A Bibliographical Guide to Self-Similar Traffic and Performance Modeling for Modern High-Speed Networks

Walter Willinger
Bellcore

Murad S. Taqqu
Boston University

Ashok Erramilli
Bellcore

Abstract
This paper provides a bibliographical guide to researchers and traffic engineers who are interested in self-similar traffic modeling and analysis. It lists some of the most recent network traffic studies and includes surveys and research papers in the areas of data analysis, statistical inference, mathematical modeling, queueing and performance analysis. It also contains references to other areas of applications (e.g., hydrology, economics, geophysics, biology and biophysics) where similar developments have taken place and where numerous results have been obtained that can often be directly applied in the network traffic context. Heavy tailed distributions, their relation to self-similar modeling, and corresponding estimation techniques are also covered in this guide.

A fundamental feature of self-similar or fractal phenomena is that they encompass a wide range of time scales. In the teletraffic literature, the notion of burstiness is often used in this context. Mathematical models that attempt to capture and describe self-similar, fractal, or bursty phenomena in a parsimonious manner include certain self-similar stochastic processes and appropriately chosen dynamical systems. A common characteristic of these models is that their space-time dynamics is governed parsimoniously by power-law distribution functions (the "Noah Effect") and hyperbolically decaying autocorrelations (the "Joseph Effect"). In sharp contrast, traditional approaches to modeling fractal phenomena typically rely on highly parameterized multilevel hierarchies of conventional models which, in turn, are characterized by distribution and autocorrelation functions that decay exponentially fast.
Although bursty or fractal phenomena have been observed in virtually all branches of science and engineering, and fractal models have been applied with some success in areas such as hydrology, financial economics, and biophysics, they are new to teletraffic theory and represent a recent addition to the already large class of alternative models for describing traffic in packet switched networks. While most applications of fractal models in science and engineering have been based on empirical findings, they have almost exclusively focused on the models’ powerful descriptive capabilities; their engineering implications and analyses have been largely ignored, mainly because fractal models are generally viewed to be very difficult to analyze. In contrast, the success of fractal models in teletraffic theory will only partly depend on how well they describe actual network traffic, but will also depend to a large degree on the ability to use these models in network analysis and control. To this end, this bibliographical guide brings together many of the references that (i) report on real-time traffic measurements from working networks and emphasize the importance of data analytic, statistical inference and mathematical modeling methods for modern traffic and performance modeling, (ii) demonstrate how traditional modeling approaches, applied with great success in conventional telephony (i.e., circuit switching), have dealt with the increasingly bursty behavior of traffic resulting from modern packet communications, and (iii) illustrate the emergence and impact of fractal processes in modern high-speed network traffic modeling and performance analysis. Thus, this guide lists some of the most recent network traffic studies, combines the most relevant survey and research papers in the areas of data analysis, statistical inference and mathematical modeling, and contains references to other areas of applications (i.e., hydrology, economics and biophysics) where similar developments have taken place. Special emphasis is given to recent works that take fractal models beyond their descriptive phase and pursue genuine engineering applications in the area of modern high-speed network design, management, and control; e.g., queueing models, performance analysis, and control theory. While we do not claim that this is a comprehensive bibliography (for example, note that we did not attempt to adequately cover the traditional teletraffic literature, and including the very latest developments in this rapidly growing field is nearly impossible), we hope that it will serve as a useful, up-to-date reference for researchers and traffic engineers who are interested in self-similar traffic modeling and analysis.

Historically, traffic modeling has its origins in conventional telephony, and has been based almost exclusively on Poisson (or, more generally, Markovian) assumptions about traffic arrival patterns and on exponential assumptions about resource holding requirements (e.g., [144, 209]). However, the emergence of modern high-speed packet networks combines drastically new and different transmission and switching technologies with dramatically heterogeneous mixtures of services and applications. As a result, packet traffic is generally expected to be more complex or bursty than voice traffic, simply because it is spanning vastly different time scales, from microseconds to seconds and minutes. Traditional traffic modeling has responded to these developments with a steady
supply of novel and increasingly sophisticated stochastic models: fluid flow models [9], Markov-Modulated Poisson processes [180], different variations of packet train models [96, 210, 349, 350], the versatile Markovian arrival processes (MAPs) [316, 317], batched Markovian arrival process models (BMAPs) [277], TES (Transform-Expand-Sample) models [305]. While the development of these and other models has been mainly driven by the desire of maintaining analytic tractability of related queueing and performance problems, the resulting models are almost never judged by how well they fit actual traffic data in a statistical sense [12] (for a critical discussion, see [327, 348, 413]).

While the availability of actual traffic measurements from working packet networks has been a serious problem in the past (see [329, 330], where it is noted that between 1966 and 1987, several thousand papers on queueing problems have been published, but only about 50 on traffic measurement results), more recently, enormous volumes of traffic data from working networks have been collected and made available to researchers: CCSN/SS7 [106, 107], ISDN [122, 304], Ethernet LANs [173–175, 255–260, 293, 378, 409–412], WANs and NSFNet [41, 68, 69, 88, 181, 220, 314, 331–333, 335], and VBR traffic [26, 145, 146, 186, 199].

Other traffic measurement studies we are aware of include [346] (Ethernet traffic to a file server), [5] (FASTPAC, an Australian high-speed data network), [59] (DQDB MAN environment), and [78] (World-Wide-Web traffic). Some of these data consist of high-resolution traffic measurements over hours and days/weeks (e.g., [26, 41, 106, 146, 174, 186, 220, 255, 256, 304, 331, 332, 346]), others provide information on coarser time scales over time periods ranging from weeks to months/years (e.g., [69, 333]). The former are typically used for traffic characterization purposes, and the latter yield insight into long-term growth trends and network utilizations. Extensive recent statistical analyses of high-time-resolution traffic measurements reported in [26, 107, 122, 146, 150, 199, 220, 256, 331, 410, 411] have provided convincing evidence that actual traffic data from working packet networks are consistent with statistical self-similarity or fractal characteristics. Moreover, these empirically observed features often distinguish clearly between traffic generated by traditional models and measured data [409, 410]. The main reason for this clear distinction is a subtle difference in the underlying dependence structure; while traditional packet traffic models are short-range dependent (i.e., have exponentially decaying autocorrelations), measured packet traffic data are consistent with long-range dependence (i.e., hyperbolically decaying autocorrelations).

The probability theory of self-similarity and long-range dependence is discussed in [22, 24, 77, 171, 177, 287, 389, 390, 402]. The books [130, 178, 337, 338] also contain large sections on self-similar processes, and extensive bibliographies can be found in [14, 22, 24, 287, 374, 389]. Self-similar stochastic processes were introduced by Kolmogorov [239] in a theoretical context and brought to the attention of probabilists and statisticians by Mandelbrot and his co-workers [287–292]. They have been used in hydrology [200–202, 214, 254, 302, 310–312], geophysics [40, 322], biophysics [143, 267–269], and biology [339, 340]. An area of application where self-similarity and long-range dependence continue to play
a significant role and where many results of practical relevance for traffic engineering have been discovered is economics or, more precisely, financial economics [15, 17, 31, 50, 51, 54, 75, 83, 167, 168, 273, 284, 297, 342, 383]. The paper [14] provides an overview. For an early application of the self-similarity concept and related topics to communications systems, see the seminal paper by Mandelbrot [283]. Enlightening philosophical discussions centering around the issues of traditional mathematical modeling (based on Markov processes) versus unconventional fractal modeling (based on concepts, such as self-similarity or long-range dependence), as well as technical issues related to the problem of stationarity and long-range dependence versus non-stationarity can be found in [103, 104, 205, 219, 246, 266, 303]. Methods for dealing with non-stationarity are developed in [18, 343, 399].

From a modeling viewpoint, the two major families of self-similar time series models are fractional Gaussian noises (i.e., the increment processes of fractional Brownian motion) [24, 288, 289, 374] and fractional ARIMA processes [166, 195, 196], a generalization of the popular ARIMA time series models [33, 37]. Techniques for identifying fractional ARIMA models (also called FARIMA or ARFIMA) are discussed in [23, 310]. Other stochastic approaches to modeling self-similar features are considered in [28, 156, 160, 274, 275, 369, 371] (based on shot-noise processes), [373] (linear models with long-range dependence), [30, 261, 262, 276, 284, 411, 412] (renewal reward processes and their superposition), [116, 283, 401] (renewal processes or “zero-rate” processes), [165] (aggregation of simple short-range dependent models), and [135, 294, 414, 416] (wavelet analysis). Further models are considered in [16, 19, 176, 313, 379, 386, 403, 417]. A radically different approach to modeling self-similar phenomena relies on ideas from the theories of chaos and fractals [73, 97, 118–120, 124, 125, 171, 250, 287, 337, 344, 345, 377]; for a general discussion on chaos, probability and statistics, see [29, 46, 47].

An overview of statistical inference methods for self-similar models and random processes with long-range dependence can be found in [22, 24], the papers [392–394] listing additional techniques. More specifically, R/S analysis is discussed in [18, 24, 26, 28, 130, 200, 258, 272, 273, 286, 288, 290–292, 302, 310, 394] (see also [10, 131]), variance-time analysis in [24, 26, 77, 258, 310, 331, 394, 399] and for spectral domain methods using periodograms, see [24, 26, 48, 84, 140, 149, 157, 159, 183, 203, 204, 206, 249, 253, 353, 357–366, 393, 407, 418].

Examples of new statistical techniques in this area include [3, 7, 20, 21, 25, 27, 52, 53, 57, 58, 80–82, 86, 98, 99, 154, 155, 189, 190, 197, 247, 258, 381, 382, 410, 415]. For a practical evaluation of the different techniques see [392–394]. The paper [76] provides a general overview on the statistical analysis of time series, and references [123, 301, 355] comment on some of the shortcomings of traditional time series analysis in the presence of large sets of traffic measurements. The problem of estimating a linear or polynomial regression when the errors have long-range dependence is considered in [85, 154, 241, 370, 418, 419]. Prediction problems in the context of long-range dependence are addressed in [24, 169, 336, 352].

The theoretical background behind many of these statistical tools is based on central and non-central limit theorems for random sequences with long-range
dependence [34,100,139,141,142,152,153,158,159,161,162,164,182,188,191–193, 280,387,388,395–397]. The proofs require understanding the structure of moments of non-linear functions of Gaussian random variables and linear processes [34,94,158,182,280,385,398]. Some of the results have been extended to random fields, that is, to processes where the "time" parameter is viewed as a "space" parameter and is multidimensional [11,148,179,185,208,278,279,341,375].

Besides the statistical and practical aspects of self-similar or fractal models, there is the ever-present desire for a physical or phenomenological "explanation" for the fractal nature of empirically observed data. For recent work on this topic in the context of high-speed network traffic modeling and how it relates to the infinite variance syndrome or heavy-tailed behavior (the "Noah Effect") of individual mechanisms that are responsible for the self-similarity property at the aggregation level, see [107,220,248,258,326,331,409–412] and related earlier results reported in [77,261,284]. Explaining and validating (with actual data) self-similarity on physical grounds in a network context result in (i) less resistance to non-traditional modeling approaches [77,103], (ii) new insight into the essential characteristics of modern high-speed network traffic, and (iii) novel approaches for dealing with problems related to network traffic management and control. In this context, statistical methods for dealing with heavy-tailed phenomena and appropriate modeling techniques are of crucial importance. Moreover, in terms of modeling, the insight gained into the relationship between the "Noah Effect" exhibited by the individual mechanisms and the "Joseph Effect" observed at the aggregation level provides convincing evidence in favor of applying the principle of parsimony or Ockham's Razor [117,147,212,411–413]; see [92,107] for a particular example involving call holding time modeling for ordinary telephony.

The relevance of heavy-tailed modeling for teletraffic data is also the subject of the recent survey paper [355]. Empirically, the heavy-tailed counterpart of the Gaussian distribution is the stable distribution [126,128,132,420]. While a Gaussian distribution is always symmetric around the mean or median, a stable distribution can be either symmetric or skewed. It has three parameters, $\alpha$, $\beta$, and $\mu$; $\alpha$ characterizes the heaviness of the tail, $\beta$ the skewness and $\mu$ the drift. The stable distribution is symmetric around $\mu$ when $\beta = 0$ and is maximally skewed to the right when $\beta = 1$. Numerical tables of stable distributions can be found in [38,108,194,296,300,324,325,374]. The books [211,374] provide a systematic treatment of stable time series and processes. There are infinite variance counterparts to fractional Gaussian noise [372–374], FARIMA [232,233,235,374] but also pulse-based models [60–62]. Infinite variance models are also of interest in economics and finance [126,127,281,282,297,307–309,347,371]. For random fields with infinite variance, see [225–227,231].

The covariance, which is used to describe the dependence, is not defined when the variance is infinite. Alternatives include the covariation which enters in formulas for regression [43,45,63–67,136,137,374] and the codifference which also characterizes the "ergodic" properties of the time series [13,170,211,230,231,237,321,367,368,374].

Estimation of the exponent $\alpha$ characterizing the heaviness of the distribution
is of central importance and is one of the main themes in [355]. One way of estimating \( \alpha \) is to use regression in a log-log plot [87,91,112,187,244,298,354,380]. Other methods are considered by [39,109–111,129,242,243,295,299,315,355]. Note that even though the variance is infinite one can still use spectral methods to estimate the unknown parameters in a model [92,221–223,228,238]. In particular, the Whittle-type estimator is still applicable [224,234,236,238,306]. One can also use M-estimators [90], in particular estimators based on least absolute deviation [89], and also the bootstrap [93]. For prediction, see [42,44,70,71,229].

While from a statistical viewpoint, the distinction between traditional traffic models and measured network traffic is significant and intriguing, there is also mounting evidence that the empirically observed fractal features of actual traffic (in particular, long-range dependence and heavy-tailed distributions) has practical implications for a wide range of network design and engineering problems. Traditional Markovian (or more general, short-range dependent) input streams to queues are known to impact queueing performance (see for example, [6,104,113,133,134,207,240,263,264,271,351,384,400]), and a range of techniques (e.g., [101,163,251,265,317,405]) are by now available to quantify these impacts and their implications for network management and control. For example, considerable attention has been paid in the recent past to the problem of call admission control in high-speed networks based on the notion of effective bandwidth, e.g., [49,55,95,114,150,151,172,216–218,406].

In contrast to the well developed field of Markovian queueing models, only few theoretical results exist to date for queueing systems with long-range dependent inputs. For recent work in this area and on the general problem of the relevance of fractal traffic in practice, see [4,5,102,105,115–117,121,122,138,146,184,199,258,318–320,328,335,356]. In this context, see also the discussions in [117,318,413] related to some practical experience with the first-generation ATM buffers reported in [79]. While there is considerable scope for future research in the area of queueing models with long-range dependent inputs, queueing in the presence of heavy-tailed service time distributions (and, in general, independent arrivals) is relatively well understood, e.g., see [1,2,56,72,213,323,408]. For specific results relating the behavior of simple queues fed by a single or by many \( ON/OFF \) sources exhibiting heavy-tailed \( ON \) - or \( OFF \)-periods to that of queues with fractional Brownian input streams (as considered in [318]), see for example, [8,35,36,270,344,345].

Given the shortage of theoretical results for long-range dependent queueing models, the ability to generate synthetic traces is of particular importance in the context of teletraffic theory and practice. There exist numerous methods to date for generating self-similar traffic traces. Exact methods, which are based on the Durbin-Levinson algorithm [37,394] are discussed in [24,146,196,198,394]. They are generally impractical for long time series. Approximate methods are described in [30,74,120,215,245,252,258,285,290,331,334,337,356,374,376,404,412]; some of these methods rely on earlier results reported in [77,165,391], derived for a different purpose and re-interpreted here in the context of synthetic traffic generation. These methods are generally very fast and feasible for even
very long time series. However, the statistical quality of the generated sequences is, in general, not well understood [252].

Acknowledgment. This work was partially supported by the NSF Grants DMS-9404093 and NCR-9404931 at Boston University.

Bibliography

30. D. C. Boes. Schemes exhibiting Hurst behavior. In J.N. Srivastava, editor,


94. A. C Davison and D. R. Cox. Some simple properties of sums of random


150. R. J. Gibbens. The statistical analysis of broadband traffic. This volume, 1996.


157. L. Giraitis, A. Samarov, and P. M. Robinson. Rate optimal semiparametric
estimation of the memory parameter of the Gaussian time series with long
range dependence. Technical report, Beiträge für Statistik, Universität

158. L. Giraitis and D. Surgailis. Multivariate Appell polynomials and the
central limit theorem. In E. Eberlein and M. S. Taqqu, editors, Dependence

159. L. Giraitis and D. Surgailis. A central limit theorem for quadratic forms in
strongly dependent linear variables and application to asymptotical nor-
mality of Whittle’s estimate. Probability Theory and Related Fields, 86:87–

160. L. Giraitis and D. Surgailis. On shot noise processes with long range de-
pendence. In B. Grigelionis, Y. U. Prohorov, V. V. Sazonov, and V. Stat-
tulevicius, editors, Probability Theory and Mathematical Statistics. Proceed-
ings of the Fifth Vilnius Conference June 25-July 1, 1989, pages 401-408,

161. L. Giraitis and M. S. Taqqu. Limit theorem for bivariate Appell polyno-

162. L. Giraitis and M. S. Taqqu. Central limit theorems for quadratic forms

163. P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state
tail probabilities in a single-server queue. Journal of Applied Probability,

164. V. V. Gorodetskii. On convergence to semi-stable Gaussian processes.

165. C. W. J. Granger. Long memory relationships and aggregation of dynamic

166. C. W. J. Granger and R. Joyceux. An introduction to long-memory time
series and fractional differencing. Journal of Time Series Analysis, 1:15–30,
1980.


169. G. Gripeenberg and I. Norros. On the prediction of fractional Brownian

170. A. Gross. Some mixing conditions for stationary symmetric stable stochas-
1994.

171. D. Guégan. Séries Chronologiques Non Linéaires à Temps Discret. Statis-
tique Mathématique et Probabilité. Economica, 49, rue Héricart, 75015

172. R. Guerin, H. Admadi, and M. Naghshineh. Equivalent capacity and its


186. D. P. Heyman, A. Tabatabai, and T. V. Lakshman. Statistical analysis and


203. C. M. Hurvich and K. I. Beltrao. Asymptotics for the low-frequency ordi-


236. P. S. Kokoszka and M. S. Taqqu. The asymptotic behavior of quadratic forms in heavy-tailed strongly dependent random variables. *Stochastic


248. T. G. Kurtz. Limit theorems for workload input models. This volume, 1996.


264. S. Q. Li and C. L. Hwang. Queue response to input correlation functions


295. J. H. McCulloch. Simple consistent estimators of stable distribution pa-


309. S. Mittnik and S. T. Rachev. Modeling asset returns with alternative stable


344. P. Pruthi. *An application of chaotic maps to packet traffic modeling*. 
A Bibliographical Guide


391. M. S. Taqqu and J. Levy. Using renewal processes to generate long-range


420. V. M. Zolotarev. *Odnomernye ustoichivye raspredeleniya*. Nauka, Moscow,
In particular, we survey here the development of self-similar traffic and performance modeling over the past 10 years, paying special attention to the radical changes that have resulted from a gradual improvement of our understanding of the causes and origins underlying the self-similarity phenomenon [32], [53], [57], [72]–[74], [77]. This work motivates the development of innovative new approaches for modeling and evaluation of the performance of large-scale complex communication networks such as today’s Internet. In the case of data networks, high time-resolution packet-level traffic measurements are generally recorded from the Self-similar traffic descriptions that capture traffic fluctuations over many time scales of interest are. A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks. Stochastic Networks: Theory and Applications (Oxford), Royal Statistical Society Lecture Notes Series, 4, 339-366. Mandelbrot, B. B. (1965). Self-similar error clusters in communications systems and the concept of conditional systems and the concept of conditional stationarity. IEEE Transactions on Communications Technology, COM, 13, 71-90. Leland, W. E., Taqqu, M. S., Willinger, W., Wilson, D. V. (1994). Computer Simulation of Chaotic Phenomena in High-Speed Communication Networks. Journal of Korean Institute of Information Technology, Vol. 11, 113-122. Hegger, R., Kantz, H., Schreiber, T. The package of TISEAN programs and concomitant documentation. iv SELF-SIMILAR NETWORK TRAFFIC: AN OVERVIEW. The long-held paradigm in the communication and performance communities has been that voice traffic and, by extension, data traffic are adequately described by certain Markovian models (e.g., Poisson) which are amenable to accurate analysis and efficient control. Put differently, physical modeling aims for models of network traffic that relate to the physics of how traffic is generated in an actual network, is capable of explaining empirically observed phenomena such as self-similarity in more elementary terms, and provides new insights into the dynamic nature of the traffic.