands $J(k)$ are given in (24) and are independent.

The conclusion now follows from standard results on limits of normalized sums of i.i.d. random variables:

$$\mathcal{L}\lim_{T\to\infty} T^{-1/\alpha_j} \sum_{k=1}^{[Tt]} U_j(k) = \ell_j^{1/\alpha_j} \Lambda^{(j)}_{\alpha_j,\sigma,1}(t),\ j = 1, 2, \quad \text{with } \sigma = C_{\alpha_j}^{-\alpha_j}$$

if $\alpha_j < 2$. (See [7], Theorem 2.6.1 and the proof of Theorem 2.2.2 for the convergence of the one-dimensional distributions). Convergence of the finite-dimensional distributions follows from the fact that the $U_j(k)$'s are i.i.d. Moreover, weak convergence holds in the Skorohod topology (see Skorohod [17]). In particular, if $\alpha = \alpha_1 = \alpha_2$, then the limit is

$$\frac{1}{(\mu_1 + \mu_2)^{1+1/\alpha}} \left[\mu_2 \ell_1^{1/\alpha_1} \Lambda^{(1)}_{\alpha,\sigma,1}(t) - \mu_1 \ell_2^{1/\alpha_2} \Lambda^{(2)}_{\alpha,\sigma,1}(t)\right]$$

where the Lévy stable motions $\Lambda^{(1)}_{\alpha,\sigma,1}$ and $\Lambda^{(2)}_{\alpha,\sigma,1}$ are independent and $C_\alpha \sigma^\alpha = 1$. These motions can be combined, yielding (7) with a skewness parameter $-1 \le \beta \le 1$ given by (8) (see Samorodnitsky and Taqqu [16], Section 1.2). The case $\alpha_1 = \alpha_2 = 2$ is straightforward.

5 An Application: Synthetic Generation of Self-Similar Traffic Traces

One of the main implications of the mathematical results presented in Section 2 is that even though today's network traffic is intrinsically complex in nature, *parsimonious modeling* is still possible. Moreover, the results give rise to a physical explanation of the observed self-similar nature of LAN traffic, that is shown in [22] to be fully consistent with actual traffic measurements from a LAN environment at the level of individual source-destination pairs. In fact, the desire for a "phenomenological" explanation of self-similarity in LAN traffic has resulted in new insights into the nature of traffic generated by the individual sources that contributed to the aggregate stream. We identified the Noah Effect of the *ON/OFF*-periods of the individual source-destination pairs as an essential ingredient, thus describing an important characteristic of the traffic in a "typical" LAN by essentially a single parameter, namely the intensity α of the Noah Effect in the *ON*- and *OFF*-periods of a "typical" network host. Whether we consider an idealized setting involving i.i.d. *ON*- and *OFF*-periods (see [21, Theorem 1]) or strictly alternating *ON/OFF* sources (see Section 2, Theorem 1) is not important for this finding. Generalizations accommodating more realistic conditions are possible (see Section 2), maintain the simplicity of the basic result, and may require the addition of only a small number of physically meaningful parameters.

Explaining, and hence modeling self-similar phenomena in network traffic in terms of the superposition of many *ON/OFF* sources with infinite variance distribution for the lengths of their *ON/OFF*-periods, leads to a straightforward method for generating long traces of self-similar traffic within reasonable (i.e., linear) time – assuming a parallel computing environment. Indeed, the results are tailor-made for parallel computing: letting every processor of a parallel machine generate traffic according to an alternating *ON/OFF* model (same α), simply adding (i.e., aggregating) the outputs over all processors produces self-similar traffic. For example, producing a synthetic trace of length 100,000 on a MasPar MP-1216, a massively parallel computer with 16,384 processors, takes on the order of a few minutes. In fact, Figure 1 shows the result of a simulation where we used this method to generate 27 hours worth of Ethernet-like traffic at the 10 millisecond time scale (i.e., a time series of approximately 10,000,000 observations). More precisely, our objective here was to experimentally "verify" our results in the context of the August 1989 Ethernet LAN traffic measurements considered in [12, 13]; i.e., we chose $\alpha = 1.2$ (corresponding to the estimated Hurst parameter of $H = 0.9$ that is consistent with the August 1989 data set), $M = 500$ (number of processors used to generate traffic, corresponding roughly to the number of active source-destination pairs during the observed period), and strictly alternating *ON/OFF* sources with the same α-value for the distributions of the *ON*- and *OFF*-periods. To check whether or not the resulting synthetic traffic trace "looks like" actual Ethernet LAN traffic as measured in August 1989, we plot in Figure 1 (right most column) the synthetic trace on 5 different time scales, the same way it was done in [13], the original traffic measurements (left most column), and a synthetic trace (middle column) generated from an appropriately matched batch Poisson process (the latter was taken as representative of traditional traffic modeling). As can be seen, our synthetic traffic passes the "visual" test easily, with the possible exception of the plot in the top row (the effect of the diurnal cycle in the 27 hour trace of Ethernet traffic on the 100s time scale becomes noticeable, especially because it is – by definition – not part of the stationary model that gave rise to the top right plot). On a more rigorous level, the trace also fits the data well in a statistical sense, i.e., the estimated Hurst parameter matches the one from the data and the marginals are approximately Gaussian. Similarly striking agreement between synthetically generated traffic and actual Ethernet LAN traces was obtained in a number of different scenarios, e.g., choosing $M = 16,000$ (close to the total number of processors on the MasPar machine), allowing for different source types (see Theorem 2), selecting different α-values for the *ON*-

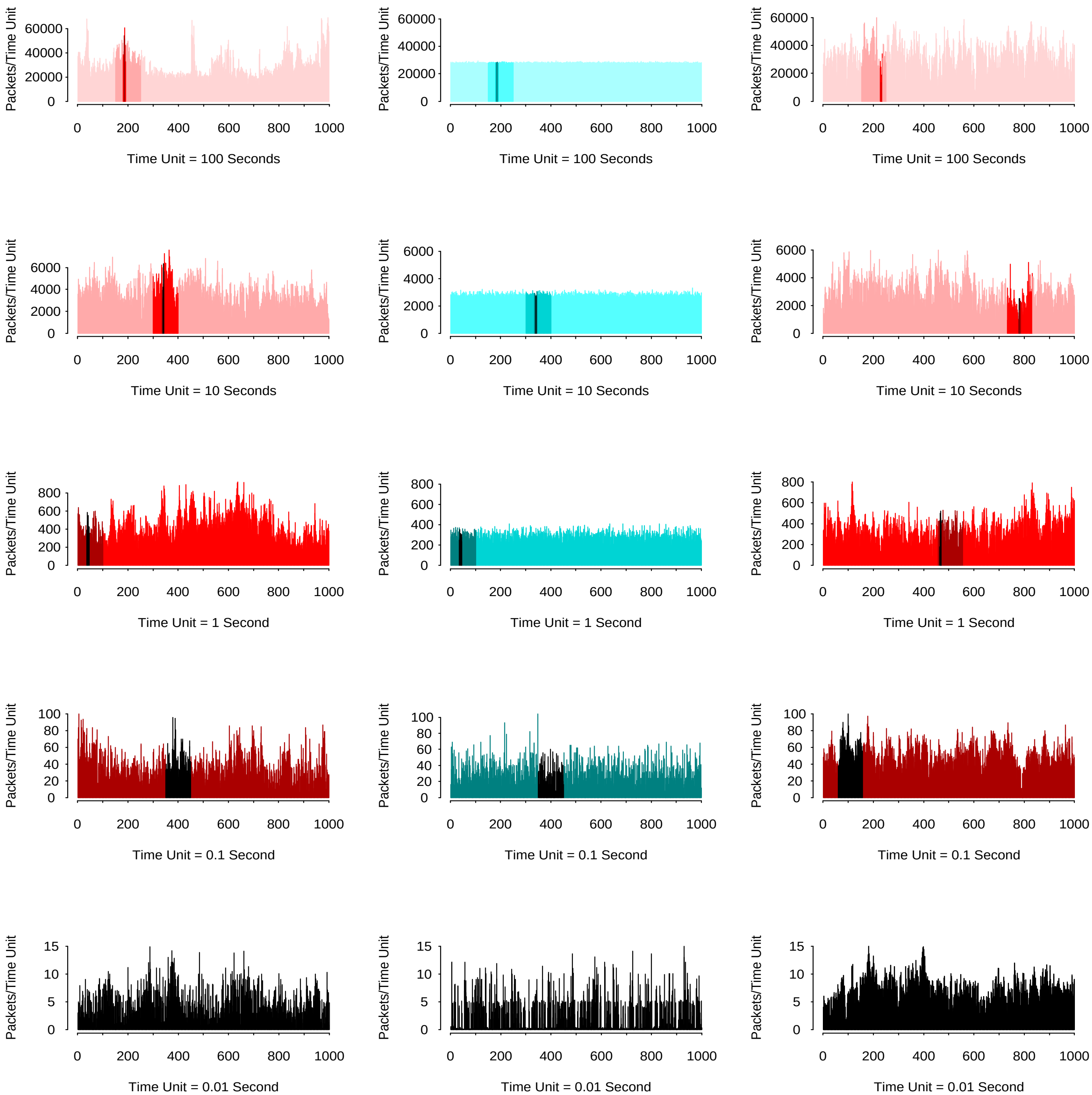

Figure 1: Actual Ethernet traffic (left column), synthetic trace generated from an appropriately chosen traditional traffic model (middle column), and synthetic trace generated from an appropriately chosen self-similar traffic model with a single parameter (right column) – on five different time scales. Different gray levels indicate the same segments of traffic on the different time scales.

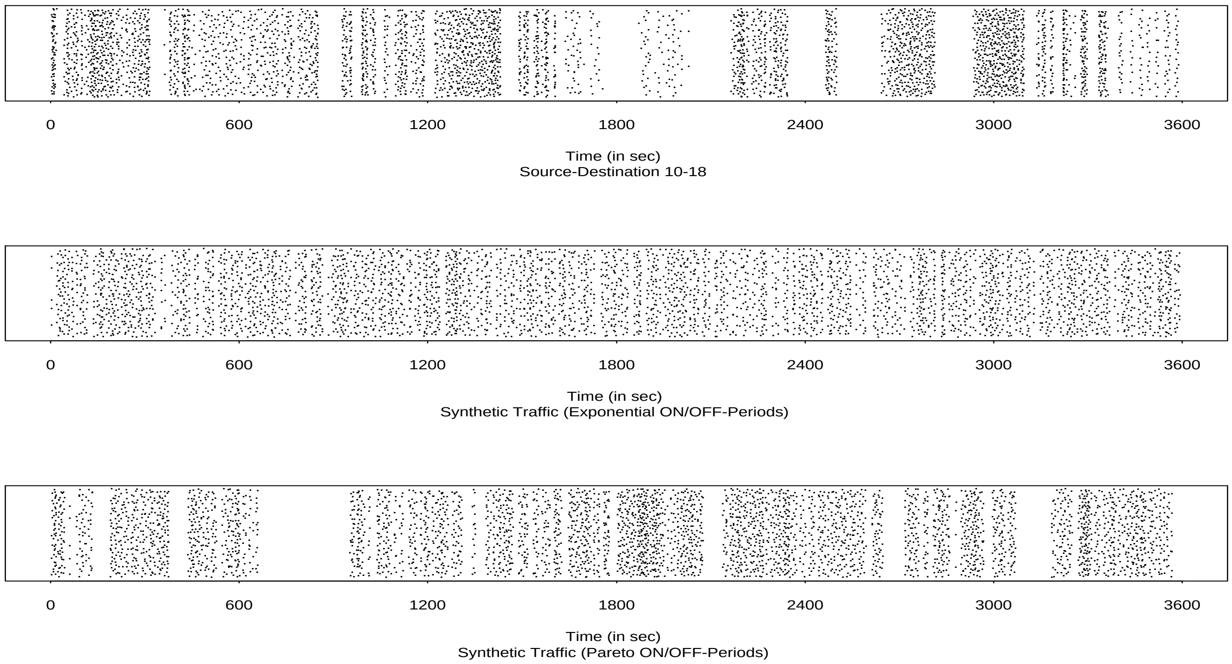

Figure 2: Textured plots of Ethernet traffic at the level of an individual source-destination pair: actual traffic (packet arrival times) from an active source-destination pair (top panel), synthetic trace generated from an appropriately chosen exponential *ON/OFF* source (middle panel), and synthetic trace generated from an appropriately chosen *ON/OFF* source with infinite variance *ON*- and *OFF*-periods (bottom panel).

and *OFF*-period distributions (including different combinations of finite/infinite variance scenarios), and generating under the i.i.d. and alternating renewal assumptions, respectively (see Section 2).

Recall that the Ethernet-like behavior of the synthetically generated trace in Figure 1 has essentially been accomplished *with only one parameter*, namely the intensity α of the Noah Effect for the *ON/OFF*-periods of the traffic generated by a "typical" user. To be precise, since we use a Pareto distribution for the *ON/OFF*-periods, we require an additional parameter indicating the lower cut-off of the Pareto distribution; yet another parameter governs the rate at which packets are generated during an *ON*-period; and last but not least, the number of sources, M, is also a parameter. While the cut-off and rate parameters are of minor importance, M plays, in general, a crucial role; however, in the present context, M was always chosen to be "large" (e.g., 500, 16 000). We leave the interesting and important problem of the quality of finite M-approximations to fractional Brownian motion or fractional Gaussian noise for further studies. In any case, the Ethernet-like trace in Figure 1 is testimony to parsimonious modeling at its best, and proof that today's complex network traffic dynamics can be modeled and described in a simple manner without requiring highly parameterized mathematical models. In fact, combining the insight gained from the new theoretical results presented in Section 2 with the practical benefits of modern high-performance computing environments, the method illustrated above enables us to quickly generate long traces of realistic network traffic by imitating on a small scale (using a multiprocessor environment) how traffic is generated on a large scale (i.e., in a real-life LAN, without accounting in detail for the possible effects that the different protocols exhibit on the traffic dynamics). To emphasize this point, Figure 2 shows (using textured plots) traffic traces generated by different types of "typical" *ON/OFF* sources: traffic generated by an actual source-destination pair during the busy hour of the August'89 data set (top panel), synthetic traffic generated by an *ON/OFF* source with exponential *ON*- and *OFF*-periods, with appropriately chosen means (matching the empirical means of the *ON*- and *OFF*-periods extracted from the trace generated by the above source-destination pair using a threshold value of $t = 1$s; see middle panel), and synthetic traffic generated by an *ON/OFF* source with Pareto *ON*- and *OFF*-periods with indices $\alpha_1 = 1.90$ and $\alpha_2 = 1.25$, respectively. A close look at Figure 2 reveals that exponential *ON*- and *OFF*-periods (middle panel) are unable to capture the highly variable nature of an actual *ON/OFF* source (top panel); while in the exponential case, the packets appear to arrive in a more or less uniform manner, the actual trace shows "gaps" (white areas) and "bursts" (shaded or dotted areas) of all sizes and intensities. Simply by replacing the exponential distributions for the *ON*- and *OFF*-periods by Pareto distributions results in a trace (bottom panel) that is practically indistinguishable from the real traffic trace. Note that the *ON*-periods in all three plots have the same mean, and similarly, the means of the *OFF*-periods in all three plots coincide. Figure 2 should be viewed as the source-destination equivalent of Figure 1, and they are closely tied together via the results in Section 2: aggregating over all actual source-destination pairs that were active during the August '89 measurement period results in actual aggregate Ethernet LAN traffic (Figure 1, left column), adding many exponential *ON/OFF* sources gives rise to the Poisson-like traffic dynamics displayed in Figure 1 (middle column), and superimposing many Pareto *ON/OFF* sources yields synthetic traffic (Figure 1, right column) that exhibits the same self-similar characteristics as measured Ethernet LAN traffic.

6 Conclusion

Motivated by the desire to provide a physical explanation for the empirically observed self-similarity property in actual network traffic, we proposed in [22] to expand the range of traditional traffic modeling at the level of individual sources (focusing mainly on the class of conventional *ON/OFF* source models with exponential or geometric *ON*- and *OFF*-periods) to account for the Noah Effect, i.e., for the ability of individual sources to exhibit characteristics that cover a wide range of time scales ("high-variability sources"). In this paper, we prove a fundamental result for self-similar traffic modeling, namely that the superposition of many *ON/OFF* models, each of which exhibits the Noah Effect, can yield aggregate packet streams that are consistent with measured LAN traffic and exhibits the same self-similar or fractal properties as can be observed in the data. Moreover, extensive statistical analyses in [22] confirm the presence of the Noah Effect in measured Ethernet LAN traffic at the source level, and demonstrate an appealing robustness property that renders objections against packet train source models (e.g., lack of a clear definition of a "train", lack of suggestions for choosing the crucial model parameters, and lack of a physical interpretation) irrelevant. The resulting new insights into the dynamics of actual LAN traffic is expected to facilitate the acceptance of self-similar traffic models as viable and practically relevant alternatives to traditional models. The benefits for doing so are immediate and include parsimonious and physically meaningful models for the seemingly very complex traffic dynamics in today's networks. As illustrated in Section 5, these physically-based models, in turn, give rise to novel and highly efficient algorithms for synthetically generating long traces of self-similar network traffic. Moreover, by combining the mathematical results proven in Section 2 with the capabilities of modern high-performance computing and communication, the resulting algorithm essentially imitates in the small (i.e., on a massively parallel machine) how traffic is generated in the large (i.e., in a LAN) – ignoring the possible impact of the network on the traffic sources (e.g., through the various protocols at the at the different network layers).

Acknowledgment

M. S. Taqqu and W. Willinger were partially supported by the NSF grant NCR-9404931. M. S. Taqqu was also partially supported by the NSF grant DMS-9404093. Part of this research was done when M. S. Taqqu was visiting Academia Sinica in Taipei. He would like to thank the Institute of Statistical Science and Dr. Hwai-Chung Ho for their hospitality. This extended version of our *ACM SIGCOMM'95* paper [21] has benefited from many helpful and specific suggestions by Vern Paxson and discussions with Joe Abate and Ward Whitt concerning the inversion problem of Laplace transforms.

References

[1] P. Billingsley. *Convergence of Probability Measures* J. Wiley & Sons, Inc., New York, 1968.

[2] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation.* Cambridge University Press, 1987.

[3] K. L. Chung. *A Course in Probability Theory.* Academic Press, New York, 2nd edition, 1974.

[4] D. R. Cox. *Renewal Theory.* Methuen & Co., London, 1967. Science Paperback Edition.

[5] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. Wiley, New York, 2nd edition, 1971.

[6] D. Heath, S. Resnick and G. Samorodnitsky. Heavy Tails and Long-Range Dependence in On/Off Processes and Associated Fluid Models. Preprint, School of ORIE, Cornell University, Ithaca, NY, 1996.

[7] I. A. Ibragimov and Yu. V. Linnik. *Independent and stationary sequences of random variables.* Wolters-Nordhoff, The Netherlands, 1971.

[8] R. Jain and S. A. Routhier. Packet Trains: Measurements and a New Model for Computer Network Traffic. *IEEE Journal on Selected Areas in Communications*, Vol. 4, pp. 986–995, 1986.

[9] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes.* Academic Press, New York, second edition, 1975.

[10] T. Konstantopoulos and S-J Lin. High variability versus long-range dependence for network performance. Preprint, 1996.

[11] K. R. Krishnan. The Hurst Parameter of non-Markovian On-Off Traffic Sources. Preprint, 1995.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic. *ACM/SIGCOMM Computer Communications Review*, Vol. 23, pp. 183–193, 1993. Proceedings of the ACM/SIGCOMM'93, San Francisco, September 1993.

[13] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, Vol. 2, pp. 1–15, 1994.

[14] S. B. Lowen and M. C. Teich. Fractal Renewal Processes Generate $1/f$ Noise. *Physical Review E*, 47:992–1001, 1993.

[15] B. B. Mandelbrot. Long-Run Linearity, Locally Gaussian Processes, H-Spectra and Infinite Variances. *International Economic Review*, Vol. 10, pp. 82–113, 1969.

[16] G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York, London, 1994.

[17] A. V. Skorohod. Limit theorems for stochastic processes with independent increments. *Theory of Probability and Its Applications*, 2:138–171, 1957.

[18] M. S. Taqqu. Weak Convergence to Fractional Brownian Motion and to the Rosenblatt Process. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31:287–302, 1975.

[19] M. S. Taqqu and J. Levy. Using Renewal Processes to Generate Long-Range Dependence and High Variability. In E. Eberlein and M. S. Taqqu, editors, *Dependence in Probability and Statistics*, pp. 73–89, Boston, 1986. Birkhäuser.

[20] J. L. Teugels. Renewal Theorems when the First or the Second Moment is Infinite. *The Annals of Mathematical Statistics*, 39, pp. 1210–1219, 1968.

[21] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *Computer Communications Review*, 25, pp. 100–113, 1995. Proceedings of the ACM/SIGCOMM'95, Cambridge, MA, August 1995.

[22] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1. pp. 1-16, 1997. Extended version of the preceding paper.