Multiscale Analysis for Wireless LAN Traffic Characterization

Jihwang Yeo, Ashok Agrawala
Department of Computer Science
University of Maryland
College Park, MD 20742
{jyeo,agrawala}@cs.umd.edu

CS-TR-4571 and UMIACS-TR-2004-16
March 1, 2004

***Abstract*— In this survey paper, we overview the various network traffic models, especially focusing on the multiscale analysis. By *multiscale analysis* we mean wavelet-based self-similar and multifractal analysis. Multiscale analysis is advantageous in that it can reveal the scaling behavior of the traffic on large time scale, at the same time characterize small-scale irregularity. We also discuss how we can apply this analysis technique to wireless LAN traffic characterization.**

I. Introduction

Historically numerous traffic models and analysis techniques have been developed for analyzing telecommunications networks [9], [10]. They, for example, include renewal models, Markov-based models, fluid models, autoregressive models, self-similar models, multiscale models and so on. Such traffic models provide ways to characterize the user, system and network behaviors, which can be observed and measured in real-world network. Furthermore with the characterized *workloads* one can predict the network performance through analytical techniques or simulation. The prediction results can contribute to the proper decisions on design and management of the networks.

Among those models, self-similar and multiscale models [43] have been paid more attentions than others over the decade. They can model the second or higher-order temporal dependence structure of actual traffics much better than traditional traffic models. Self-similar model can characterize the traffic in parsimonious manner with so-called Hurst parameter H. In other words, self-similar model can capture the *monofractal* nature of network traffic, asymptotically on large time scale.

Multiscale model [36], [41] overcomes the limitation of self-similar model, that self-similar model focuses on large time scale behaviors. Multiscale analysis can also capture small-scale temporal dependence , which self-similar model simply ignores. With wavelet tools [24], [25], [27] self-similar and multiscale models can be efficiently applied on network traffic data. Based on these two models, many analyses have been performed on various kinds of real-world network traffics. The examples of such traffics are video traffic [18], [20], *Local Area Network* (LAN) traffic [13], [15], *Wide Area Network* (WAN) traffic [16], [38] and *World Wide Web* (WWW) application traffic [17].

Recently the IEEE 802.11 wireless LANs [1] have been widely deployed and more and more mobile computers are being equipped with the IEEE 802.11 compatible wireless network devices. Wireless LAN traffics are typically affected by *non-ideal channel condition* [2], underlying MAC protocol and user mobility. Several measurement and analysis studies [3], [4], [5] examined traffic characteristics in the IEEE 802.11 wireless LAN. In these studies the measurements have been conducted on the *wired* portion of the network at TCP/IP level or above. Their analyses have focused on characterizing the *patterns* for usages and performance. Another experimental study in [8] has conducted similar measurements on wireless LAN traffic, but in more controlled and restricted scenarios. They have examined *asymptotic* self-similar properties of wireless LAN traffic on large time scale and found that channel quality degradation, measured separately from the traffic, reduces the degree of self-similarity. They argued that the decrease in the degree of the self-similarity was due to the buffering effect of the poor, slow link, which can make the self-similar bursts smoothed down.

In the previous wireless LAN studies described above, the measurements at TCP/IP level or above and at wired vantage points can hardly reveal the effects of the IEEE 802.11 MAC protocol and lossy wireless links on network traffics. Measurement at *wireless* vantage point, e.g. by placing so-called *wireless sniffers* in between wireless stations and Access Point (AP), is presented in [6], [7]. This *wireless* technique can provide per-frame PHY/MAC information therefore is capable to expose the effects of MAC and lossy channel more clearly [6], [7].

Such wireless measurement technique, combined with measurement at wired vantage points, can provide complete picture of wireless LAN traffic. In perspective of traffic analysis, such technique gives an opportunity for us to examine the effect of MAC and lossy channel on the wireless LAN traffic. In this paper assuming that such detailed PHY/MAC measurement data are available, we discuss what models and analysis techniques can reveal the effect of the IEEE 802.11 MAC protocol and lossy wireless links on network traffics, preferably on *wide* range of scales in very efficient manner.

We expect that wireless LAN traffic have similar statistical properties to Ethernet LAN on large time scale. Statistical behaviors in Ethernet LAN traffic have been well captured with self-similar models [13], [15]. On the other hand wireless LAN traffic may exhibit different properties on small time scale due to the IEEE 802.11 MAC protocol and lossy wireless links. Multiscale analysis has successfully explained small-scale behaviors in WAN traffic [38]. Therefore in this paper we argue that self-similar and multiscale analysis are the best candidates for analysis of wireless LAN traffics. To support this argument we survey existing network traffic models and specifically elaborate on self-similar models and multiscale analysis. We introduce the concept of wavelets and discuss their applications to self-similar and multiscale analysis. Finally we discuss how we can apply those models and techniques to the wireless traffic data.

This paper is organized as follows: In the next section we overview traditional traffic models. Then we discuss self-similar traffic models in Section III and their relevance in actual network traffics in Section IV respectively. We introduce wavelet analysis as a useful mathematical tool in Section V. Followingly we explain multiscale analysis in detail in Section VI. Finally we discuss our approaches to multiscale analysis on wireless LAN traffic in Section VII, then conclude this paper in Section VIII.

II. Overview of Stochastic Processes and Traffic Models

In this section we first summarize the basics of stochastic processes, then survey various stochastic traffic models. The models we describe in this section abstract the *first-order properties* (also called *marginal or time-independent* properties) and short-range dependence of a given traffic. For more detailed descriptions of stochastic processes and traffic modeling, readers are recommended to refer to [9], [10], [30]. Specifically the discussions on Sections (II-B)-(II-H) are based on [9].

A. Stochastic Process

A stochastic process $x(t)$ is a rule for assigning to every outcome of an experiment, a function $x(t)$. Therefore a stochastic process is a function of time and the experimental outcomes. If the domain of t is a set R of real numbers, then $x(t)$ is a *continuous-time* process [30]. If R is the set of integers, then $x(t)$ is a *discrete-time* process. Similarly $x(t)$ is a *discrete-state* process if its values are countable. Otherwise, it is a *continuous-state* process.

To describe the distribution of the stochastic process $x(t)$, *n-th-order distribution* of the process $x(t)$ is defined as follows:

$$F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = P\{x(t_1) \le x_1, x(t_2) \le x_2 \ldots, x(t_n) \le x_n\}.$$

The corresponding *n-th-order density* equals to

$$f(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

For many applications, to represent statistics of stochastic processes the expected value of $x(t)$ and of $x^2(t)$ are used. These time-dependent quantities can be expressed in terms of the *second-order properties* of $x(t)$ defined as follows.

The *mean* $\eta(t)$ of $x(t)$ is the expected value of the random variable $x(t)$:

$$\eta(t) = E\{x(t)\} = \int_{-\infty}^{\infty} x f(x,t) dx \tag{1}$$

The *autocorrelation* $r(t_1, t_2)$ of $x(t)$ is the expected value of the product $x(t_1)x(t_2)$:

$$\begin{aligned} r(t_1, t_2) &= E\{x(t_1)x(t_2)\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2; t_1, t_2) dx_1 dx_2 \end{aligned} \tag{2}$$

The *autocovariance* $C(t_1, t_2)$ of $x(t)$ is the covariance of the random variables $x(t_1)$ and $x(t_2)$:

$$C(t_1, t_2) = r(t_1, t_2) - \eta(t_1)\eta(t_2)$$

and its value $C(t,t)$ on the diagonal $t_1 = t_2 = t$ equals the variance of $x(t)$ [30].

Note that the second-order property is different from the *second-order distribution*. The second-order distribution of the process $x(t)$ is the joint distribution

$$F(x_1, x_2; t_1, t_2) = P\{x(t_1) \le x_1, x(t_2) \le x_2\}$$

of the random variables $x(t_1)$ and $x(t_2)$. The corresponding *second-order density* equals

$$f(x_1, x_2; t_1, t_2) = \frac{\partial^2 F(x_1, x_2; t_1, t_2)}{\partial x_1 \partial x_2}$$

Another important statistical property of stochastic process is the *stationarity*. A stochastic process $x(t)$ is called *strict-sense stationary* if its statistical properties are invariant to a shift of the origin. This means that the processes $x(t)$ and $x(t+c)$ have the same statistics for any c. It follows that *the n-th-order density* of a strict-sense stationary process must be such that

$$f(x_1, \ldots, x_n; t_1, \ldots, t_n) = f(x_1, \ldots, x_n; t_1 + c, \ldots, t_n + c).$$

A stochastic process $x(t)$ is called *wide-sense stationary* if its mean in (1) is constant and its autocorrelation in (2) depends only on $\tau = t_2 - t_1$ as follows [30]:

$$r(t, t+\tau) = r(\tau)$$

In this paper unless otherwise stated, by *stationary* we mean wide-sense stationary.

The *spectral density* or (*power spectrum*) of a wide-sense stationary process $x(t)$, real or complex, is the Fourier transform $S(\omega)$ of its autocorrelation $r(\tau) = E\{x(t+\tau)x^*(t)\}$:

$$S(\omega) = \int_{-\infty}^{\infty} r(\tau) e^{-j\omega\tau} d\tau$$

Since $r(-\tau) = r^*(\tau)$ it follows that $S(\omega)$ is a real function of ω. From the Fourier inversion formula, it follows that

$$r(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) e^{j\omega\tau} d\omega$$

If $x(t)$ is a real process, then $r(\tau)$ is real and even; hence $S(\omega)$ is also real and even. In this case,

$$S(\omega) = \int_{-\infty}^{\infty} r(\tau) cos\omega\tau d\tau = 2\int_{0}^{\infty} r(\tau) cos\omega\tau d\tau$$

B. Generic Traffic Model

A network traffic basically consists of an arrival of traffic entities (e.g. packets, cells, etc). Mathematically this arrival process is described as a *point process*. A point process $\{T_n\}$ consists of a sequence of arrival instants $T_1, T_2, \ldots, T_n, \ldots$, measured from the origin 0. Equivalently there are two additional processes, *counting process* $N(t)$ and *interarrival time process* $\{A_n\}$. A counting process $N(t)$ is a continuous-time, non-negative integer-valued stochastic process, where $N(t) = max\{n : T_n \leq t\}$ represents the number of arrivals in the interval $(0, t]$. An interarrival time process $\{A_n\}$ is a non-negative random sequence $\{A_n\}_{n=1}^{\infty}$, where $A_n = T_n - T_{n-1}$. The equivalence of these three processes follows from the equality of events:

$$\begin{aligned} \{N(t) = n\} &= \{T_n \leq t < T_{n+1}\} \\ &= \{\sum_{k=1}^{n} A_k \leq t < \sum_{k=1}^{n+1} A_k\} \end{aligned} \qquad (3)$$

Batch arrivals are also important in compound traffic; arrivals may consist of more than one unit at an arrival instant T. A non-negative random sequence $\{B_n\}_{n=1}^{\infty}$ describes batch arrivals, where B_n is the (random) number of units in the batch [9].

In addition to the arrival processes and batch sizes, the *workload* can be incorporated into traffic models. The workload is a general concept describing the amount of work $\{W_n\}$ brought to a system by the n-th arriving unit. The examples of workloads include service time requirements of arrivals at a queueing system and packet size (in bits or bytes) [9].

C. Burstiness

Burstiness is present in a traffic process if the arrival points $\{T_n\}$ appear to form visual clusters; equivalently $\{A_n\}$ tends to give rise to runs of several relatively short interarrival times followed by a relatively long one. Two main sources of burstiness are due to the shapes of the marginal distribution and autocorrelation function of $\{A_n\}$ [9].

Traffic burstiness can be measured from the marginal distribution of interarrival time. Examples of such measurements are the ratio of peak rate to mean rate and the ratio of standard deviation to mean of interarrival times, as $\sigma[A_n]/E[A_n]$. In contrast, the measurements like the index-of-dispersion for counts (IDC) and self-similarity take account of temporal dependence in traffic. For a given time interval of length τ, the IDC is the function $I_c(\tau) = Var[N(\tau)]/E[N(\tau)]$; i.e., the variance-to-mean ratio of the number of arrivals in the interval $[0, \tau]$. Since the number of arrivals is related to the sum of interarrival times as in (3), the numerator of the IDC includes the autocorrelations of $\{A_n\}$. Self-similarity test through so-called Hurst parameter will be discussed in detail in Section III.

D. Renewal Model

In a renewal traffic process, the $\{A_n\}$ are independent, identically distributed (IID), but their distribution is allowed to be general. Renewal processes, while simple analytically, have a severe modeling drawback - the autocorrelation function of $\{A_n\}$ vanishes identically for all nonzero lags. In other words, renewal processes cannot capture the temporal dependence structure of bursty traffic, which typically dominates broadband networks [9].

The most important example of renewal process is Poisson process. Poisson models are the oldest traffic models, whose interarrival times $\{A_n\}$ are exponentially distributed with rate parameter λ. Specifically the interarrival time process $\{A_n\}$ and the corresponding counting process $N(t)$ satisfy the following equations.

$$P\{A_n \leq t\} = 1 - e^{-\lambda t} \qquad (4)$$

$$P\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad (5)$$

Poisson processes have some elegant mathematical properties. First, the superposition of independent Poisson processes results in a new Poisson process whose rate is the sum of the component rates. Second, the independent increment property renders Poisson a memoryless process. Third, under suitable but mild regularity conditions, multiplexed streams approach a Poisson process as the number of streams grows, but the individual rates decrease so as to keep the aggregate rate constant. This theory is known as Palm's Theorem [31]. In real-world traffics e.g. Ethernet LAN, however, the aggregate traffic does not approach Poisson process, instead its burstiness typically intensifies as the number of active traffic sources increases [13].

Another special class of renewal processes is phase-type renewal process. Phase-type interarrival times can be modeled as the time to absorption in a continuous-time Markov process $C = \{C(t)\}_{t=0}^{\infty}$ with state space $\{0, 1, \ldots, m\}$. Here, state 0 is absorbing. To determine A_n, start the process C with some initial distribution π. When absorption occurs, i.e., when the process enters state 0, stop the process. The elapsed time is A_n, which is the sums of exponentials [9].

E. Markov and Markov Renewal Models

Markov and Markov-renewal traffic models introduce dependence into the random sequence $\{A_n\}$. Consequently, they can potentially capture traffic burstiness, because of nonzero autocorrelations in $\{A_n\}$ [9].

In continuous-time Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with a discrete state space, M stays in a state i for an exponentially distributed *holding time* with parameter λ_i; it then jumps to state j with probability p_{ij}. In a simple Markov traffic model, each jump of the Markov process is interpreted as signaling an arrival, so interarrival times are exponential, and their rate parameters depend on the state from where the jump occurred. This result in dependence among interarrival times as a consequence of the Markov properties [9].

Markov-renewal models are more general than discrete-state Markov processes. A Markov renewal process $R = \{(M_n, \tau_n)\}_{n=0}^{\infty}$ is defined by a Markov chain $\{M_n\}$ and its associated jump times $\{\tau_n\}$, subject to the following constraint: the pair (M_{n+1}, τ_{n+1}) of next state and inter-jump time depends only on the current state M_n, but not on previous states nor on previous inter-jump times. If we interpret jumps (transitions) of $\{M_n\}$ as signaling arrivals, then we would have dependence in the arrival process. Here the distribution of the interarrival time can be arbitrary and depends on both states at the start and end of each interarrival time [9].

F. Markov-Modulated Traffic Models

The idea of markov-modulated models is to introduce an explicit notion of state into the description of a traffic system. More precisely, in Markov-modulated traffic models a *state* does determine a probability law that governs a sequence of interarrival times generated in the state. A state also determines the duration the Markov process stays in the state [9].

Let $M = \{M(t)\}_{t=0}^{\infty}$ be a continuous-time Markov process, with state space $\{1, 2, \ldots, m\}$. Now assume that while M is in state k, the probability law of traffic arrivals is completely determined by k, and this holds for every $1 \leq k \leq m$. Note that when M undergoes a transition to, say, state j, then a new probability law for arrivals takes effect for the duration of state j, and so on. Thus the probability law for arrivals is *modulated* by the state M [9].

The most commonly used Markov-modulated model is the Markov-Modulated Poisson Process (MMPP) model, which combines the simplicity of the modulating (Markov) process with that of the modulated (Poisson) process. In state k of M, arrivals occur according to a Poisson process at rate λ_k. As the state changes, so does the rate [9].

As a simple example of MMPP model, consider a two-state MMPP model, where one state is an "on" state with an associated positive Poisson rate, and the other is an "off" state with associated rate zero. These models have been widely used to model voice traffic sources; the "on" state corresponds to a talk spurt, and the "off" state corresponds to a silence [9].

G. Fluid Traffic Models

The fluid traffic models view traffic as a stream of fluid, characterized by a flow rate (such as bits per second), so that a traffic count is replaced by a traffic volume. Fluid models are appropriate to cases where individual units are numerous relative to a chosen time scale. For example, in the context of ATM (Asynchronous Transfer Mode), all packets are fixed-size cells of relatively short length (53 bytes); in addition the high transmission speeds render the transmission impact of an individual cell negligible [9].

The fluid traffic models not only provide such conceptual simplicity, but also have a benefit of its efficiency as a simulation model. A fluid simulation assumes that the incoming fluid flow remains constant over long time periods, therefore the simulation events only occur at changes of flow rates. These changes can be assumed to happen far less frequently than individual cell arrivals, therefore the simulation can achieve a great saving in computing. Furthermore, the waiting time concept simply becomes the time it takes to serve the current buffer, and loss probabilities can be calculated in terms of overflow volumes [9].

Despite the analytical simplicity and the simulation efficiency, the fluid traffic models are only suitable for specific applications, e.g. ATM cell traffic. For the aims of investigating the traffics in small time scale, e.g. in milli- or micro-second, the fluid traffic models can hardly be applicable.

H. Autoregressive Traffic Models

Autoregressive models define the next random variable in the sequence as an explicit function of previous ones within a time window stretching from the present into the past. Such models are particularly suitable for modeling VBR (Variable Bit Rate)-coded video. The nature of video frames is such that successive frames within a video scene vary visually very little. Only scene changes can cause abrupt changes in frame bit rate. Thus, the sequence of bit rates (frame sizes) comprising a video scene may be modeled by an autoregressive scheme, while scene changes can be modeled by some modulating mechanism, such as a Markov chain [10].

The Autoregressive model of order p, denoted as $AR(p)$, has the following form:

$$\begin{aligned} X(t) &= \phi_1 X(t-1) + \phi_2 X(t-2) + \cdots \\ &+ \phi_p X(t-p) + \epsilon(t), \end{aligned} \qquad (6)$$

where $\epsilon(t)$ is white noise, ϕ_j's are real numbers, and $X(t)$'s are prescribed correlated random variables. If $\epsilon(t)$ is a white Gaussian noise with variance ${\sigma_{\epsilon(t)}}^2$, then $X(t)$'s will be normally distributed random variables. Let us define a lag operator B as $X(t-1) = BX(t)$, and let $\phi(B)$ be a polynomial in the operator B, defined as follows: $\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$. Then $AR(p)$ process can be represented as [10]:

$$\phi(B)X(t) = \epsilon(t). \qquad (7)$$

The autocorrelation $r(k)$ can be computed by multiplying (6) with $X(t-k)$, taking the expectation:

$$r(k) = \phi_1 r(k-1) + \phi_2 r(k-2) + \cdots + \phi_p r(k-p),$$

where $k > 0$. Thus, the general solution is

$$r(k) = A_1 {R_1}^k + A_2 {R_2}^k + \cdots + A_p {R_p}^k,$$

where ${R_i}^{-1}$s are the roots of $\phi(B)$. Therefore, the autocorrelation of $AR(p)$ process will consist, in general, of damped exponentials, and/or damped sine waves depending on whether the roots are real or imaginary [10].

The self-recursive structure of AR schemes make suitable for modeling autocorrelated traffic. There are many other autoregressive models, such as MA, ARMA and ARIMA [30]. Autoregressive models are typically used to fit the empirical autocorrelation function, but they cannot generally fit the empirical marginal distribution.

III. Overview of Self-Similar Traffic Models

In this section we overview self-similar and long-range dependence models. The concept of self-similarity had originated from Mandelbrot's seminal works [11], [12]. He applied this *fractal*-like concepts to such areas as hydrology and geophysics. In computer network modeling area, Leland *et al.* show that Ethernet LAN traffic is statistically self-similar in [13]. The causes of such self-similarity in Ethernet LAN traffic is plausibly explained in [15]. A number of researches have been performed for demonstrating the evidences of self-similarity in various network traffics, such as VBR-video traffic [14], WAN traffic [16] and WWW traffic [17].

In this section we first review the concept of self-similarity and its equivalent notion of long-range dependence. Then we describe two well-known analytic models (FGN and fARIMA) for self-similarity. We also discuss how we can infer the self-similar structure from the traffic data, for example by estimating so-called *Hurst parameter* H. Finally we discuss the results of self-similar analyses in various traffics.

A. Definition of Self-Similarity

Let $X = (X(t) : t = 0, 1, 2, \ldots)$ be a covariance stationary stochastic process with mean μ, variance σ^2 and autocorrelation function $r(k), k \geq 0$. In particular, we assume that X has an autocorrelation function of the form

$$r(k) \sim k^{-\beta} L(t), \text{ as } k \to \infty, \tag{8}$$

where $0 < \beta < 1$ and L is slowly varying at infinity, i.e., $\lim_{t\to\infty} L(tx)/L(t) = 1$, for all $x > 0$. For each $m = 1, 2, 3, \ldots$, let $X^{(m)} = (X_k^{(m)} : k = 1, 2, 3, \ldots)$ denote the new covariance stationary time series (with corresponding autocorrelation function $r^{(m)}$) obtained by averaging the original series X over non-overlapping blocks of size m. That is, for each $m = 1, 2, 3, \ldots$, $X^{(m)}$ is given by $X_k^{(m)} = 1/m(X_{km-m+1} + \cdots + X_{km}), k \geq 1$. The process X is called *(exactly) second-order self-similar* with self-similar parameter $H = 1 - \beta/2$ if for all $m = 1, 2, \ldots$, $var(X^{(m)}) = \sigma^2 m^{-\beta}$ and

$$r^{(m)}(k) = r(k), k \geq 0. \tag{9}$$

X is called *(asymptotically) second-order self-similar* with self-similarity parameter $H = 1 - \beta/2$ if for all k large enough,

$$r^{(m)}(k) \to r(k), \text{ as } m \to \infty \tag{10}$$

with $r(k)$ given by (8). In other words, X is exactly or asymptotically second-order self-similar if the corresponding aggregated process $X^{(m)}$ are the same as X or become indistinguishable from X—at least with respect to their autocorrelation functions [13].

Stochastic self-similar processes retain the same statistics over a range of scales, and they satisfy the following relation:

$$X(at) \equiv a^H X(t), \tag{11}$$

where $\equiv$ denotes equality in distribution and H is called the Hurst parameter. Therefore, the sample paths appear to be qualitatively the same, irrespective of the time scale.

B. Properties of Self-Similarity

Mathematically, self-similarity manifests itself in a number of equivalent ways: (i) the variance of the sample mean decreases more slowly than the reciprocal of the sample size *(slowly decaying variances)*, i.e., $var(X^{(m)}) \sim a_2 m^{-\beta}$, as $m \to \infty$, with $0 < \beta < 1$ (here and below, $a_2, a_3, \ldots$ denote finite positive constants); (ii) the autocorrelations decay hyperbolically rather than exponentially fast, implying a non-summable autocorrelation function $\sum_k r(k) = \infty$ *(long-range dependence)*, i.e., $r(k)$ satisfies relation (8); (iii) the spectral density $f(\cdot)$ obeys a power-law near the origin ($1/f$*-noise*), according to the following Fourier transform pair [13].

$$\frac{|\tau|^{\gamma-1}}{2\Gamma(\gamma)cos(\gamma\pi/2)} \leftrightarrow \frac{1}{|\omega|^{\gamma}}$$

i.e., $f(\omega) \sim a_3 \omega^{-\gamma}$, as $\omega \to 0$, with $0 < \gamma < 1$ and $\gamma = 1 - \beta$.

The existence of a nondegenerate correlation structure for the processes $X^{(m)}$, as $m \to \infty$, is in stark contrast to typical packet traffic models currently considered in the literature, all of which have the property that their aggregated processes $X^{(m)}$ tend to second-order pure noise, i.e., for all $k \geq 1$,

$$r^{(m)}(k) \to 0, \text{ as } m \to \infty \tag{12}$$

Equivalently, such *traditional* packet traffic models can be characterized by (i) a variance of the sample mean that decreases like the reciprocal of the sample mean, i.e., $var(X^{(m)}) \sim a_4 m^{-1}$, as $m \to \infty$; (ii) an autocorrelation function that decreases exponentially fast (i.e., $r(k) \sim \rho^k, 0 < \rho < 1$), implying a summable autocorrelation function $\sum_k r(k) < \infty$ (short-range dependence); (iii) a spectral density that is bounded at the origin [13].

Historically, the importance of self-similar processes lies in the fact that they provide an elegant explanation and interpretation of an empirical law that is commonly referred to the *Hurst effect*. Briefly, for a given set of observations $(X_k : k = 1, 2, \ldots, n)$ with sample mean $\overline{X}(n)$ and sample variance $S^2(n)$, the *rescaled adjusted range statistic* (or *R/S statistic*) is given by $R(n)/S(n) = 1/S(n)[max(0, W_1, W_2, \ldots, W_n) - min(0, W_1, W_2, \ldots, W_n)]$, with $W_k = (X_1 + X_2 + \cdots + X_k) - k\overline{X}(n)(k \geq 1)$. While many naturally occurring time series appear to be well represented by the relation $E[R(n)/S(n)] \sim a_5 n^H$, as $n \to \infty$, with *Hurst parameter H typically* about

0.7, observations X_k from a short-range dependent model are known to satisfy $E[R(n)/S(n)] \sim a_6 n^{0.5}$, as $n \to \infty$. This discrepancy is generally referred to as the *Hurst effect* [13].

C. Modeling of Self-Similar Phenomena

Fractional Gaussian noise is an example of an exactly self-similar process and *Fractional ARIMA* is an example of an asymptotically self-similar process.

1) Fractional Gaussian noise: Brownian motion is a stochastic process, denoted, $B(t)$, for $t \geq 0$. It is characterized by the following properties [10]:

- the increments $B(t+t_0) - B(t_0)$ are normally distributed with mean 0 and variance $\sigma^2 t$.
- the increments in non-overlapping time intervals $[t_1, t_2]$ and $[t_3, t_4]$, i.e., $B(t_4) - B(t_3)$ and $B(t_2) - B(t_1)$ are independent random variables.
- $B(0) = 0$ and $B(t)$ is continuous at $t = 0$.

The *fractional Brownian motion* $fB(t)$ is a Gaussian self-similar process with self-similarity parameter $H \in [0.5, 1)$. Frational Brownian motion differs from the Brownian motion by having increments with variance $\sigma^2 t^{2H}$. Define $\sigma^2 = E\{(fB(t) - fB(t-1))^2\} = E\{(fB(1) - fB(0))^2\}$ the variance of the increment process (Note that $fB_0 = 0$). Then:

$$E\{(fB(t_2) - fB(t_1))^2\} = E\{(fB_{t_2-t_1} - fB_0)^2\} = \sigma^2 (t_2 - t_1)^{2H}.$$

Also:

$$\begin{aligned} & E\{(fB(t_2) - fB(t_1))^2\} \\ = \ & E\{fB^2(t_2)\} + E\{fB^2(t_1)\} - 2E\{fB(t_2) fB(t_1)\} \\ = \ & \sigma^2 {t_2}^{2H} + \sigma^2 {t_1}^{2H} - 2r(fB(t_1), fB(t_2)), \end{aligned}$$

, where $r(X(t_1), X(t_2))$ is an autocorrelation between $X(t_1)$ and $X(t_2)$.

Therefore:

$$r(fB(t_1), fB(t_2)) = 1/2\sigma^2 ({t_2}^{2H} - (t_2 - t_1)^{2H} + {t_1}^{2H}).$$

Hence, the correlation of increments in two non-overlapping intervals is given by:

$$\begin{aligned} & r(fB(t_4) - fB(t_3), fB(t_2) - fB(t_1)) \qquad (13) \\ = \ & r(fB(t_4), fB(t_2)) - r(fB(t_4), fB(t_1)) \\ - \ & r(fB(t_3), fB(t_2)) + r(fB(t_3), fB(t_1)) \\ = \ & \sigma^2/2\{(t_4 - t_1)^{2H} - (t_3 - t_1)^{2H} + (t_3 - t_2)^{2H} \\ - \ & (t_4 - t_2)^{2H}\}. \end{aligned}$$

In the discrete case, the autocorrelation of the increment sequence is obtained by replacing t_1, t_2, t_3 and t_4 in (13) by $n, n+1, n+k$ and $n+k+1$ respectively:

$$r(k) = \sigma^2/2\{(k+1)^{2H} - 2(k)^{2H} + (k-1)^{2H}\}. \qquad (14)$$

The increment sequence is called *fractional Gaussian noise*. The autocorrelation in (14) exhibits long-range dependence, since $r(k) \sim k^{2H-2} = k^{-\beta}$ (follows by Taylor expansion) [10].

2) Fractional ARIMA: The fractional Autoregressive Integrated Moving Average process, F-ARIMA(p, d, q) with $0 < d < 1/2$, is an example of a stationary process with long-range dependence. It is an extension to ARIMA(p, d, q) and defined as [10]:

$$\phi(B)\nabla^d(B)X(t) = \theta(B)\epsilon(t),$$

where ϕ and B are as defined previously, and $\theta(B) = (1 - \theta_1 B - \cdots - \theta_q B^q)$. d can take the values between 0 and $1/2$. The operator $\nabla^d(B) = (1-B)^d$ can be expressed using the binomial expansion

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k B^k, \qquad (15)$$

$$\binom{d}{k} = \frac{d!}{k!(d-k)!} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}, \qquad (16)$$

where $\Gamma(x)$ denotes the gamma function. Note that for all positive integers, only the first $d+1$ terms are non-zero in (16). That is because the gamma function has poles for negative integers and hence the binomial coefficients are zero if $k > d$ and d is an integer. F-ARIMA$(0, d, 0)$ process with $0 < d < 1/2$ is stationary and long-range dependent, with an autocorrelation function

$$r(k) \sim \frac{\Gamma(1-d)}{\Gamma(d)} k^{2d-1} \text{ as } k \to \infty \qquad (17)$$

Observe that for $0 < d < 1/2$, the hyperbolic decay will produce persistence. By comparing (17) to (8), $d = (1 - \alpha)/2 = H - 0.5$. F-ARIMA processes can model short-range and long-range dependence. If Gaussian white noise is used, then the F-ARIMA has a Gaussian distribution. This limits the ability of F-ARIMA to model the processes that have an approximately Gaussian distributions. The Gaussian white noise is used because the sum of two Gaussian random variables is a Gaussian random variable. F-ARIMA was used to model VBR video traffic [14] [10].

D. Inference for Self-Similar Processes

To investigate the self-similarity in actual traffic data, we need the methods for testing and estimating the degree of self-similarity. There are various ways to estimate the degree of self-similarity, H (*Hurst parameter*): (1) analysis of the variances of the aggregated process $X^{(m)}$, (2) periodogram-based analysis in the frequency-domain, (3) time-domain analysis based on the R/S-statistics, (4) parameter estimation based on MLE (Maximum-likelihood Estimation), called *Whittle's method*, and (5) wavelet analysis in the frequency-domain. These methods are based on the self-similar models and properties, described in previous sections.

Among those methods, wavelet estimator has been known to outperform others because it is an unbiased method, robust to nonstationary trend, without any underlying self-similar models assumed [22], [23], [33]. By the *unbiased method*, we mean that the method is not affected by a specific setting of estimation, e.g. a selection of the frequency of analyzing window on which the target parameter is measured and estimated. We postpone the description of the wavelet estimator until we introduce the concept of wavelet in Section V. In this section we discuss the other four methods of estimating H.

1) Variance-Time plot: The variances of the aggregated processes $X(m), m \geq 1$, decrease linearly (for large m) in log-log plots against m with slopes arbitrarily flatter than -1. The so-called *variance-time plots* are obtained by plotting $log(var(X(m)))$ against $log(m)$ ("time") and by fitting a simple least squares line through the resulting points in the plain, ignoring the small values for m. Values of the estimate $\hat{\beta}$ of the asymptotic slope between -1 and 0 suggest self-similarity, and an estimate for the degree of self-similarity is given by $\hat{H} = 1 - \hat{\beta}/2$ [13].

2) Periodogram plot: As described in the previous section, the spectral density $f(\cdot)$ obeys a power-law near the origin. The *periodogram* $I(x)$ of $X = (X_1, X_2, \ldots, X_n)$ is obtained such as

$$I(x) = (2\pi n)^{-1} | \sum_{j=1}^{n} X_j e^{ijx} |^2, 0 \leq x \leq \pi. \qquad (18)$$

Values of the estimate $\hat{\gamma}$ of the asymptotic slope near ($x \rightarrow 0$ therefore, $log(x) \rightarrow -\infty$) suggest self-similarity. An estimate for the degree of self-similarity is given by $\hat{H} = (1 + \hat{\gamma})/2$.

3) R/S plot: The objective of the R/S analysis of an empirical record is to infer the degree of self-similarity H via Hurst effect. Graphical R/S analysis consists of taking logarithmically spaced values of n (starting with $n \approx 10$), and plotting $log(R(n)/S(n))$ versus $log(n)$ results in the *rescaled adjusted range plot* (also called the *pox diagram of R/S*). When H is well defined, a typical rescaled adjusted range plot starts with a transient zone representing the nature of short-range dependence in the sample, but eventually settles down and fluctuates in a straight line of a certain slope. Graphical R/S analysis is used to determine whether such asymptotic behavior appears supported by the data. In case the behavior exists, an estimate $\hat{H}$ of H is given by the line's asymptotic slope which can take any value between $1/2$ and 1. For practical purposes, the most useful and attractive feature of the R/S analysis is its relative robustness against changes of the marginal distribution. This feature allows for practically separate investigations of the self-similarity property of a given data set and of its distributional characteristics [13].

4) Whittle estimator: Maximum-likelihood Estimation (MLE), the best known fully parametric method, offers a coherent approach to estimator design which is capable of producing an unbiased, asymptotically efficient estimator for H (as well as for other parameters). The Whittle estimator consists of two analytic approximations to the exact Gaussian MLE, suggested by Whittle in 1953 in order to avoid the huge computational complexity of the exact algorithm. In the 1980's it was shown that nothing is lost in this approximation, in the sense that asymptotically the estimator is unbiased and efficient, just as in the exact case. The approximation essentially replaces the covariance matrix by an integral of a function of the spectrum. Still computational difficulties remain, motivating a further approximation: the discretization of the frequency-domain integration rewritten in terms of the periodogram (as in (18)). This discrete version is called *discrete Whittle estimator*. Being based on a parametric estimator, a specific family of process must be assumed, like fGN or fractional ARIMA. With Whittle estimator, we can obtain confidence intervals for the self-similarity parameter H [13].

E. Self-similarity in Ethernet LAN Traffic

The first rigorous analysis study on self-similarity of Ethernet Local Area Network was made in [13]. They applied H parameter estimation techniques, e.g. variance-time , periodogram, R/S plot and Whittle's method, on the well-known Ethernet traffic measurements collected between 1989 and 1992 at Bellcore laboratory. They found that (i) Ethernet LAN traffic is statistically self-similar (as H typically lies between 0.8 and 0.95), irrespective of when during the four-year data collection period and where they were collected in the network, (ii) the degree of self-similarity measured in terms of the Hurst parameter H is typically a function of the overall utilization of the Ethernet and can be used for measuring the "burstiness" of the traffic (the higher the load on the Ethernet the higher the estimated H). Especially as human-generated traffic increases, H becomes higher than that of machine-generated traffic. They also found that (iii) major components of Ethernet LAN traffic such as external LAN traffic or external TCP traffic share the same self-similar characteristics as the overall LAN traffic, and they concluded that (iv) the packet traffic models at that time considered in the literature are not capable to capture the self-similarity property and can therefore be clearly distinguished from their measured data [13].

Their following-up work [15] investigated the plausible physical explanation for the observed self-similar nature of the measured traffic. They show that the superposition of many (strictly alternating) independent and identically distributed (i.i.d.) *ON/OFF* sources, each of which exhibits a phenomenon called the "Noah Effect," results in self-similar aggregate traffic, also called "Joseph Effect". The Noah Effect is synonymous with the *infinite variance syndrome*. They use a *heavy-tailed* distribution of *ON/OFF* period x with infinite variance (e.g. Pareto) to account for the Noah Effect as the following:

$$F_c(x) \sim l x^{-\alpha} L(x), \text{ as } x \rightarrow \infty, 1 < \alpha < 2,$$

where α is called the intensity of Noah Effect ($H = (3-\alpha)/2$), $l > 0$ is a constant and $L(x) > 0$ is a slowly varying function at infinity. With such heavy-tailed *ON/OFF* period distribution, they mathematically show that superposition of strictly alternating *ON/OFF* produces aggregate self-similar process. Their statistical analysis of Ethernet LAN traffic traces, involving a few hundred active source-destination pairs, confirms that the data at the level of individual sources or source-destination pairs are consistent with the Noah Effect [15].

F. Self-similarity in WAN Traffic

Self-similarity in Wide Area Network was examined with 24 wide-area traces in [16]. They investigated a number of wide-area TCP arrival processes (session and connection arrivals, FTP data connection arrivals within FTP sessions, and TELNET packet arrivals) to determine the error introduced by modeling them using Poisson processes. They found that user-initiated TCP session arrivals (measured with TCP SYN/FIN), such as remote-login and file-transfer, are well-modeled as Poisson processes with fixed hourly rates. But they found that other connection arrivals deviate considerably from Poisson, being rather self-similar [16].

G. Self-similarity in WWW Traffic

Crovella *et al.* [17] showed evidence that World Wide Web traffic exhibits behavior that is consistent with self-similar traffic models. They also showed that transmission times may be heavy-tailed, primarily due to the distribution of Web file sizes. In addition they showed evidence that silent times also may be heavy-tailed, primarily due to the influence of user "think time". Their results included that the distribution of user requests is lighter-tailed than the set of available files; but that the action of caching serves to make the distribution of actual files transferred similar to the more heavy-tailed distribution of available files.

They argued that these results seem to trace the causes of Web traffic self-similarity back to basic characteristics of information organization and retrieval. They noted that heavy-tailed distribution is more like human or social phenomena, for example, the distribution of lengths of books on library shelves, and the distribution of word lengths in sample texts. Therefore these results suggest that the self-similarity of Web traffic is not a machine-induced artifact; in particular, changes in protocol processing and document display are not likely to fundamentally remove the self-similarity of Web traffic, as they concluded [17].

IV. Relevance of Long-Range Dependence Models

As discussions on self-similarity, or equivalently long-range dependence (LRD, abbreviated), became prevalent in research community, questions had been raised on the relevance of the model. The issues they brought up are three-fold: First, what are the implications of LRD models on actual network system performance, especially what are the impacts on the buffer behaviors? Second, how well do the synthetic LRD models, e.g. fractional Gaussian noise and fractional ARIMA, capture the correlation structures of actual traffic? Third, are the conventional LRD estimation methods, which are used for estimating Hurst parameter, robust to the data where LRD does not exist or nonstationarity dominates? In this section we overview various answers to those questions and discuss the implications to our works.

A. Implications of LRD Model for Practical System Performance

Heyman and Lakshman [18] examined how the distribution of $X^{(m)}$ (aggregated time series with block size m, see Section III) determines the buffer occupancy of variable bit rate(VBR)-video sources. From this model, they show that long-range dependence does not affect the buffer occupancy when the busy periods are not large. They identified a *reset effect* that shows that only those averages taken within a busy period functionally affect the buffer size. Another effect, the *truncating effect* of finite buffers, enhances the reset effect. They used Markov chain models, which can capture short-range dependence, making excellent estimates of cell-loss rates mean buffer sizes, especially when buffer is not too large.

Grossglauser and Bolot [19] argued that self-similar modeling has failed to consider the impact of two important parameters, namely the finite range of time scales of interest in performance evaluation and prediction problems, and the first-order statistics such as the marginal distribution of the process. They used a modulated fluid model which can control such time scales of interest, called *cutoff lags*, to examine the fluid loss rate in terms of marginal distribution, Hurst parameter, cutoff lags and buffer sizes. Their findings are as the following: First they find that the amount of correlation that needs to be taken into account for performance evaluation depends not only on the correlation structure of the source traffic, but also on time scales specific to the system under study. Thus, for finite buffer queues, they find that the impact on loss of the correlation in the arrival process becomes *nil* beyond the time scale, they refer to as the *correlation horizon*. Second, they find that loss can depend in a crucial way on the marginal distribution of the fluid rate process. Third, the results suggest that reducing loss by buffering is hard for traffic with correlation over many time scales.

Another related study by Ryu and Elwalid [20] investigated the practical implications of LRD in the context of realistic ATM traffic engineering by studying ATM multiplexers of VBR video sources over a range of desirable cell loss rates and buffer sizes. They reached the similar conclusion that even in the presence of LRD, long-term correlations do not have significant impact on the cell loss rate beyond some time scale, they called *Critical Time Scale*. They also argued that short-range correlations have dominant effect on cell loss rate, and therefore, well-designed Markov traffic models are effective for predicting Quality of Services (QoS) of LRD VBR-video traffic.

B. Validity of LRD Model on Real-World Traffics

Researches described in the previous section argued that Markovian models are enough to predict the buffer behaviors because the correlation structure does not impact the performance beyond some buffer-size dependent finite time scale. However, Krunz and Makowski [21] argued that empirically the autocorrelation function $r(k)$ is better captured by $r(k) = e^{-\beta\sqrt{k}}$ than by $r(k) = k^{-\beta} = e^{\beta log k}$ (long-range dependence) or $r(k) = e^{-\beta k}$ (Markovian), therefore a *third* approach is necessary which can incorporate the autocorrelation in such form as well as those of both LRD and SRD.

They introduced a video model based on so-called $M/G/\infty$ input processes. The $M/G/\infty$ process is a stationary version of the busy-server process of a discrete-time $M/G/\infty$ queue, where by varying G, many forms of time dependence can be displayed. They derived the appropriate G that gives the desired correlation function $r(k) = e^{-\beta\sqrt{k}}$. They argued that this model is shown to exhibit short-range dependence, but capture both short-range and long-range correlations, hence combining the goodness of Markovian models at small lags with that of LRD models at large lags [21].

C. Reliability of Hurst Parameter Estimation

In Section III, we introduced several methods for estimating Hurst parameter H. In estimating H, the test methods may in itself contain *biases* that may result in wrong conclusions and unreliable estimation. There are several categories of biases that may exist in inference methods for estimating Hurst parameter. One is the bias sensitive to the specific settings of the measurement, e.g. the frequency length of analyzing window, etc. The second category is the bias which may misinterpret the *unexpected* nonstationary trends or short-range dependence as long-range dependence. In this section, we examine this problem in detail and discuss what is the best estimator in terms of such biases.

Molnar and Dang [22] conducted analytical and simulation study, showing that the presence of different nonstationary effects (level shifts, linear and polynomial trends) in the data can deceive several LRD tests. In their results, variance-time plot and the periodogram lead to a poor estimate of the Hurst parameter. Moreover, the estimated results can be confused with the results of the processes having short-range dependence with nonstationary effects; similar result was reported by Krunz and Matta [23] that a variance-type estimator often indicates, falsely, the existence of an LRD structure (i.e. $H > 0.5$) in synthetically generated traces from the two SRD models. The R/S analysis can reveal the presence of the level shifts, but it would mislead the result without removing points caused by the level shifts. The Wavelet-based method provides a very robust estimation of H in the presence of level shifts or trends. They recommended wavelet-based method for the estimation of Hurst parameter of LRD processes in the possible presence of nonstationary trends.

Abry and Veitch [32], [33] introduced a wavelet-based tool for the analysis of long-range dependence and a related semi-parametric estimator of the Hurst parameter. The estimator is based on time-scaled spectral estimation of $1/f$ noise process. As described in Section III-B, spectral density of LRD process indicates $1/f$ noise in the following form

$$S_x(\nu) \sim c_f|\nu|^{\gamma} = c_f|\nu|^{1-2H}, \text{ as } \nu \to 0 \tag{19}$$

where

$$c_f = \frac{c_\gamma \Gamma(2H-1) sin(\pi - \pi H)}{\pi}$$

and Γ is the Gamma function. As shown in [32], if an $1/f$ process has the spectral density of $\sigma^2|\nu|^{-\gamma}$, then the *estimated* spectral density based on time-frequency analysis reads:

$$\hat{S}_x(\nu) = \sigma^2|\nu|^{-\gamma}\int_{-\infty}^{\infty}|1+\omega/\nu|^{-\gamma}\tilde{\pi}(0,\omega)d\omega$$

where $\tilde{\pi}(\zeta,\omega)$ is the Fourier transform of an arbitrary 2D function $\Pi(\tau,\omega)$ in terms of τ. Here, the estimated spectrum is generally affected by a (multiplicative) bias term, which depends on estimating frequency ν, preventing us from easily estimating the exponent γ. This means that the conventional periodogram method (refer to Section III-D) suffers this kind of bias.

On the contrary, the estimated spectral density based on time-scale analysis reads:

$$\hat{S}_x(\nu) = \sigma^2|\nu|^{-\gamma}\int_{-\infty}^{\infty}|\omega|^{-\gamma}\tilde{\pi}(0,\omega)d\omega.$$

Here, the multiplicative bias on the spectrum is no longer frequency-dependent, and it follows that the spectral exponent γ can be estimated without bias from a linear fitting of the data $\hat{S}_x(\nu)$ versus frequency ν in log-log plot. Exploiting this advantage of time-scaled spectral analysis, Abry and Veitch developed wavelet-based estimator [33]. We will present a detailed procedure for calculating wavelet-based estimator in the following section.

In this section, we overviewed the issues on the relevance of LRD models. The performance related results show that in actual VBR-video traffics LRD does not impact on buffer occupancy in a crucial manner, therefore buffer behaviors can be sufficiently well modeled with Markovian or $M/G/\infty$ queue. Those results are quite appealing but it is questionable whether the results can be generalized to the cases of other traffics than VBR-video traffic. Another result that conventional LRD estimator suffers serious biases indicates that we should carefully choose the estimator to obtain more reliable results. Wavelet-based estimator, as recommended in many researches, is bias-free and robust even when LRD exists mixed with SRD or nonstationarity. More importantly wavelet-based tools enable us to investigate multiscale analysis, which we explain in Section VI.

V. Wavelet Analysis

In this section, we introduce the concept of wavelets and wavelet analysis, then explain wavelet-based Hurst parameter estimator as an application of the wavelet. For more readings on introduction of the wavelets, readers are recommended to refer to [24], [27]

A. What is Wavelet?

Wavelet is a mathematical tool for representing signals as the sum of "small waves" (so it is called *wavelet*). Wavelet can be thought of as a better substitute or extension of the *Fourier transform*. The Fourier transform is a mathematical procedure that breaks up a function (or a given signal) into the frequencies that compose it, represents it as a weighted sum of periodic sine and cosine functions. With $e^{i\theta} = cos\theta + isin\theta$, the Fourier transform $F(x)$ of a given function $f(x)$ is obtained from $f(x)$ and reversely $f(x)$ can be reconstructed from $F(x)$ by the following equations:

$$f(x) = \int_{-\infty}^{\infty}F(k)e^{2\pi ikx}dk \leftrightarrow F(x) = \int_{-\infty}^{\infty}f(x)e^{-2\pi ikx}dx.$$

The Fourier transform basically aims to find coefficients of sine and cosine functions mainly on *frequency domain*. However Fourier analysis is poorly suited to very brief signals, or signals that change suddenly and unpredictably. This is because in the Fourier transform a local characteristic of the signal becomes a global characteristic of the transform. A discontinuity, for example, is represented by a superposition of all possible frequencies. In addition the lack of time information makes a Fourier transform terribly vulnerable to errors [24].

To analyze a signal in both time and frequency, a new method, called *windowed Fourier transform* was introduced. The idea is to study the frequencies of a signal segment by segment. The "window" that defines the size of the segment to be analyzed. While the classical Fourier transform compares the entire signal successively to infinite sine and cosines of different frequencies, windowed (or *short-time*) Fourier analysis compares a segment of the signal to bits of oscillating curves, first of one frequency, then of another, and so on. However, this method also has the disadvantages; the smaller the window the better we can locate sudden changes, such as peaks or discontinuity, but at the same time the more we lose the lower frequency of the given signal [24].

The wavelets overcome such drawbacks of the Fourier transforms. A wavelet contains fixed number of oscillations, which can be *dilated* (which, in wavelet context, means either squeezed or stretched) to analyze various (low or high) frequencies of a given signal. Moreover, wavelets can be *translated* (shifted) to represent various intervals on time domain. In other words, the wavelets enable us to analyze the signal on *both* frequency and time domain. Therefore, the wavelets are suited for analysis of highly non-stationary signals with sudden peaks or discontinuities, without losing any lower frequency information [24].

B. Admissibility of Wavelet

We have been introduced the general concept of wavelets in the previous section. But what condition should a function satisfy to be used as the wavelets? The *admissibility condition* of the wavelets is expressed by [28]:

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty, \tag{20}$$

where Ψ is the Fourier transform of a wavelet function ψ. This inequality in (20) implies two things: First, the transform is invertible. That is, all square integrable functions that satisfy the admissibility condition can be used to analyze and reconstruct any signal. Second, the wavelet function ψ must have a value of zero at zero frequency. In other words, the wavelet must be an oscillatory signal, where the positive and negative values cancel each other.

C. Continuous Wavelet Transform

In the continuous wavelet transform, a function ψ, which in practice looks like a little wave, is used to create a family of wavelets $\psi(at+b)$ where a and b are real numbers, a dilating (compressing or stretching) the function ψ and b translating (displacing) it [24].

The continuous wavelet transform turns a signal $f(t)$ into a function with two variables (scale and time), which one can call $c(a,b)$:

$$c(a,b) = \int_{-\infty}^{\infty} f(t)\psi(at+b)dt.$$

This transformation is in theory infinitely redundant, but it can be useful in recognizing certain characteristics of a signal. In addition, the extreme redundancy is less of a problem than one might imagine; a number of researchers have found the ways of rapidly extracting the essential information from these redundant transforms [24].

While the continuous wavelet transform may be beneficial sometimes, its redundancy incurs inefficiency and cost anyway. In a *discrete wavelet transform* a wavelet is translated and dilated only by discrete values. Most often dilation is by a power of 2 (thus called *dyadic* dilation). That is, one uses wavelets only of the form:

$$\psi(2^k t + l), \text{ with } k \text{ and } l \text{whole numbers}$$

Orthogonal wavelets (see the next section, *multiresolution*) are special cases of discrete wavelets. They give a representation without redundancy and lend themselves to fast algorithms [24].

D. Multiresolution Analysis (MRA)

A multiresolution analysis (MRA) consists in a collection of nested subspaces $\{V_j\}_{j\in\mathcal{Z}}$, satisfying the following set of properties [25], [24], [27], [33], [38]:

i) $\bigcap_{j\in\mathcal{Z}} V_j = \{0\}$, $\bigcup_{j\in\mathcal{Z}} V_j$ is dense in $L^2(\mathcal{R})$

ii) $V_j \subset V_{j-1}$

iii) $x(t) \in V_j \longleftrightarrow x(2^j t) \in V_0$

iv) There exists a function $\phi(t)$ in V_0, called the *scaling function* such that the collection $\{\phi(t-k), k \in \mathcal{Z}\}$ is an unconditional Riesz basis for V_0,

where $L^2(\mathcal{R})$ is a set of square-integrable functions over $\mathcal{R}$ and the subset of A of X is *dense* if its set closure $cl(A) = X$. Readers can refer to Appendix A to see the definition of Riesz basis [33].

Similarly, the scaled and shifted functions

$$\{\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k), k \in \mathcal{Z}\}$$

constitute a Riesz basis for the space V_j. Performing a multiresolution analysis of the signal x means successively projecting it into each of the approximation subspaces V_j

$$approx_j(t) = (Proj_{V_j} x)(t) = \sum_k a_x(j,k)\phi_{j,k}(t)$$

Since $V_j \subset V_{j-1}$, $approx_j$ is a coarser approximation of x than is $approx_{j-1}$ and, therefore, the key idea of the MRA consists in examining the loss of information, that is, the detail, when going from one approximation to the next, coarser one $detail_j(t) = approx_{j-1}(t) - approx_j(t)$. The MRA analysis shows that the detail signals $detail_j$ can be directly obtained from projections of x onto a collection of subspaces, the W_j, called the wavelet subspaces. Moreover, the MRA theory shows that there exists a function ψ, called the mother wavelet or wavelet, to be derived from ϕ, such that its templates $\{\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t-k), k \in \mathcal{Z}\}$ constitute a Riesz basis for W_j

$$detail_j(t) = (Proj_{W_j}x)(t) = \sum_k d_x(j,k)\psi_{j,k}(t)$$

Basically, the MRA consists in rewriting the information in x as a collection of details at different resolutions and a low-resolution approximation

$$\begin{aligned} x(t) &= approx_J(t) + \sum_{j\leq J}^{j=J} detail_j(t) \\ &= \sum_k a_x(J,k)\phi_{J,k}(t) \\ &+ \sum_{j\leq J}^{J}\sum_k d_x(j,k)\psi_{j,k}(t) \end{aligned} \tag{21}$$

The representation in (21) is called the *wavelet decomposition* of the signal X, and $d_x(j,k) = \langle X, \psi_{j,k}\rangle$ is commonly referred to as the *wavelet coefficient* at scale j and time $2^j k$. The quantity $|d_x(j,k)|^2$ measures the amount of energy in the signal X about the time $t_0 = 2^j k$ and about the frequency $2^{-j}\lambda_0$, where λ_0 is a reference frequency which depends on the wavelet ψ. The set of all wavelet coefficients $\{d_x(j,k) : j \in \mathcal{Z}, k \in \mathcal{Z}\}$ is called the *discrete wavelet transform* (DWT) of the signal X and its key feature is that it contains the same information as the signal X; i.e., it allows us to reconstruct X completely from its wavelet coefficients by setting $X(t) = \sum_{j\in\mathcal{Z}}\sum_{k\in\mathcal{Z}} d_x(j,k)\psi_{j,k}(t)$ [33].

Intuitively, the discrete wavelet transform divides a signal into different frequency components and analyzes each component with a resolution matched to its scale. We can use the wavelet coefficients to study directly either scale- or time-dependent properties of a given signal X. For example, by fixing a given scale j and studying X at that scale across time, we can obtain information about the scaling behavior of X, as a function of j. On the other hand, fixing a point t_0 in tie and investigating the wavelet coefficients across finer and finer scales results in powerful techniques for investigating the nature of local irregularities or singularities in the signal, as a function of t_0. While the former method results in scaling properties that hold globally (across the whole signal), the latter technique captures the idea behind the notion of "the wavelet transform as a mathematical microscope", provides (local) information about the fine structure of the signal at a given point in time, and thus opens up new ways for studying the intrinsic nature of "bursts" in measured network traffic [38].

E. Wavelets and Filters

With MRA property (iv) in the previous section and $W_j \oplus V_j = V_{j-1}$, scaling function $\phi(t)$ and wavelet $\psi(t)$ can be a combination of the translates of $\phi(2t-k)$, which is at a resolution twice as fine.

$$\begin{aligned} \phi(t) &= \sum_{k=0}^{\infty} h(k)\phi(2t-k), \text{ (MRA property (iv))} \\ \psi(t) &= \sum_{k=0}^{\infty} g(k)\phi(2t-k), (W_j \subset V_{j-1}) \end{aligned}$$

where h is a low-pass filter and g is the conjugate mirror filter of h, i.e., $g(k)$ are identical to $h(k)$ but in reverse order and with alternating signs [24].

We can think of the Fourier Series of $h(k)$ to be $A(\omega)$:

$$A(\omega) = \sum_{k=0}^{\infty} h(k)e^{2\pi ik\omega}.$$

$A(\omega)$ satisfies the following condition for low-pass filter:

$$A(0) = 1 \text{ and } |A(\omega)|^2 + |A(\omega + 1/2)|^2 = 1.$$

Similarly, the Fourier Series $D(\omega)$ of $g(k)$ can be defined as:

$$D(\omega) = \sum_{k=0}^{\infty} g(k)e^{2\pi ik\omega},$$

and satisfying the condition for high-pass filter:

$$D(0) = 0 \text{ and } |D(\omega)|^2 + |D(\omega + 1/2)|^2 = 1.$$

Conversely, from A and D we can generate Φ (the Fourier Transform of ϕ) and Ψ (the Fourier Transform of ψ) [24]:

$$\begin{aligned} \Phi(\omega) &= \Pi_{j=1}^{\infty} A(\frac{\omega}{2^j}) \\ \Psi(\omega) &= D(\frac{\omega}{2})\Pi_{j=2}^{\infty} A(\frac{\omega}{2^j}). \end{aligned}$$

F. Properties of Wavelets

For the successful application of wavelets, choosing a proper wavelet is important. To determine "how proper" a wavelet is, we need to know the properties of wavelets, such as *compact support* and *vanishing moments*. If a wavelet is with compact support, wherever the function is not defined, it will have a value of zero. Therefore, compact support represents the locality of the wavelet in the time domain. Compact support provides the advantages in numerical calculation of numerous wavelet coefficients. Another important property of wavelets is *regularity*. Regularity roughly means how the wavelet function ψ is locally smooth and concentrated in both the time and frequency domains. Such regularity is related with the differentiability of the wavelet, which is quantitatively defined with a *vanishing moment* M. A vanishing moment M of the wavelet ψ is defined by [24]:

$$\int x^m \Psi(x)dx = 0, m \in [0, M],$$

where Ψ is a Fourier transform of ψ.

Vanishing moment simply means how many differentiable ψ is. For example, a three-times differentiable ψ must have three vanishing moments. As the vanishing moment M increases:

- the corresponding filter becomes to have more non-zero values, i.e., non-zero $h(k)$ or $g(k)$,
- the wavelet can see more; e.g. the wavelet with $M = 1$ does not see linear functions, the wavelet with $M = 2$ is blind to quadratic functions, and that with three is blind to cubic functions, and so on.
- small number of coefficients are needed for the same quality of the approximation, therefore more useful for compression as well as for analyzing signals with singularities and discontinuities.

But how many vanishing moments are desirable depends on the application [24].

G. Fast Wavelet Transform

There exist many algorithms to calculate wavelet coefficients efficiently. In this section, we introduce one of those algorithms, called *the pyramidal algorithm* ([26], [24], [29]). The pyramidal algorithm calculates the wavelet coefficients the wavelet coefficients for any number of scales using octave filter banks given the initial approximation coefficients. Therefore, the detail coefficients need only be computed at the initial scale. The detail coefficients at higher (larger) scales are computed from these initial coefficients via the pyramidal algorithm, which uses only the approximate coefficients of the preceding scale for calculating the detail coefficients of the next scale.

Let us denote the projection of a function $f \in L^2(\mathrm{R})$ onto V_j and W_j respectively by

$$a_f(j,k) =< f, \phi_{j,k} > \text{ and } d_f(j,k) =< f, \psi_{j,k} >, k \in \mathrm{Z},$$

where $< \cdot, \cdot >$ denotes the standard L^2 inner product. The pyramidal algorithm calculates these coefficients efficiently with a cascade of discrete convolutions and subsamplings. Denote time reversal by $\bar{x}(n) = x(-n)$ and upsampling by

$$\breve{x}(n) = \begin{cases} x(n) & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

The pyramidal algorithm is then given by the following theorem:

Theorem 1: (Mallat [26]) Decomposition:

$$\begin{aligned} a_{j+1,p} &= \sum_{n=-\infty}^{\infty} h(n-2p)a_{j,n} = a_j \star \bar{h}(2p) \\ d_{j+1,p} &= \sum_{n=-\infty}^{\infty} g(n-2p)a_{j,n} = a_j \star \bar{g}(2p) \end{aligned}$$

Reconstruction:

$$\begin{aligned} a_{j,p} &= \sum_{n=-\infty}^{\infty} h(p-2n)a_{j+1,n} + \sum_{n=-\infty}^{\infty} g(p-2n)d_{j+1,n} \\ &= \breve{a}_{j+1} \star h(p) + \breve{d}_{j+1} \star g(p) \end{aligned}$$

To compute the detail coefficient at scale j, we use only the approximation coefficient at the previous scale a_{j-1}. Note that the domain of h and g are compact if we use Daubechies wavelets with compact supports. Computing a fast wavelet transform of a signal with n points takes about $2cn$ computations, where c depends on the wavelet used (which must have compact support) [26], [29].

H. Wavelet-based H Estimator

We introduce in this section the wavelet-based H-estimator [33]. The wavelet-based estimator is known to be bias-free and robust when no LRD exists or nonstationary trend appears.

1) Definition of the Estimator: The coefficient $|d_X(j,k)|^2$ measures the amount of energy in the analyzed signal about the time instant $2^j k$ and frequency $2^{-j}\nu_0$, where ν_0 is an arbitrary reference frequency selected by the choice of ψ. It has been suggested [32] a useful spectral estimator can be designed by performing a time average of the $|d_x(j,k)|^2$ at a given scale, that is,

$$\hat{S}_x(2^{-j}\nu_0) = \frac{1}{n_j}\sum_k |d_x(j,k)|^2$$

where n_j is the available number of wavelet coefficients at octave j. Essentially $n_j = 2^{-j}n$ where n is the length of the data. $\hat{S}_x(\nu)$ is therefore a measure of the amount of energy that lies within a given bandwidth around the frequency ν and can therefore be regarded as a statistical estimator for the spectrum $S_x(\nu)$ of x. In fact, one can show [32] that when x is wide-sense-stationary process, then the expectation of $\hat{S}$ is

$$\mathrm{E}[\hat{S}_x(2^{-j}\nu_0)] = \int S_x(\nu)2^j|\Psi(2^j\nu)|^2 d\nu \tag{22}$$

where Ψ denote the Fourier transform of the analyzing wavelet ψ. From this relation, one sees that $\hat{S}_x$ suffers from the standard convolutive bias, that is, the spectrum to be estimated is mixed within a frequency range corresponding to the frequency width of the analyzing window at scale j. The crucial point here is that for LRD signals this bias reduces naturally to a simple form, enabling an unbiased estimation of H. Recall that the spectral density of LRD signal follows the power low in (19). Therefore, we can rewrite (22) as

$$\mathrm{E}[\hat{S}_x(2^{-j}\nu_0)] = c_f|2^{-j}|^{(1-2H)}\int |\nu|^{(1-2H)}|\Psi(\nu)|^2 d\nu \tag{23}$$

From (23), one sees that in the case of $1/|\nu|^{\gamma}$ process the standard convolutive bias turns into a multiplicative one. Moreover, this multiplicative constant is independent of the analyzing scale j. It is, therefore, possible to design an estimator $\hat{H}$ for the parameter H from a simple linear regression of $log_2(\hat{S}_x(2^{-j}\nu_0))$ on j as the following:

$$\begin{aligned} log_2(\hat{S}_x(2^{-j}\nu_0)) &= log_2\left(\frac{1}{n_j}\sum_k |d_x(j,k)|^2\right) \\ &= (2\hat{H}-1)j + \hat{c} \end{aligned}$$

where $\hat{c}$ estimates

$$log_2\left(c_f\int |\nu|^{(1-2H)}|\Psi(\nu)|^2 d\nu\right).$$

provided that the integral

$$\int |\nu|^{(1-2H)} |\Psi(\nu)|^2 d\nu \tag{24}$$

converges.

2) Bias of $\hat{H}$: The above definition for $\hat{H}$ holds provided that (19) holds for all frequencies and that (22) converges. We can relax the first condition since in (23) we are free to choose only the range of scales over which (19) holds [33].

Now consider the convergence of (24). In fact, estimation problems in the presence of LRD often arise from the singular behavior of $1/\nu^\gamma$ spectra at $\nu = 0$ which causes such integrals to diverge. When designing the wavelet ψ, one is free to select one of its important characteristics, namely, the number M of vanishing moments. Clearly, the vanishing moment controls the behavior of the Fourier transform of the wavelet about $\nu = 0$

$$|\Psi(\nu)| = O(\nu^M), \nu \to 0.$$

It is easy to check that provided

$$M > H - 1$$

the behavior of $|\Psi(\nu)|^2$ at the origin will be flat enough to balance the singularity of the LRD spectrum, thus ensuring the convergence of (24). When this inequality is satisfied, we have shown that the log-log regression-based H estimator is asymptotically unbiased, and in practice has very low bias even for short data sets [33].

3) Confidence Interval: It has been shown [33] that under Gaussian and quasi-decorrelation of the wavelet coefficient hypotheses and in the asymptotic limit, a closed form for the variance of the estimator $\hat{H}$ can be obtained and is given by

$$\begin{aligned} \sigma^2_{\hat{H}} &= var\hat{H}(j_1, j_2) \\ &= \frac{2}{n_{j_1} ln^2 2} \frac{1 - 2^J}{1 - 2^{-(J+1)}(J^2 + 4) + 2^{-2J}} \end{aligned}$$

where $J = j_2 - j_1$ is the number of octaves involved in the linear fit and $n_{j_1} = 2^{-j_1} n$ is the number of available coefficients at scale j_1. From this closed form for the variance estimation, one can derive a confidence interval

$$\hat{H} - \sigma_{\hat{H}} z_\beta \le H \le \hat{H} + \sigma_{\hat{H}} z_\beta$$

where z_β is the $1 - \beta$ quantile of the standard Gaussian distribution, i.e., $P(z \ge z_\beta) = \beta$. For example, if $\beta = 0.025$, then we can obtain 95% confidence intervals [33].

4) Robustness against Nonstationarity: Assume that the signal $x(t)$ consists of stationary "data" $s(t)$, plus some contaminating deterministic function of time $p(t)$ such that $x(t) = s(t) + p(t)$. Abry and Veitch [33] tried to measure H correctly for the data, and detect and identify the trend with wavelet-based H estimator. They showed that the wavelet-based method accurately estimates H provided that M of the wavelet is tuned to the degree P of the polynomial trend $p(t)$, that is, provided

$$N \ge P + 1.$$

To explain this, recall that a wavelet ψ with M vanishing moments is, by definition of (22), orthogonal to the space of polynomials of degree less than or equal to $M - 1$. Hence the details $d_p(j,k)$ corresponding to $p(t)$ vanish provided that $M \ge P + 1$. It follows that the estimation of H will not be affected by the presence of the trend, as it is entirely absent from the details of the signal, i.e., $d_x(j,k) = d_s(j,k)$. This can also be given a useful spectral interpretation. The Fourier transform of a polynomial of order P consists, within the distribution theory framework, in the P-th derivative of the Dirac impulse function $\delta^{(P)}(\nu)$. The frequency content of a polynomial is therefore concentrated at the null frequency and since wavelets are bandpass functions, in fact satisfying $|\Psi(\nu)| = O(\nu^M), \nu \to 0$, they will be blind to a given polynomial for M sufficiently large [33].

They also showed that even when the trend is not polynomial but some smooth function, increasing M still helps to cancel its influence and very accurate estimates for H. They explain this in two ways. First, selecting an M will effectively cancel the part of the trend, which can be efficiently approximated by polynomials of degree $M - 1$. Increasing M would therefore approximately cancel any smooth function. The second explanation is again from the spectral viewpoint. Smooth trends have, in most cases, an important frequency content near $\nu = 0$. The LRD phenomenon basically consists of a power-law behavior of the spectrum near $\nu = 0$ and in general this overlap significantly complicates analysis and estimation. To see why choosing wavelets with high M significantly improves this situation consider the power-law trend $p(t) = at^\gamma$, with a a constant. The wavelet coefficients read

$$d_p(j,k) = 2^{j(\gamma+1/2)} C \int |\nu|^{-(\gamma+1)} \Psi(\nu) e^{(i2\pi k\nu)} d\nu$$

where C is a constant independent of the scale j. It can be checked numerically that for a given j, the magnitude of these coefficients decreases with increasing M. Increasing M therefore enlarges the range of scales where $|d_p(j,k)| \ll |d_s(j,k)|$, that is, where the effect of the trend is negligible [33].

VI. Multifractal Analysis

In this section, we introduce multifractal analysis and its application to network traffic analysis.

A. Fractals, Fractal Dimensions and Random Fractals

Fractals are geometric objects exhibiting an intricate, highly irregular appearance on all resolutions [34], [35] Geometric (nonrandom) fractals have a characteristic called *self-similarity*, which indicates that they are made of smaller copies of themselves iteratively.

Another important characteristic is that fractals have *fractional* dimensions. For example, Sierpinski gasket, a geometric fractal, has a dimension of $log_2 3 \approx 1.59$, while the filled-in square has a dimension of 2. We can apply so-called *Box-Counting* method to measure the dimension of a geometric fractal object. In Box-Counting method, for different side lengths r we count $N(r)$, the smallest number of *boxes* of side length r needed to cover the shape. For example, if the shape is 1-dimensional, such as the line segment, we see $N(r) = 1/r$, and if the shape is 2-dimensional, such as the (filled-in) unit square, $N(r) = (1/r)^2$, and so on. Once we obtain $N(r)$ for various values of r, then we can fit their relations into this power law equation:

$$N(r) = k \left(\frac{1}{r}\right)^d$$

where d is the fractal dimension of a given object [34].

Fractals fall naturally into two categories, *nonrandom* and *random*. Nonrandom fractals are made by iteration of a simple *growth rule*. For example, Sierpinski gasket is defined operationally as an "aggregation process" obtained by a simple iterative process. Real systems in nature, however, do not resemble the Sierpinski gasket - in fact, *nonrandom* fractals are not found in nature. Nature exhibits numerous examples of objects which by themselves are not fractals but which have the remarkable feature that, if we form a *statistical average* of some property such as the density, we find a quantity that decreases linearly with the length scale when plotted on double-logarithmic scales. Examples of such *random* fractals include (fractional) Brownian Motion, bacteria growth in stressful condition, large scale distribution of galaxies and stock market. The network traffic, which is our main interest, is another example of random fractals [34], [35].

B. Multifractal Analysis

In self-similar analysis, we focus on the scaling behavior at large time scale, *asymptotically*, simply ignoring the small-scale behavior. Moreover, we examine only the second-moment behavior, i.e. autocorrelation, to obtain a single scaling exponent H. According to the definition of *exact* self-similarity, power law should hold for any moment q, which can be higher than 2:

$$\mathrm{E}|X(t)|^q = \mathrm{E}|X(1)|^q |t|^{qH}$$

In actual network traffic, however, no longer a single exponent can determine the self-similarity at higher moments.

Multifractal analysis [41], [36], [37], [43] aims to overcome such limitations of self-similar (monofractal) analysis. In multifractal analysis, we rather focus on the distribution of the *local irregularity* at small time scale. In addition, scaling exponents in multifractal analysis are not a single constant, but a *spectrum* $H(q)$:

$$\mathrm{E}|X(t)|^q = \mathrm{E}|X(1)|^q |t|^{qH(q)}$$

In the following sections, we first introduce the concept of local irregularity and then describe how we can characterize it with a function, called *multifractal spectrum*.

1) Local Hölder Exponent h: For each point (x, y) of the fractal, and for each radius r, the *measure* $m((x, y), r)$ of the part of the fractal within a distance r of (x, y) is the sum of all the probabilities of those parts of the fractal. For each point (x, y) of the fractal, the *local Hölder exponent* $\hat{\alpha}(x, y)$ at (x, y) is:

$$\hat{\alpha}(x, y) = \lim_{r \to 0} \frac{\log m((x, y), r)}{\log r}$$

if the limit exists. The local Hölder exponent $\hat{\alpha}(x, y)$ indicates how *irregular* the mass is distributed around the point (x, y). Note that high local Hölder exponent corresponds low irregularity, therefore high regularity [34].

If α is the local Hölder exponent of a point of the fractal, then

$$E_\alpha = \{(x, y) : \hat{\alpha}(x, y) = \alpha\}$$

is the collection of all points of the fractal having local exponent α. As α takes on all values of the local hölder exponents, we decompose the fractal into these sets E_α.

Because each local hölder exponent α is the exponent for a power law, a multifractal is a process exhibiting scaling for a range of different power laws. The multifractal structure is revealed by plotting $dim(E_\alpha)$ as a function of α, which is called $f(\alpha)$ curve, or multifractal spectrum $f(\alpha)$ [34].

2) Multifractal Spectrum $f(\alpha)$: In this section, we describe how we can calculate the multifractal spectrum from some mass distribution rule, called the *Iterated Function System (IFS)* rule. We first introduce IFS then describe the way to construct multifractals from IFS and finally the procedure to obtain multifractal spectrum $f(\alpha)$ [34].

Generating fractals by iterating a collection of transformations is the Iterated Function System (IFS) method. A *contraction* is a transformation T that reduces the distance between every pair of points. That is, there is a number $r < 1$ with

$$d(T(x, y), T(x', y')) \leq r * dist((x, y), (x', y'))$$

for all pairs of points (x, y) and (x', y'), where d denotes the Euclidean distance between points:

$$d((x, y), (x', y')) = \sqrt{(x - x')^2 + (y - y')^2}$$

The *contraction factor or ratio* of T is the smallest r satisfying

$$d(T(x, y), T(x', y')) \leq r * d((x, y), (x', y'))$$

for all pairs of points $(x, y), (x', y')$. Given IFS rules, the Deterministic Algorithm renders a picture of the fractal by

1) applying all the rules to any (compact) initial picture,
2) then applying all the rules to the resulting picture,
3) and continuing this process.

Specifically, suppose $T_1, \ldots, T_n$ are contractions, the following rule renders Sierpinski gasket with any initial picture [34].

$$\begin{aligned} T_1(x, y) &= (x/2, y/2), \\ T_2(x, y) &= (x/2, y/2) + (1/2, 0), \\ T_3(x, y) &= (x/2, y/2) + (0, 1/2). \end{aligned}$$

In the *Random IFS* Algorithm the transformations T_i are applied in random order, but they need not be applied equally often. Associated with each T_i is a probability $p_i, 0 < p_i < 1$, representing how often each transformation is applied. That is, when N points are generated, each T_i is applied about $N * p_i$ times [34].

A simple way to construct multifractals is to use an IFS with transformations $T_1, \ldots, T_N$, contraction ratios $r_1, \ldots, r_N$, and probabilities $p_1, ..., p_N$. To find a measure of the complexity of the multifractal built this way, we use the $f(\alpha)$ curve, which we determine through an auxiliary function B. Traditionally, the variable for B is called q, and the function $B(q)$ is the solution of the equation

$$p_1{}^q r_1{}^{B(q)} + ... + p_N{}^q r_N{}^{B(q)} = 1$$

We begin with the equation

$$\sum_{i=1}^{n} p_i{}^q r_i{}^{B(q)} = 1$$

Differentiating once with respect to q gives

$$\sum_{i=1}^{n} (p_i{}^q ln(p_i) r_i{}^{B(q)} + p_i{}^q r_i{}^{B(q)} ln(r_i) B'(q)) = 0 \tag{25}$$

Solving (25) for B' gives

$$\frac{dB}{dq} = B' = -\frac{\sum_{i=1}^{n} p_i{}^q r_i{}^{B(q)} ln(p_i)}{\sum_{i=1}^{n} p_i{}^q r_i{}^{B(q)} ln(r_i)} \tag{26}$$

therefore, $B' < 0$. Differentiating (25) again and simplifying gives

$$\sum_{i=1}^{n} p_i{}^q r_i{}^{B(q)} ((ln(p_i) + ln(r_i) B'(q))^2 + ln(r_i) B''(q)) = 0 \tag{27}$$

Solving for B'' gives

$$B'' = -\frac{\sum_{i=1}^{n} p_i{}^q r_i{}^{B(q)} (ln(p_i) + ln(r_i) B')^2}{\sum_{i=1}^{n} p_i{}^q r_i{}^{B}(q) ln(r_i)}$$

therefore $B'' > 0$.

Define α by $\alpha = -dB/dq$, then

$$\alpha \geq \alpha_{min} = min\{\frac{ln(p_i)}{ln(r_i)} : i = 1, \ldots, n\}$$

$$\alpha \leq \alpha_{max} = max\{\frac{ln(p_i)}{ln(r_i)} : i = 1, \ldots, n\}$$

Finally, $B = -\alpha_{max} q$ and $B = -\alpha_{min} q$ are the $q \to -\infty$ and $q \to \infty$ asymptotes of the $B(q)$ curve. Through each point $(q, B(q))$ on the $B(q)$-curve there is a unique tangent line. Because the $B(q)$-curve decreases, the tangent line slopes downward, so we denote the slope of the tangent line by $-\alpha$. The tangent line intersects the vertical axis at the point $B(q) + \alpha q$; call this intercept $f(\alpha)$. Proving that $f(\alpha) = dim(E_\alpha)$ is quite delicate, therefore, is omitted here [34].

C. Multifractal Analysis for Traffic Models

In the previous section we introduce the general notions of multifractals and multifractal spectrum. Now we explain such notions in the context of the network traffic analysis according to Riedi *et al.* [36] and Gilbert *et al.* [41].

1) Spikiness: The concept of *irregularity* is interpreted as *spikiness* in network traffic. The strength of growth, called the local Hölder exponent (which is also termed *degree of Hölder continuity*), of an increasing process Y at time t can be characterized by

$$\hat{\alpha}(t) := \lim_{k^n 2^{-n} \to t} \alpha_{k_n}^n \tag{28}$$

with

$$\alpha_{k_n}^n := -\frac{1}{n} \log_2 \Delta_{k_n}^n [Y] \tag{29}$$

$$\Delta_{k_n}^n [Y] := |Y((k_n + 1)2^{-n}) - Y(k_n 2^{-n})| \tag{30}$$

$$= 2^{-n\alpha_{k_n}^n} \tag{31}$$

and $k_n = 0, \ldots, 2^n - 1$. The smaller the $\hat{\alpha}(t)$, the larger the increment of Y around t, and the "burstier" it is at t. Note that high burstiness here corresponds small regularity, therefore high irregularity. Considering only $t \in [0, 1]$ for simplicity, the frequency of occurrence of a given strength α can be measured by the *multifractal spectrum*

$$f(\alpha) := \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log_2 \#\{\alpha_{k_n}^n \in (\alpha - \epsilon, \alpha + \epsilon)\} \tag{32}$$

In this setting, f takes values between zero and one and is often shaped like $\cap$ (concave). The smaller $f(\alpha)$ is, the "fewer" points t act like $\hat{\alpha}(t) \simeq \alpha$. If $\bar{\alpha}$ denotes the value $\hat{\alpha}(t)$ assumed by "most" points t, then $f(\bar{\alpha}) = 1$ [36].

2) Scaling at Higher Order Moments: It is often quite difficult to calculate the dimension (in this context, *Housdorff dimension*) of random fractals. Another way to obtain the multifractal spectrum of a measure $\alpha_{k_n}^n$ which is more statistical rather than geometric is to consider the *scaling of moments*. Being especially interested in the scaling of moments, the *partition function* $Z(q, n)$ can be defined:

$$Z(q, n) := \mathrm{E}\left[\sum_{k_n=0}^{2^n-1} (\Delta_{k_n}^n [Y])^q\right]. \tag{33}$$

Also the *structure function* $T(q)$ can be defined, which indicates the scaling behavior of $Z(q, n)$ at *finer* time scale (as $n \to \infty$):

$$T(q) := \lim_{n \to \infty} \frac{1}{-n} \log_2 Z(q, n). \tag{34}$$

Note that T is always concave [36].

3) The Multifractal Formalism: The multifractal spectrum $f(\alpha)$ and the structure function $T(q)$ are closely related, as the following quick argument shows. Omitting in the sum of (33) all terms but the ones with $\alpha_{k_n}^n \approx \alpha$ and using (32), we obtain

$$\begin{aligned} \sum_{k_n=0}^{2^n-1} (\Delta_{k_n}^n [Y])^q &\geq \sum_{\alpha_{k_n}^n \sim \alpha} (2^{-n\alpha})^q \\ &\simeq 2^{nf(\alpha)} 2^{-nq\alpha} \\ &= 2^{-n(q\alpha - f(\alpha))} \end{aligned} \tag{35}$$

where $2^{nf(\alpha)}$ indicates the number of points t with $\hat{\alpha}(t) = \alpha$. According to [36], for all α and q,

$$T(q) \leq f^*(q) := \min_{\alpha}(q\alpha - f(\alpha))$$

and

$$f(\alpha) \leq T^*(\alpha) := \min_{q}(q\alpha - T(q)). \tag{36}$$

The transform $T^*(\alpha)$ appearing in (36) is called the *Legendre transform*. If $T''(q) < 0$, then we can find by simple calculus that

$$\begin{aligned} T^*(\alpha) &= q\alpha - T(q) \text{ and} \\ (T^*)'(\alpha) &= q \text{ at } \alpha = T'(q). \end{aligned} \tag{37}$$

We may write this equivalently as the dual formula

$$T(q) = q\alpha - T^*(\alpha)\ ,\ T'(q) = \alpha \text{ at } q = (T^*)'(\alpha).$$

The procedure we explained in this section is called *multifractal formalism* [36], [42].

4) Multifractal Analysis and Wavelets: In this section, we discuss wavelet-based multifractal analysis, assuming we aim to analyze an actual trace with wavelet tools. Wavelet decompositions contain considerable information on the singularity behavior of a process Y. Recall the wavelet decomposition in (21) with $J = 0$:

$$\begin{aligned} Y(t) &= \sum_k a_x(0,k)\phi_{0,k}(t) \\ &+ \sum_{j\leq 0}\sum_k d_x(j,k)\psi_{j,k}(t). \end{aligned} \tag{38}$$

In [36] they showed that $|Y(s)-Y(t)| = O(|s-t|^{\alpha})$ implies that for $j \leq 0$

$$2^{-j/2}|\int Y(s)\psi_{j,k_j}(s)ds| = 2^{-j/2}d_{j,k} = O(2^{j\alpha(t)})$$

if k_j is chosen as usual to satisfy $k_j 2^j \leq t \leq (k_j+1)2^j$. This holds for any $\alpha > 0$ and any compactly supported wavelet. Therefore, the wavelet coefficients all behave like $d_{j,k} \approx 2^{j(\alpha+1/2)}$ as j tends to $-\infty$. In this case evaluating the so-called *wavelet-based partition function* $S(q,j)$, defined by summing across each level j the q-th moments (with $q \leq 0$) of the absolute value of the normalized wavelet coefficients $\hat{d}_{j,k} = 2^{-j/2}d_{j,k}$; i.e., setting

$$S(q,j) = \sum_k |\hat{d}_{j,k}|^q, \tag{39}$$

we obtain $S(q,j) \approx 2^{-j}2^{jq\alpha(t)} = 2^{-j(1-\alpha q)}$. For $q > 2$, the function $S(q,j)$ takes into account the effects of higher-order statistics that may be present in a trace and hence may be contained in the DWT of the trace. Moreover, because wavelet coefficients tend to decorrelate quickly within a given scale as well across different scales [40], it can be expected that hardly any information about possibly strong correlations within the trace is lost by defining the partition function $S(q,j)$ as in (39) [36].

To examine the scaling behavior of $S(q,j)$ as the time scale or resolution level becomes finer and finer (i.e. $j \to -\infty$), we consider the corresponding *wavelet-based structure function* $\tau(q)$ defined as the scaling exponent of $S(q,j)$, as $j \to -\infty$; that is,

$$\tau(q) = \lim_{j\to-\infty}\frac{\log S(q,j)}{j\log 2}, \tag{40}$$

In other words, we check whether or not the partition function behaves like $S(q,j) \approx 2^{j\tau(q)}$ as we look at finer and finer time scales. For example at hand, it is easy to see that $\tau(q) = \alpha q - 1$; i.e., the structure function of a monofractal signal is linear in q. In particular, if the trace is self-similar with Hurst parameter H, the $\tau(q) = Hq - 1$ and H can be easily inferred from the structure function. A more or less linear $\tau(q)$ function is consistent with monofractal scaling and rules out multifractality. On the other hand, the more concave the shape of $\tau(q)$, the wider the range of local scaling exponents found in the signal; in particular, a concave shape of the structure function is consistent with multifractality [38].

In general, α is a set of local Hölder exponents, and we can derive $f(\alpha)$ applying Legendre transformation of $\tau(q)$ with the equations (36) and (37). We can think of the multifractal spectrum $f(\alpha)$ as a "histogram", which measures the number of instants in the trace which have local scaling exponent α.

D. Multifractal Analysis of WAN Traffic

In [38], they apply the Haar wavelet-based DWT structure function method to a number of LAN, WAN and test traces. For each trace, they picked 10 milliseconds as the finest resolution level (i.e., $j = -18$) and examined the scaling behavior of the partition function $S(q,j)$ over a range of fine resolution levels, i.e., for j-values bigger than -18. All log-log partition function plots suggest the presence of well-defined fine-time scaling regions over fine time scales where reading off the slopes of the different lines with different q values, to determine the value of the structure function $\tau(q)$ at different q's.

The log-log partition function plots for Bellcore LAN trace, self-similar trace and a Poisson trace show that all three traces result in linear $\tau(q)$ functions of the form $\tau(q) = Hq - 1$, and are hence fully consistent with monofractal scaling behavior. In contrast, the $\tau(q)$ functions for the three WAN traces show indications of nonlinear, i.e., concave shapes that are inconsistent with monofractal behavior and suggest multifractal structure over small time scales [38].

They successfully applied the multifractal analysis techniques on actual WAN traces, however, they provided only the evidence of multifractality in WAN, did not interpret it in terms of the specific characteristic of each traffic trace. Instead they offered a general theory on why such WAN traffic exhibits multifractality. They plausibly argued that the cause of multifractality is the multiplicative cascade structure of WAN traffic, which we will go over in the following sections.

E. Multiplicative Cascade Model

A construction that fragments a given set into smaller and smaller pieces according to some geometric rule and, at the same time divides the measure of these pieces according to some other (deterministic or random) rule is called a *multiplicative process* or *cascade*. In this section, we explain how the multiplicative cascade model can generate multifractal structure according to the works in [38], [41], [36], [42].

1) Conservative Cascade Model and Multifractal Analysis: The limiting object generated by such a multiplicative process defines, in general, a singular measure or multifractal and describes the highly irregular way the mass of the initial set gets redistributed during this simple fragmentation procedure. In [41], [36], they construct a binomial conservative cascade or measure μ on the interval $I := [0, 1]$. By *conservative cascades* they mean the class of cascades which can be characterized by a generator (mass fragmentation rule) that preserves the total mass of the initial set at every stage of the construction (i.e. mass conservation). More precisely, they construct the distribution function $Y(t) = \mu([0, t])$ and since the underlying generator will be random. Y defines a stochastic process. By construction it will have positive increments and $Y(0) = 0$ almost surely.

This iterative construction starts with a uniform distribution on the unit interval of total mass M^0 and then "redistributes" this mass by splitting it among the two subintervals of half size in the ratio M_0^1 to M_1^1 where $M_0^1 + M_1^1 = 1$. Proceeding iteratively on obtains after n steps a distribution which is uniform on intervals $I_{k_n}^n := [k_n 2^{-n}, (k_n + 1)2^{-n}]$. The mass lying in $I_{k_n}^n$ is redistributed among its two dyadic subintervals $I_{2k_n}^{n+1}$ and $I_{2k_n+1}^{n+1}$ in the proportions $M_{2k_n}^{n+1}$ and $M_{2k_n}^{n+1}$ where $M_{2k_n}^{n+1} + M_{2k_n}^{n+1} = 1$ almost surely [36].

To summarize, for any n let us choose a sequence $k_1, k_2, \ldots, k_n$ such that the interval $i_{k_l}^l$ lies in $I_{k_i}^i$ whenever $i < l$. In other words, the k_i are the n first binary digits of any point $t \in I_{k_n}^n$. Therefore it is called a *nested sequence*, and it is uniquely defined by the value k_n [42]. Then we have

$$\begin{aligned} Y(*k_n + 1)2^{-n}) - Y((k_n)2^{-n}) &= \mu(I_{k_n}^n) \\ &= M_{k_n}^n \cdot M_{k_{n-1}}^{n-1} \cdots M_{k_1}^1 \cdot M_{k_0}^0 \end{aligned} \quad (41)$$

The various M_l^i, which collectively define the generator of the conservative cascade, may have distributions which depend on i and l and which are arbitrary, as long as they are positive and provided that for all i and all m,

$$M_{2m}^i + M_{2m+1}^i = 1. \quad (42)$$

Note that this mass conservation condition introduces a strong dependence between the two "children" of any parent node. Furthermore, it is required that for all n and k_n ($n = 1, 2, \ldots$), all the multipliers appearing in (41) are mutually independent. This property is called *nested independence*. As long as these two requirements on dependency are satisfied one is completely free in how to introduce further correlation structure [42].

It is obvious from this iterative construction and from relation (41) that a multiplicatively generated "multifractal process" has approximately *lognormal* marginals. Indeed, as a sum of independent random variables, the logarithms of the increments of Y are approximately Gaussian, provided that the random variables $\log M_l^i$ have finite second moments [42].

To obtain the singularity structure of Y using $\alpha(t)$, we can calculate the structure function $T(q)$ of the binomial conservative measure "in expectation". To this end, we can assume that the $M_k^n (k = 0, \ldots, 2^n - 1)$ are identically distributed with $M^{(n)}$. Note that $M^{(n)}$ is necessarily symmetrically distributed around $1/2$ due to (42). Then, (41) is equally distributed as $M^{(n)} \cdots M^{(1)} \cdot M^{(0)}$ for each of the 2^n nested sequences $k_1, \ldots, k_n$ of length n. Using the "nested" independence we find

$$\mathrm{E}[S_n(q)] = 2^n \cdot \mathrm{E}((M^{(n)})^q \cdots \mathrm{E}(M^{(1)})^q \cdot \mathrm{E}(M^{(0)})^q \quad (43)$$

Assuming now further that the $M^{(n)}$ converge in distribution, say to M, we have

$$T(q) := \lim_{n \to \infty} \frac{-1}{n} \log_2 \mathrm{E}S_n(q) = -1 - \log_1 \mathrm{E}[M^q] \quad (44)$$

We can calculate for every α [42],

$$dim(E_\alpha) = f(\alpha) = T^*(\alpha) \quad (45)$$

2) Wavelet Analysis of Conservative Cascade: In [38], [41] they summarize the main results of a (Haar) wavelet-based global and local scaling analysis applied to the class of conservative cascades. They show that the DWT of conservative cascades gives rise to a set of analysis and inference tools that allows us to detect and identify the global and local scaling properties of multifractal objects generated by the conservative cascades.

To start, consider a conservative cascade with fixed generator M; i.e. M has mean $1/2$, takes on values in $(0, 1)$ and is symmetric about its mean. If Y denotes the limiting multifractal generated by this conservative cascade, then Y has global linear scaling; that is the logarithm of the expected value of the energy E_l in X around level l in the cascade construction of Y depends linearly on l (plotted from large l, fine scales, to small l, coarse scales for $l \leq 0$) and has the form [38].

$$\log_2 \mathrm{E}[E_l] = (1 + \log_2 \mathrm{E}[M^2])l + \log_2 \mathrm{E}[(2M - 1)^2]. \quad (46)$$

Note that the slope $1 + \log_2 \mathrm{E}[M^2]$ depends only on $\mathrm{E}[M^2]$, the second moment of the generator. thus, if we want a nonlinear global scaling behavior for Y, we need to change the second moment of the generator M at each level (or within a range of levels) in the cascade construction. One way to achieve this is to let $M^{(l)}$ be equal in distribution to $\lambda_l M + 1/2(1 - \lambda_l)$ where $\lambda_l^2 = Var(M^{(l)})/Var(M) \leq 1$. The limiting object Y resulting from a conservative cascade construction with this type of variable generator can be shown to exhibit non-linear global scaling behavior. In particular, λ_l^2 of the generators at each stage in the construction process increase (decrease) monotonically (as we go from the coarsest scale to the finest) then the slope of the global scaling analysis increases (decreases) monotonically from finest scale to the coarsest [38].

Turning to the local scaling analysis of a multifractal Y generated by a conservative cascade with fixed generator M, the DWT structure function $\tau(q)$ defined in (40), that is, the scaling exponent of the partition function $S(q, j)$, given by (39), can be computed as

$$\tau(q) = -1 - \log_2 \mathrm{E}[M^q], q > 0$$

Moreover, the multifractal spectrum $f(\alpha)$ of Y can be obtained from $\tau(q)$ by setting

$$f(\alpha) = \min_q (q\alpha - \tau(q)).$$

The same results hold if the fixed generator M is replaced by a variable generator of the type considered above, with a more complicated expression for the DWT structure function associated with the underlying limiting multifractal [38].

F. Physical Explanation of Multifractal Nature of WAN

In [38], [41], [42], they attempted a plausible explanation on multifractal nature of Wide Area Network, based on multiplicative cascade structure of network protocols.

1) Additive vs. Multiplicative Natures: The mathematical results and physical explanations of the observed self-similarity or monofractal nature of measured traffic traces state explicitly that self-similarity is an *additive* property of network traffic. That is, self-similarity arises from aggregating many *ON/OFF*-streams [15] or from superposing many renewal-type connections, provided the individual *ON/OFF*-periods or connection durations/sizes exhibit extreme variability (i.e., are heavy-tailed with infinite variance). As such, self-similarity is plausibly the result of user behavior, application-specific features (e.g., layout of web pages), and the inherent properties of the objects (e.g., sizes of text, picture, video, audio files) that are sent across the network. I particular, these findings imply that the precise nature of the local traffic structure within individual *ON*-periods or connections is not essential for self-similarity of the aggregate traffic stream. Being additive in nature, aggregate network traffic will be approximately normal when viewed over sufficiently large time scales, provided certain weak conditions on the individual traffic streams hold for the central limit theorem to apply [38].

On the other hand, the observed multifractal nature of network traffic over small time scales and the empirical evidence in support of an underlying cascade mechanism implies that over those fine time scales, network traffic is *multiplicatively* generated. In other words, at the microscopic level, the traffic rate process has an approximate lognormal shape because it is the product of a large number of more or less independent "multipliers" [38].

In [38], they showed that the clearest distinction between the additive aspect of measured network traffic (over large time scales) and its multiplicative property (over small time scales) can be seen at the level of port-to-port flows. They define a *port-to-port flow* as *consisting of all packets flowing in either direction between two IP hosts that use the same source and destination port numbers and that are separated in time by less that 60 seconds*. They also found that the change in the global scaling behavior from small-time to large-time scales, around time scales on the order of a few hundred milliseconds or seconds. They related the location of such "knee" to properties of the round-trip time in the network [38].

2) Multifractal Workload Models: Based on the measured multifractal phenomenon and its analysis, Feldmann *et al.* [38] discussed the workload modeling approach as the following:

1) User sessions arrive in accordance to a Poisson process,
2) bring with them a workload (e.g., number of packets or bytes, number of port-to-port flows, length of session) that is heavy-tailed with infinite variances, and
3) distribute the workload over the lifetime of the session according to a multiplicatively generated multifractal with a conservative cascade generator.

This workload model is a generalization of Kurtz's model [39], by allowing the within-session traffic rate process to be generated by a conservative cascade model.

Riedi *et al.* [42] pointed out that the above generalization of Kurtz's workload model is not consistent with measured data. They noticed that the multiplicative structure in measured traffic is clearly isolated at TCP level and moreover, the overall number of TCP connections per time unit exhibits self-similar scaling behavior for time scales on the order of seconds and beyond. Therefore, they simply modify the multifractal version of Kurtz's process and require that:

1) User-initiated sessions continue to arrive in a Poisson fashion, but the workload is now expressed in terms of the number of TCP connections that make up a particular session and remains to be heavy-tailed with infinite variances and
2) the TCP connections' workload is heavy-tailed with infinite variance; and
3) the workload of a TCP connection is distributed over the connection's lifetime in a multifractal fashion, i.e., according to a conservative binomial cascade.

It is then easy to see that this two-tier approach to describing aggregate WAN traffic yields the additive traffic component via the TCP-connection-within-session structure and the multiplicative component via the dynamics prescribed for the packets within individual TCP-connections. Moreover, this two-tier approach is also fully consistent with measured Internet traffic at the different layers in the TCP/IP protocol hierarchy [42].

VII. Multiscale Analysis for the IEEE 802.11 Wireless LAN Traffic Characterization

In this section, we describe the approaches we can apply the multiscale analysis techniques to wireless LAN traffic characterization. By *multiscale analysis* we mean wavelet-based self-similar and multifractal analysis. We expect that multiscale analysis can reveal the traffic characteristics of the IEEE 802.11 wireless LAN, especially the difference from other traffics, which might be due to the MAC protocol and lossy wireless links.

A. Motivation

Previous measurement-based analysis studies on wireless LAN [3], [4], [5] have focused on characterization of traffic *patterns* in terms of usages (e.g. daily, monthly average traffic sizes) and performances (e.g. daily, monthly throughput variability). Those results are useful for management and deployment of the wireless LAN, but still not enough especially when we want to look at the detailed performance behaviors with various workloads and conditions. Therefore in our work, we take the approach of *workload characterization* based on more rigorous traffic model frameworks. For this purpose we overviewed various models and analysis techniques in this paper.

On measurement issues, in the previous studies the measurements were made at some *wired* locations to capture the wireless traffics. We instead measure the traffic at some *wireless* vantage points, by placing so-called *wireless sniffers* in between the stations and the Access Point [6]. This *wireless* measurement techniques enable us to examine the IEEE 802.11 frames, therefore, provide the various information on MAC protocols and wireless link conditions. With the MAC frames available, we can investigate the impacts of the IEEE 802.11 MAC protocol and the lossy wireless links.

With these approach and setup, we have the following motivations on the study of wireless LAN characterization: First, we want to investigate the effect of IEEE 802.11 MAC protocol and lossy wireless links on (higher layer) traffic characteristics. As mentioned above, our measurements at wireless vantage points can provide enough information on the MAC protocol and wireless links. Second, as we take the approach of workload characterization, we need to analyze higher-order statistics (e.g. second-order) than just first-order statistics (e.g. mean, variance and marginal distributions). Third, we need to analyze the data on large time scale to investigate the presence of self-similarity, and at the same time we want to examine small-time-scale data to look at the microscopic characteristics of wireless LAN traffic. Fortunately we have small-time-scale data, e.g. those of the IEEE 802.11 frames in microsecond resolution. Forth, according to the works [38], [42], the workload can be better characterized in multi-tier fashion than in only one network layer. As seen in Paxson's work [16], the same traffic can be differently characterized in different layers. We need to examine the wireless traffic at the level of users (user sessions), port-to-port flow (TCP/UDP connection), network protocol (IP packet) and MAC protocol (the IEEE 802.11 frames) respectively.

For these motivations, what traffic model is the most proper? We believe that the multiscale analysis can reveal the unknowns in wireless LAN traffics as it can satisfy the requirement of each motivation described above. We can apply multifractal analysis technique to the trace to obtain so-called multifractal spectrum on small-time-scale behaviors. By comparing it with that of Ethernet LAN [38] we can examine how similar or different the wireless LAN traffic is from the Ethernet LAN traffic, which in turn can reveal the impact of the IEEE 802.11 MAC protocol. We can also analyze the trace at source-level, in the similar way to [15], classify the sources with wireless link conditions, then examine the impact of lossy wireless links on the traffic. Using wavelet-based mathematical tools, we can estimate the second-order properties and Hurst parameter H [33] on both large and small time scales. We can apply those techniques on wide range of layers, e.g. from user sessions to the IEEE 802.11 frames.

B. Description of Wireless LAN Traffic Processes

In this section we define the various processes at each protocol layer for the analysis of wireless LAN traffics.

We first think of a point process $\{P_t, t \in \mathrm{R}^+\}$ which has 1 at a frame arrival and 0 otherwise. Then we can obtain $\{C_{\delta,n}, n = 1, 2, \ldots\}$, a discrete version of $\{P_t\}$ by integrating $\{P_t\}$ over the duration $[n\delta, (n+1)\delta]$. We can think of another point process $\{X_t, t \in \mathrm{R}^+\}$ where the state space is the set $\{0, 1\}$, corresponding to the presence or absence of a frame byte. Then we can obtain $\{W_{\delta,n}, n = 1, 2, \ldots\}$, a discrete version of $\{X_t\}$ by integrating $\{X_t\}$ over the duration $[n\delta, (n + 1)\delta]$. We can also define $\{A_n, n = 1, 2, \ldots\}$, the sequence of interarrival times and $\{F_n, n = 1, 2, \ldots\}$, the sequence of frame sizes.

For example, consider a traffic arrival scenario where initially, a frame of 100 bytes is transmitted during 50 microseconds, followed by 30 microsecond silence, then followed by two arrivals of 20 byte frames during 20 microseconds (10 microseconds each) and finally followed by silence of 40 microseconds (assumed to be ended by another arrival), and so on. Then $\{C_{40\mu s,n}\}$ and $\{W_{40\mu s,n}\}$ will be $\{1, 0, 2, 0\ldots\}$ and $\{80, 20, 40, 0\ldots\}$ respectively. We also have $\{A_n\} = \{80, 10, 50, \ldots\}$ and $\{F_n\} = \{100, 20, 20, \ldots\}$.

Now we define the actual processes of wireless LAN traffic with the above definition. In the following, the process $\{Y^{(L)}\}$ denotes the process defined on the data obtained from the traffic layer L.

- (User session) By *user session*, we mean the duration one user (station) keeps associated with the AP (similar to the definition in [4]. To model the user session, we define the corresponding user session interarrival process, $\{A_n^{(US)}\}$, and have the distribution of *user session duration* and distribution of the number and transmitted byte size of *port-to-port flows* (which are defined below) within a session.
- (Port-to-port flow) A port-to-port flow, called a *flow*, abridged, is defined as that consisting of all packets flowing in either direction between two IP hosts that use the same source and destination port numbers and that are separated in time by less than 60 seconds [38]. Flow is preferred rather than TCP connection because it is applicable to non-TCP traffic. We can define the corresponding flow processes such as $\{C_{\delta,n}^{(PF)}\}$ (e.g. $\delta =$ 1 second) and $\{W_{\delta,n}^{(PF)}\}$, where $\{W^{(PF)}\}$ is defined on the total transmitted byte size of a flow, which is defined

as the sum of all the packet (or frame) sizes within the flow.

- (IP packet) We consider only the IP packets which are *successfully* transmitted. We define the corresponding processes, $\{C_{\delta,n}^{(IP)}\}$ (e.g. $\delta = 1$ millisecond), $\{W_{\delta,n}^{(IP)}\}$, $\{A_n^{(IP)}\}$, and $\{F_n^{(IP)}\}$.
- (The IEEE 802.11 data frame) We consider only the IEEE 802.11 data frames. Then we can define the corresponding 802.11 frame data traffic processes, $\{C_{\delta,n}^{(FD)}\}$ (e.g. $\delta = 1$ millisecond), $\{W_{\delta,n}^{(FD)}\}$, $\{A_n^{(FD)}\}$, and $\{F_n^{(FD)}\}$.
- (The IEEE 802.11 frame) We consider all the IEEE 802.11 frames, including control and management frames as well as data frames. Then we can define the corresponding 802.11 frame traffic processes, $\{C_{\delta,n}^{(FA)}\}$ (e.g. $\delta = 1$ millisecond), $\{W_{\delta,n}^{(FA)}\}$, $\{A_n^{(FA)}\}$, and $\{F_n^{(FA)}\}$.

C. Data Available from Wireless LAN Measurement

In this section we describe our detailed measurement setup and the data available from the measurement.

In a given BSS (Basic Service Set, consisting of one AP and multiple stations), we assume that there are a single or multiple *wireless* sniffers in between the user stations and the AP, and a *wired* sniffer just behind the AP. Therefore, assuming no measurement loss at BSS-level, we have all layers of data, e.g. user sessions, flows, IP packets and the IEEE 802.11 frames, which are aggregated from a small number of users.

From the measurement point just behind the AP, we have the flows each consisting of IP packets, which are successfully transmitted from a BSS to the wired part over the wireless link.

In a given DS (Distribution System, a set of BSSs), we assume that there are a *wired* sniffers at the bridge between the DS and backbone network (other Ethernet LAN or WAN). Therefore, at DS-level, we have the flows each consisting of IP packets, which are successfully transmitted from a number of BSS's, therefore are aggregated from a large number of users.

D. Analysis Strategies

In this section we describe our analysis strategies for each target traffic that we want to model.

1) Modeling User Sessions: In user session traffic, one of the main interests is whether we can model user session interarrival time $\{A_n^{(US)}\}$ with Poisson or MMPP (Markov-Modulated Poisson Process). We are also investigating how we can model the workload of each user session.

To characterize the workload we need to examine the marginal distribution of user session duration and investigate whether and how heavy-tailed it is. The workload of user session can be modeled by the number and transmitted byte size of the flows in a user session. For this purpose, we need to examine the marginal distribution of the number and size of flows of each user session.

For detailed analysis, traffic modeling with the counting process $\{C_{\delta,n}^{(PF)}\}$ and work process $\{W_{\delta,n}^{(PF)}\}$ of the flows in each user session, can reveal the workload structure of user sessions. We need to perform wavelet-based large-scale analysis to investigate the presence of self-similarity in the flow traffics within a user session.

2) Modeling Port-to-Port Flows: Each flow consists of either IP packets or the IEEE 802.11 frames. We need to categorize the flows and select the specific sets of flows. Then we can examine the additive and multiplicative nature of the traffics.

We are mainly interested in the impact of lossy wireless links, therefore we can select only the flows that are established over two wireless links with extreme link conditions, namely "Good" or "Bad" wireless link. Note that we assume that we have nearly complete 802.11 traffic trace which includes the information on PHY/MAC layers and above. Therefore we can select those flows based on the information we capture from the trace.

Suppose we select six sets of flows, i.e., FLOW-IP-GOOD, FLOW-FD-GOOD, FLOW-FA-GOOD, FLOW-IP-BAD, FLOW-FD-BAD and FLOW-FA-BAD, which denote the set of the flows consisting of IP packets over Good wireless link, the set of the flows consisting of wireless *data* frames over Good wireless link, the set of the flows consisting of *all* the wireless frames over Good wireless link, and so on. On each set, we define $\{C_{\delta,n}^{(L)}\}$, $\{W_{\delta,n}^{(L)}\}$, $\{A_n^{(L)}\}$ and $\{F_n^{(L)}\}$, where L can be either IP, FD or FA. Then we apply wavelet based multiscale analysis on those processes of each set.

We need to examine additive and multiplicative nature of such categorized flow traffics and compare them to investigate the impacts of lossy wireless links. Marginal distributions of flow byte size can also affect the analysis results, therefore we need to examine and compare them between the flow sets.

3) Modeling Aggregate Traffic: Aggregate traffic can be captured at one BSS or one DS (described in the previous section), which consists of either IP packets or the IEEE 802.11 data frames or the IEEE 802.11 frames (defined previously). Then, we have four available sets of aggregate traffics, i.e., AGGR-IP-BSS, AGGR-FD-BSS, AGGR-FA-BSS and AGGR-IP-DS, which denote the aggregate IP packets at BSS level, the aggregate wireless data frames at BSS level, the aggregate wireless frames at BSS level, and so on. On each traffic, we define $\{C_{\delta,n}^{(L)}\}$, $\{W_{\delta,n}^{(L)}\}$, $\{A_n^{(L)}\}$ and $\{F_n^{(L)}\}$, where L can be either IP, FD or FA. We, then, apply wavelet based multiscale analysis.

We need to examine additive and multiplicative nature, as in [13], [15], [38], of such aggregated traffic (but due to the aggregated nature, additiveness is expected to be stronger than multiplicativeness). First-order (marginal) distributions of packet (or frame) size also need to be examined and compared between the aggregate traffic sets.

Throughout the analysis, we are mainly interested in the impact of wireless MAC protocol and links, and presence of additive nature with well-defined H parameter.

VIII. Conclusion

In this survey paper, we discuss various traffic models, which include traditional models like Markovian model, and recently noticed models such as multifractal model. As we need a proper model for wireless LAN traffic characterization, we describe each model in context of *workload characterization*.

Among such numerous models, we rather focus on self-similar and multifractal models and related wavelet analysis techniques. The reason we think that such models are proper for wireless LAN, is that they can reveal the similarity and difference of wireless LAN traffic from other traffics. Ethernet LAN traffics have been well known to have self-similarity [13], therefore, we can easily conjecture that wireless LAN traffic also has the same characteristic. Moreover, wireless LAN traffic is supposed to be affected by microscopic behaviors of the IEEE 802.11 MAC protocol and lossy wireless links. For this reason, we expect that multifractal analysis can reveal the different small-time-scale behaviors of wireless LAN traffic from other traffics. Wavelet techniques are widely used for their multiscale applicability and the computing efficiency.

In SectionVII, we describe how we can apply the multiscale models and techniques in our wireless LAN setting. There are several complicated issues in measurement on wireless LAN traffic. In this paper, however, we focus on how to model and analyze the traffic, assuming that we have nearly perfect wireless LAN traces, i.e., where nearly all the frames on the air are captured. We expect that by applying the multiscale analysis, we can fully understand and explain the characteristics of wireless LAN traffic. Moreover, we expect that we can accurately characterize the workload for wireless LAN over wide range of time scales and throughout the overall layers, from user sessions to the IEEE 802.11 MAC frames.

REFERENCES

[1] IEEE Computer Society LAN MAN Standards Committee. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. In *IEEE Std 802.11-1999*, 1999.

[2] T.S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall, 2002.

[3] D. Tang and M. Baker Analysis of a Local-Area Wireless Network In *Proc. the Sixth Annual International Conference on Mobile Computing and Networking (MOBICOM 2000)*, Boston, MA, August 2000.

[4] A. Balachandran, G.M. Voelker, P. Bahl and V. Rangan. Characterizing User Behavior and Network Performance in a Public Wireless LAN In *Proc. ACM SIGMETRICS 2002*, Marina Del Rey, CA, June 2002.

[5] D. Kotz and K. Essien. Analysis of a Campus-wide Wireless Network. In *Proc. the Eighth Annual International Conference on Mobile Computing and Networking (MOBICOM 2002)*, Atlanta, GA, September 2002.

[6] J. Yeo, S. Banerjee and A. Agrawala Measuring traffic on the wireless medium: experience and pitfalls. Technical Report, CS-TR 4421, Department of Computer Science, University of Maryland, College Park, December 2002. (available at http://www.cs.umd.edu/ jyeo/TR.pdf)

[7] J. Yeo, M. Youssef, A. Agrawala Characterizing the IEEE 802.11 Traffic: The Wireless Side. Technical Report, CS-TR 4570, Department of Computer Science, University of Maryland, College Park, March 2004.

[8] J. Potemans, J. Theunis, M. Teughels, E.V. Lil and A.V. Capelle Measuring Self-Similar Wireless Data Traffic for Multimedia Applications. In *Proc. 2001 International Conference on Third Generation Wireless and Beyond*, San Francisco, CA, May 2001.

[9] V.S. Frost and B. Melamed. Traffic Modeling for Telecommunications Networks. In *IEEE Communications Magazine*, March 1994.

[10] A. Adas. Traffic Models in Broadband Networks. In *IEEE Communications Magazine*, July 1997.

[11] B.B. Mandelbrot and J.W.V. Ness. Fractional Brownian Motions, Fractional Noises and Applications. In *SIAM Rev.*, Vol. 10, pp. 422-437, 1968.

[12] B.B. Mandelbrot and J.R. Wallis. Computer Experiments with Fractional Gaussian Noises. In *Water Resources Research*, Vol. 5, pp. 228-267, 1969.

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

[14] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger. Long-Range Dependence in Variable-Bit-Rate Video Traffic. In *IEEE Transactions on Communications*, Vol. 43, pp. 1566-1579, 1995.

[15] 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. In *IEEE/ACM Transactions on Networking*, Vol. 5, pp. 77-86, 1997.

[16] V. Paxson and S. Floyd. Wide-area Traffic: the Failure of Poisson Modeling. In *IEEE/ACM Transactions on Networking*, Vol. 3, pp. 226-244, 1995.

[17] M.E. Crovella and A. Bestavros. Self-similarity in World Wide Web Traffic: Evidence and Possible Causes. In *Proc. ACM SIGMETRICS 1996*, Philadelphia, PA, May 1996.

[18] D.P. Heyman and T.V. Lakshman. What are the Implications of Long-Range Dependence for VBR-Video Traffic Engineering? In *IEEE/ACM Transactions on Networking*, Vol 4, No. 3, June 1996.

[19] M. Grossglauser and J. Bolot. On the Relevance of Long-Range Dependence in Network Traffic. In *Proc. ACM SIGCOMM 1996*, Stanford University, CA, August 1996.

[20] B. Ryu and A. Elwalid. The Importance of Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities. In *Proc. ACM SIGCOMM 1996* , Stanford University, CA, August 1996.

[21] M. Krunz and A. Makowski. Modeling Video Traffic Using M—G—Infinity Input Processes: A Compromise between Markovian and LRD Models. In *IEEE Journal on Selected Areas in Communications*, 16(5):733-748, June 1998.

[22] S. Molnar and T.D. Dang. Pitfalls in Long-Range Dependence Testing and Estimation. In *Proc. GLOBECOM 2000*, San Francisco, CA, November 2000.

[23] M. Krunz and I. Matta. Analytical Investigation of the Bias Effect in Variance-Type Estimators for Inference of Long-Range Dependence. In *Computer Networks*, 40(3):445-458, 2002.

[24] B.B. Hubbard. *The World according to Wavelets*. A K Petters, 1995.

[25] I. Daubechies, editor. *Ten Lectures on Wavelets*. S.I.A.M, 1992.

[26] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 2001.

[27] C.S. Burrus, R.A. Gopinath and H. Guo. *Introduction to Wavelets and Wavelet Transforms: Primer*. Prentice Hall, 1998.

[28] G. Erlebacher, M.Y. Hussaini and L.M. Jameson, editor. *Wavelets: Theory and Applications*. Oxford University Press, 1996.

[29] E. Bayraktar, H.V. Poor and K.R. Sircar. Estimating the Fractal Dimension of the S&P 500 Index Using Wavelet Analysis. Department of Electrical Engineering, Princeton University, 2003.

[30] A. Papoulis and S.U. Pillai. *Probability, Random Variables and Stochastic Processes (4th Ed.)*. McGraw-Hill, 2002.

[31] K.S. Trivedi. *Probability and Statistics with Reliability, Queueing and Computer Science Applications (2nd Ed.)*. John Wiley & Sons, 2002.

[32] P. Abry, P. Goncalves and P. Flandrin. Wavelet-Based Spectral Analysis of 1/f Processes. In *Proc. IEEE ICASSP 1993*, Minneapolis, 1993.

[33] P. Abry and D. Veitch. Wavelet Analysis of Long-Range Dependent Traffic. In *IEEE Transactions on Information Theory*, Vol. 44, No. 1, January 1998.

[34] M. Frame, B. Mandelbrot and N. Neger. Fractal Geometry http://classes.yale.edu/fractals, Yale University.

[35] A. Bunde and S. Havlin, editor. Fractals and Disordered Systems Springer-Verlag, 1996.

[36] R.H. Riedi, M.S. Crouse, V.J. Ribeiro and R.G. Baraniuk. A Multifractal Wavelet Model with Application to Network Traffic. In *IEEE Transactions on Information Theory*, Vol. 45, No. 1, April 1999.

[37] C. Stathis and B. Maglaris. Multifractal Experiments with Internet Traffic. In *Proceedings of the 7th IFIP Workshop of Performance Modeling and Evaluation of ATM & IP Networks*, Antwerp, Belgium, June 1999.

[38] A. Feldmann, A.C. Gilbert and W. Willinger. Data Networks as Cascades: Investigating the Multifractal Nature of Internet WAN Traffic. In *Proc. ACM SIGCOMM 1998*, Vancouver, British Columbia, September 1998.

[39] T.G. Kurtz. Limit Theorems for Workload Input Models. In *Stochastic Networks: Theory and Applications: Kelly et al. editors*, Clarendon Press, Oxford, 1996.

[40] A.H Tewfik and M. Kim. Correlation Structure of the Discrete Wavelet Coefficients of Fractional Brownian Motion. In *IEEE Transactions on Information Theory*, Vol. 38, No. 2, 1992.

[41] A.C. Gilbert and W. Willinger. Scaling Analysis of Conservative Cascades, with Applications to Network Traffic. In *IEEE Transactions on Information Theory*, Vol. 45, No. 3, April 1999.

[42] R.H. Riedi and W. Willinger. Toward an Improved Understanding of Network Traffic Dynamics. In *Self-similar Network Traffic and Performance Evaluation*, Wiley, June 2000.

[43] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi and D. Veitch. The Multiscale Nature of Network Traffic: Discovery, Analysis, and Modelling. In *IEEE Signal Processing Magazine*, May 2002.

APPENDIX A
DEFINITION OF RIESZ BASIS

Riesz basis is defined as follows: Suppose we have a *Hilbert* space $\mathcal{H}$ with the inner product $\langle f, g\rangle$, where the norm defined by $|f| = \sqrt{\langle f, f\rangle}$ turns $\mathcal{H}$ into distance-defined space (*metric space*). Suppose we also have a (countably infinite) sequence of vectors $\{\tilde{\phi}_n\} \subset \mathcal{H}$. Define the operator $\tilde{\phi} : l^2 \to \mathcal{H}$ as

$$\tilde{\phi}(\{\alpha_n\}) \equiv \sum_{n=0}^{\infty} \alpha_n \tilde{\phi}_n \tag{47}$$

Then $\{\tilde{\phi}_n\}$ is a Riesz basis of $\mathcal{H}$ if and only if:

1) The series (47) converges for all $\{\alpha_n\} \in l^2$.
2) The operator $\tilde{\phi}$ is bounded.
3) The inverse $\tilde{\phi}^{-1} : \mathcal{H} \to l^2$ exists.
4) The inverse $\tilde{\phi}^{-1}$ is also bounded.