Performance Evaluation 107 (2017) 1–33

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Multifractal and Gaussian fractional sum–difference models for Internet traffic

David Anderson [a], William S. Cleveland [b,*], Bowei Xi [b]

[a] Department of Mathematics, Xavier University of Louisiana, New Orleans, LA, United States
[b] Department of Statistics, Purdue University, West Lafayette, IN, United States

ARTICLE INFO

Article history:
Received 19 January 2016
Received in revised form 28 October 2016
Accepted 5 November 2016
Available online 15 November 2016

Keywords:
FARIMA
Multiplexing
Self-similar
Long-range dependence
Synthetic traffic generation
Network simulation

ABSTRACT

A multifractal fractional sum–difference model (MFSD) is a monotone transformation of a Gaussian fractional sum–difference model (GFSD). The GFSD is the sum of two independent components: a moving sum of length two of discrete fractional Gaussian noise (fGn); and white noise. Internet traffic packet interarrival times are very well modeled by an MFSD in which the marginal distribution is Weibull; this is validated by extensive model checking for 715,665,213 measured arrival times on three Internet links. The simplicity of the model provides a mathematical tractability that results in a foundation for understanding the statistical properties of the arrival process. The current foundation is time scaling: properties of aggregate arrivals in successive equal-length time intervals and how the properties change with the interval length. This scaling is also the basis for the widely discussed multifractal wavelet models. The MFSD provides a more fundamental foundation that is based on how changes in the fGn and white noise components result in changes in the arrival process as various factors change such as the aggregation time length or the traffic packet rate. Logistic models relate the MFSD model parameters to the packet rate, so only the rate needs to be specified in using the MFSD model to generate synthetic packet arrivals for network engineering simulation studies.

1. Background

1.1. Internet technology and network engineering

Internet traffic results from the transfers of information between pairs of computers, or hosts, across the Internet [1–3]. For simplicity we refer to the information as a "file". The file is broken up into packets with sizes typically up to 1460 bytes = 11 680 bits. The packets are sent from the source host over a path consisting of routers connected by transmission links, and the file is reassembled at the destination host. The two hosts establish a connection to carry out the transfer, which means each is listening for the arrival of packets from the other. Headers, typically 40 bytes in size, are added to each packet to manage the file transmission and packet routing. In addition, both hosts can send control packets with no file data, just headers, as part of the transmission management. This means that packet sizes range from 40 bytes to 1500 bytes. Each router has input links and output links; when a packet arrives on an input link, the router reads a field in the header to determine the destination host, and looks in a table to determine the output link over which the packet should be sent to get to the destination.

---

[*] Corresponding author.
E-mail addresses: danders2@xula.edu (D. Anderson), wsc@purdue.edu (W.S. Cleveland), xbw@purdue.edu (B. Xi).

http://dx.doi.org/10.1016/j.peva.2016.11.001
0166-5316/

Each transmission link on the Internet at each point in time can be servicing many ongoing connections. The packet arrival times for transmission on the link are a superposition of the packet arrival times of the individual ongoing connections. Interestingly, the term "superposition" is used in statistics, but in network engineering, the term is "statistical multiplexing". We use the latter here to remind us that this area of Internet research is about statistics. If a packet arrives for transmission and the link is busy transmitting, then the arriving packet is put in a queue. The interface that writes the packet to the link has a speed in bits/second that determines the service time, the packet size in bits divided by the link speed. The queueing is a major factor in quality-of-service (QoS) for Internet connections; if queueing delay is too large, QoS degrades [4].

This work addresses a common type of traffic being carried on most Internet links. The traffic consists of a very wide range of applications such as downloading Web pages, sending email, logging in remotely to a computer, and streaming video or audio. We name this "multi-application traffic". We shall see from the work here that the statistical multiplexing homogenizes the traffic once there are enough connections using a link. This is just a beginning of the well known central limit theorem of point processes: they tend toward Poisson as the number of multiplexed processes increases. The detailed arrival behavior of individual applications is washed out by this, making statistical modeling possible. As we will see, this tends to happen at quite low traffic rates. More precisely we want to model the "offered load" on a router. This is traffic that arrives onto the back-plane of the router from input interface links and is destined for a specific output interface link.

1.2. The need for a validated statistical model for the multiplexed arrival process of the offered load

A statistical model for the offered load can be used to generate traffic for simulations which determine the maximum traffic load on the link that can achieve quality of service (QoS) standards. Packets from the back-plane going to the link are held in a queue to await transmission on the link if the link is busy. Traffic, in bits/second, is written to the link at the link speed, which is also in bits per second. So wait times are in the queue are determined by the relative sizes of the traffic rate and the link speed. The traffic load is chosen to achieve acceptable queue wait-time distributions.

One might think that this traffic engineering would be easy. Just institute a measurement program that collects arrival times and packet sizes of each packet in many streams of packets on many routers for different rates and types of traffic. The problem is that the Internet is not instrumented to do this. Even if it were, the complexity of finding just the right traffic conditions needed in an experiment would be quite hard, and impossible when the conditions do not yet exist. The most practical route for network engineering study is to run computer simulations with packet arrival times as inputs to a queue, and queueing delay as the output.

A simulation for network engineering requires a model for the arrival process. The queueing properties are determined by the statistical properties of the process [5–8]. We take the arrival process to be the sequence of interarrival times, t_u. The interarrivals have very complex statistical properties when studied directly, which in the past has made modeling complex. Three properties account for the complexity—long-range dependence, non-Gaussian behavior, and changing statistical properties with the packet arrival rate $\alpha = 1/\mathrm{E}(t_u)$.

It is critical that a model be valid. Validation must be carried out in substantial detail. In addition, to be most useful for simulation, there needs to be a way to accomplish fast generation of packet arrivals, especially at high packet rates. Interestingly, while there has been much past work in describing the statistical properties of the arrival process, cited in coming sections, there is still not a validated model that provides fast generation. The barrier has been the complexity due to the long-range dependence, nonlinearity, and changing statistics with the arrival rate. This paper contains a very substantial amount of validation. Not only do we look just at the model itself, but rather we drive properties of the model, and then check the derivations by comparing with empirical estimation of the properties based on the data.

The long-range dependence was discovered in the 1990s and reported in two pioneering articles [9,10]. Here, we take long-range dependence to mean that as the frequency f goes to zero, the power spectrum increases like f^{-2d} for $0 < d < 0.5$, which means as the lag k gets large, the autocorrelation function decreases like k^{2d-1}. These statistical properties make the arrival process "bursty", in the language of network engineering. Compared with Poisson arrivals that have the same arrival rate, the upper tail of queueing delays is longer, and the average amount of traffic that can be put on the link and maintain QoS is less [11–15].

One must, however, treat the notion of burstiness with immense care because there is more to the story. Consider the number of packet arrivals in a fixed interval of time. It has a mean and a standard deviation. Consider the ratio of the standard deviation to the mean, the coefficient of variation. We will see clearly the following properties here. As the traffic rate increases, the long-range dependence remains, but the coefficient gets smaller and smaller. So the traffic remains "bursty" in the technical sense, but eventually is not salient because the traffic gets smoother and smoother. In the early days after the discovery of long-range dependence, this smoothing was not appreciated and led to the wrong engineering concepts for the core of the Internet where traffic rates are high.

As one would expect, the interarrival sequence is non-Gaussian. It is expected because it is a waiting time until an event occurs and is a positive random variable. So we can expect the process to be nonlinear (non-Gaussian) since just the marginal distribution of the process is non-Gaussian.

One other matter must be considered for traffic engineering simulation. The statistical properties of the arrival process changes with the expected number of ongoing connections. The change is not just a change in α, the rate, but rather a profound change in the multivariate distributions of any sequence of m consecutive interarrivals. This is a very general result for point processes [16]. Traffic on a link has a "deterministic" component. By this we simply mean that the expected value