Mutifractal Study of Internet Traffic

Z. GÁL, Gy. TERDIK, E. IGLÓI
Center for Informatics and Computing,
University of Debrecen,
Debrecen H-4010, PF. 58,
HUNGARY,
E-mail: {ZGal,Terdik, Igloi}@cic.unideb.hu

Abstract

The self-similar/multifractal nature of Internet traffic has been observed by several measurements and statistical studies. It has not been decided yet whether the traces are following either self-similar or multifractal lows. In this paper we analyzed Internet data traffics, both the classical Bellcore-data and data measured at our MAN by local protocol analyzer. A new multifractal stochastic model were applied and showed its relevance in these cases.

Key-Words: Multifractal, Long-range dependence, Self-similarity, Ethernet, Internet traffic.

1 Introduction

The convergence of the telecom and datacom world onto the infocom era is becoming a reality. New technologies are emerging, government regulations are being relaxed, and the industry is rapidly globalizing. The efficient transport of information is becoming a key element is today's society. By some estimates, bandwidth usage of the Internet is doubling every six to twelve months. For the first time data network capacities are surpassing voice network capacities and the growing demand for network bandwidth is expected to continue in the coming years. Since Internet traffic will more and more dominate traditional telecom traffic, the understanding of its characteristics is crucial for a reasonable design of the future infocom network. Specially three main issues are worth highlighting: server-bound congestion, routing and data flow asymmetry, and the self-similar or multifractal nature of Internet traffic.

It is widely experienced that traffic flows on the Internet are limited by the servers providing data to requests from users, rather than by the network itself. In the presence of large bandwidth it is increasingly likely that the server flow control window will be the dominant control element in traffic throughput rather than today's congestion window. As a result, with larger pipes Internet throughput will be increasingly server-bound even beyond the 50 percent server-bound congestion experienced today. Certainly, servers will increase their flow control window size as they increase in CPU power. The overall network capacity increases faster than the average CPU performance of the most servers. The question of server-bound vs. network-bound will depend on the relative growth of bandwidth vs. CPU power. If bandwidth growth is faster than Moore's law for CPU power and capacity, then ultimately server congestion will be the controlling element in the future networks.

The phenomenon of IP data flow asymmetry has been much observer on both national and international links. Such asymmetry is attributed to larger server farms sending out large amounts of data in response to small requests and to the preponderance of users who download Web pages. Web server farms tend to be clustered near large Internet service points, while users are randomly distributed around the edges of the network. Consequently, near large interconnection points where Web servers are located, there is a large asymmetry in transmit/receive data flows in favor of the transmit path exiting the servers. The main consequence of asymmetry is that a considerable amount of Internet bandwidth is idle, and at the same time the bandwidth on the other side of transmit/receive is totally congested. Such a condition exists because contemporary telecom systems are still designed to primarily support voice traffic.

The self-similar/multifractal nature of Internet traffic has been observed by several measurements and statistical studies starting by Taqqu et al. [6] and a recent one with a god reference list is [3]. A stationary process X_k is multifractal if

$$\log\left|\mathrm{cum}_m\left(X_k^{(n)}\right)\right| = m\left[H\left(m\right)-1\right]\log\left(n\right)+c\left(m\right), \tag{1}$$

i.e. allowing that the Hurst exponent H changes together with the order m. Note that Taqqu [4], considers absolute moments instead of cumulants for the aggregated processes. As far as one concerns on monofractals the two definitions are equivalent. We prefer cumulants because the scaling properties should not change with additive constants and summing up independent processes. The aggregated series is meant by

$$X_k^{(n)} = \sum_{j=1}^{n-1} X_{kn-j}, \quad k\in\mathbb{N}.$$

It was observed that traffic on Internet networks exhibits the same characteristics regardless of the number of simultaneous sessions on a given physical link. At the same time it was pointed out that

- Many signals possess significant LRD, but display short term correlations and behavior inconsistent with strict self-similarity.
- In many signals, the scaling behavior of moments as the signal is aggregated is a nontrivial (nonlinear) function of the moment order.
- Many signals have increments that are inherently positive and hence non-Gaussian.

There is an additional property which is motivated by our experimental study of ATM traces, see Terdik-Gál-Molnár-Iglói [5], providing strong evidence of presence of Gamma distribution and real valued bispectrum.

- Marginal distribution of many signals of ATM traces is close to Gamma.
- Many signals of ATM traces have real valued bispectrum

If we compare Internet traffic vs. Poisson voice traffic for different numbers of aggregated users, it can be seen that as the number of voice flows increases, the traffic becomes more and more smoothed. In other words, the variance of voice traffic rapidly decreases with the increase of flow aggregation. In case of Internet traffic the variance of the process decreases with much lower seed. This property is usually named long-range dependence (LRD) of Internet traffic. We recall that a stationary X_k is **long-range dependent** (LRD) if the autocovariance series

$$r_k \mathrel{\hat{=}} \mathrm{cov}\left(X_i, X_{i+k}\right) \simeq L(k)k^{2h-1}, \quad k\to\infty.$$

Several papers addresses the question of wether for a more accurate description of Internet traffic, a broader model class, namely that a multifractal process, has to be considered [4]. In our paper we propose multifractal process for Internet traffic description.

2 A model for the high speed network traffic

We have managed to construct a multifractal process as a limit, see [2] for detailes. The fact is that there exists an a.s. continuous process J in $\mathcal{C}[0,T]$, it is the Limit of the Integrated Superposition of diffusion processes with linear differencial generator (DLG). It will be called LISDLG process with the following basic properties:

- The LISDLG process $J(t)$ has the following cumulants

  $$\mathrm{cum}_1(J(t)) = 0,$$

  and

  $$\mathrm{cum}_m(J(t)) = (m-1)!\frac{2^{1-2h}}{1-2h}c_0\sigma_0^{2m-2} \int\limits_{[0,1]^m} D_{\tau_0}(\underline{s})^{2h-1}d\underline{s}\; t^{m+2h-1},$$

  for $m\geq 2$, where $h\in(0,1/2)$

  $$D_\tau(\underline{t}) \mathrel{\hat{=}} |t_{i_1}-t_1| + |t_{i_2}-t_{i_1}|\cdots + \left|t_1 - t_{i_{m-1}}\right|.$$

$\tau = (i_1, \ldots, i_{m-1}) \in$Perm$(2,3,\ldots,m)$, $\tau_0 = (2,3,\ldots,m)$. For $m=1$, $D_\tau(t) \doteq 0$

- The LISDLG process $J(t)$ has the same covariance structure as that of the fractional Brownian motion (FBM). Namely, for $t_1, t_2 > 0$,

$$\operatorname{cov}(J(t_1), J(t_2)) = c_g \left(t_1^{2h+1} + t_2^{2h+1} - |t_2 - t_1|^{2h+1} \right),$$

- The discrete time increment process of $J(t)$

$$\Delta J(t) = J(t+1) - J(t), \quad t = 0, 1, 2, \ldots,$$

is stationary and long-range dependent (LRD) with long-range parameter h.

- The $2 \leq m^{th}$ order joint cumulants of the LISDLG process $J(t)$ are

$$\operatorname{cum}(J(t_1), \ldots, J(t_m)) = \text{const}.\sigma_0^{2m}(m-1)! \operatorname*{sym}_{\underline{t}} \left(\int_0^{\underline{t}} D_{\tau_0}(\underline{s})^{2h-1} d\underline{s} \right),$$

- The logarithm of the characteristic functional of the limit process J has the series expansion

$$\log \varphi_J(G) = \frac{2^{1-2h}}{1-2h} \frac{c_0}{\sigma_0^2} \sum_{m=2}^{\infty} \frac{i^m}{m} \sigma_0^{2m} \int_{[0,T]^m} D_{\tau_0}(\underline{t})^{2h-1} \prod_{k=1}^{m} (G(T) - G(t_k))\, d\underline{t},$$

for

$$\operatorname{var}(G) < \frac{1}{\sigma_0^2 T},$$

where τ_0 is the identity permutation and var stands for the total variation.

- The distribution of the LISDLG process $J(t)$ is uniquely determined by its cumulants.

- The increments of the LISDLG process $J(t)$ are stationary and multifractal i.e. the Hust exponent depends on the order of the cumulants. Put the Hurst exponent $H = h + 1/2$ then

$$\operatorname{cum}_m \left(\Delta J(k)^{(n)} \right) = k_1(m) n^{m\left(\frac{2H-2+2m}{m} - 1\right)} \quad (2)$$
$$H(m) = \frac{2H-2}{m} + 2.$$

- The spectrum of the process $\Delta J(t)$ exists and it is given by

$$S_{2,\Delta J}(\omega) = \frac{2\Gamma(2h)\cos(h\pi) c_0 \sigma_0^2}{1-2h} \left| e^{i\omega} - 1 \right|^2 \sum_{k=-\infty}^{\infty} \left| \omega^{(k)} \right|^{-2-2h}, \quad (3)$$

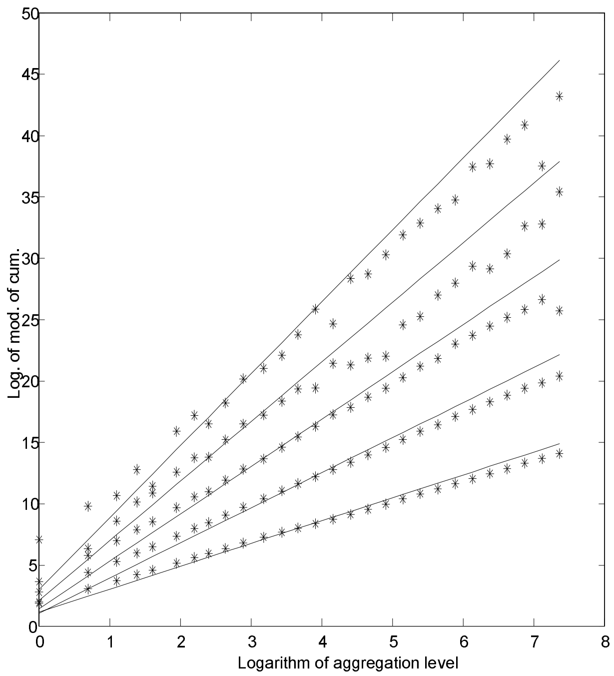
for $\omega \in (0, 2\pi)$, where $\omega^{(k)} \doteq \omega + 2k\pi$, $k \in \mathbb{Z}$. The same as the DFGN spectrum (see Geweke-Porter-Hudak [1]).

- The bispectrum of the process $\Delta J(t)$ exists and it is real valued and nonnegative, i.e.

$$S_{3,\Delta J}\left(\underline{\omega}_{(2)}\right) = \frac{-18 i\pi c_0 \sigma_0^4}{\cos(h\pi)\Gamma(2-2h)} \prod_{j=1}^{3} (1 - e^{i\omega_j}) \quad (4)$$
$$\times \operatorname*{sym}_{\left(\underline{\omega}_{(3)}^{(k,\ell)}\right)} \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} \left[\frac{\left|\omega_1^{(k)}\right|^{1-2h}}{\omega_1^{(k)} \left(\omega_2^{(\ell)} \omega_3^{(k+\ell)} \right)^2} \right],$$

where $\omega_j \in (0, 2\pi)$, $\omega_j^{(k)} \doteq \omega_j + 2k\pi$, $k \in \mathbb{Z}$, $j = 1, 2$ and $\omega_1^{(k)} + \omega_2^{(k)} + \omega_3^{(k)} = 0$.

The discrete time increment process of $J(t)$ is applied for the modeling of different type of high speed Internet traffics.



Plot of modulus of logarithm of cumulants against logarithm of aggregation level, order of cumulants are changing from 2 to 6 starting at bottom, Bellcore trace, see formulae (2), (1).

3 Data and statistics

There were analyzed data sets from two different sources. First data set contains a million packet arrivals seen on an Ethernet at the Bellcore Morristown Research and Engineering facility. The measurement hardware did not drop any packets, but corrupted packets (e.g., Ethernet collisions) are not included in the traces. 99.5% of the encapsulated packets were IP (Internet Protocol). The trace BC-pAug89 began at 11:25 on August 29, 1989, and ran for about 3142.82 seconds (until 1,000,000 packets had been captured). Ethernet, which carried primarily local traffic, but also all traffic between Bellcore and the Internet.

The second data trace was collected exclusively between the University of Debrecen MAN (UDMAN) and the rest of the world. There are more than 5,000 Internet nodes connected to the UDMAN. The number of users is more than 10,000, which implies a high number of parallel network sessions. The data sets contains aggregated Ethernet traffic with one second sampling period. There were sampled frame numbers and byte numbers. The measurement time interval was 260,000 seconds (approxmatelly three days) and the trace began at 17:00 (CET) on October 27, 2000. The measurement period was a week-end when the majority of running network sessions were provided by batch applications (e-mail, batch FTP) and only a relatively small number of interactive network sessions (telnet, irc, etc.) were running by dial-up users. The measurement hardware was a Tekelec protocol analyzer connected to one Fast Ethernet port of the backbone switching router. It is necessary to mention that the traces collected from UDMAN contain data for incoming and outgoing traffic together because of the bus topology of the Ethernet transmission technique.

For both data trace coming from Bellcore and UDMAN, Ethernet protocol forces all packets to have at least a minimum size of 64 bytes and at most the maximum size of 1518 bytes.

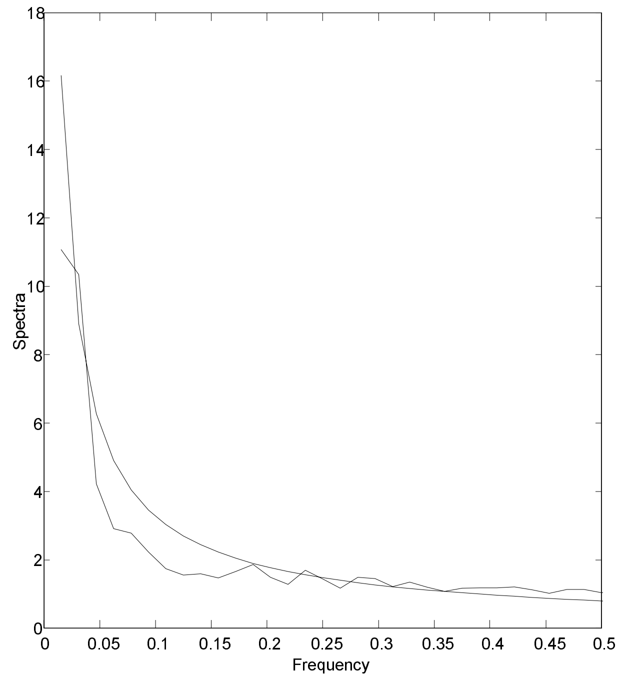


Figure 1: Theoretical and estimated spectra for Bellcore trace, see formula (3).

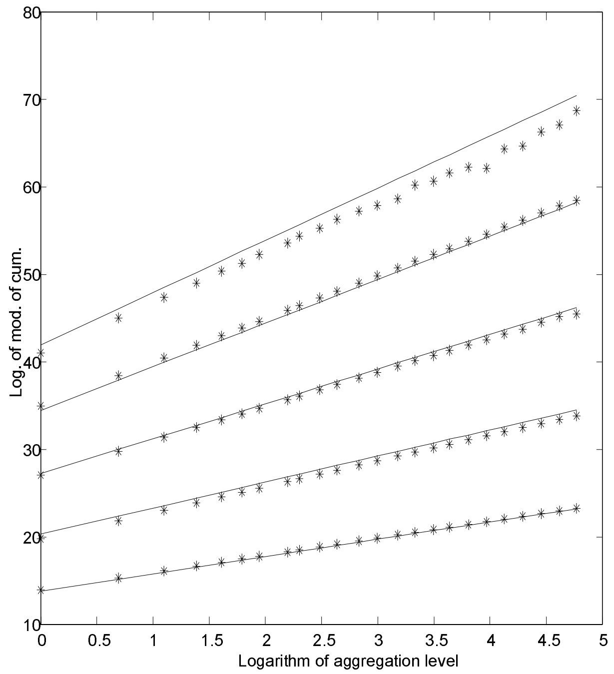
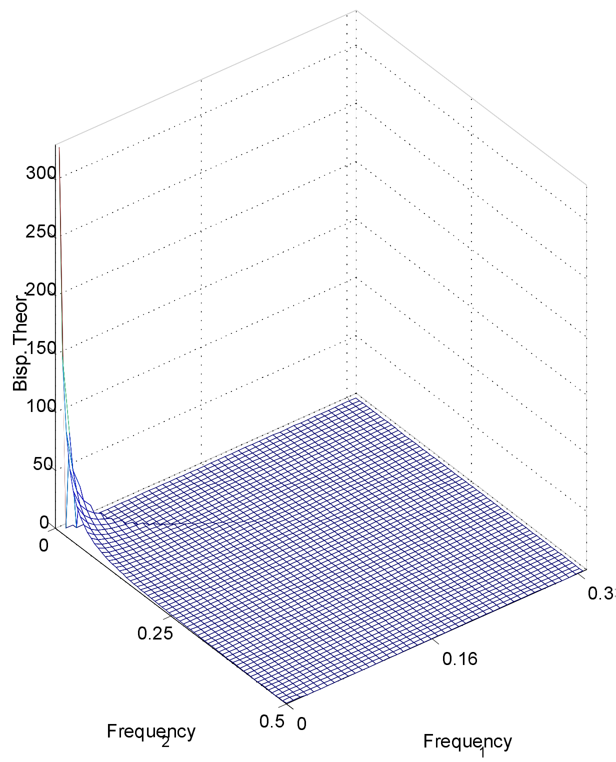


Figure 2: Plot of modulus of logarithm of cumulants against logarithm of aggregation level, order of cumulants are changing from 2 to 6 starting at bottom, UDMAN trace, see formulae (2), (1).



Theoretical bispectrum for the Ethernet frame size of UDMAN-Internet traffic, see formula (4).

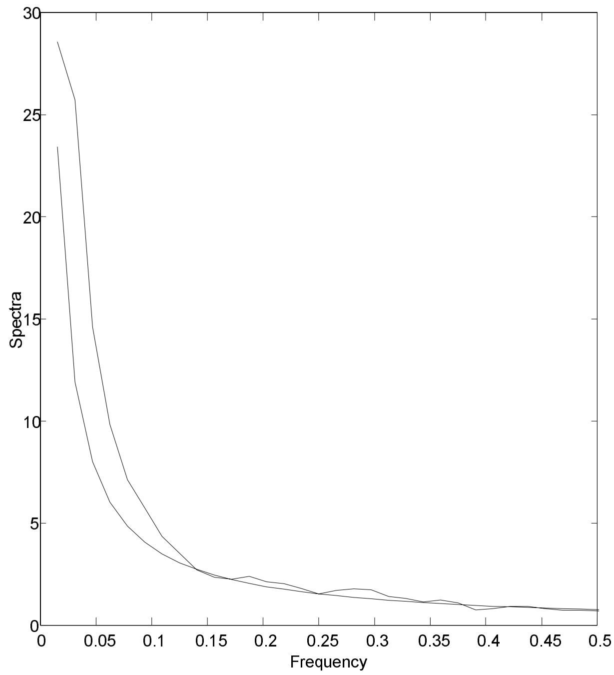
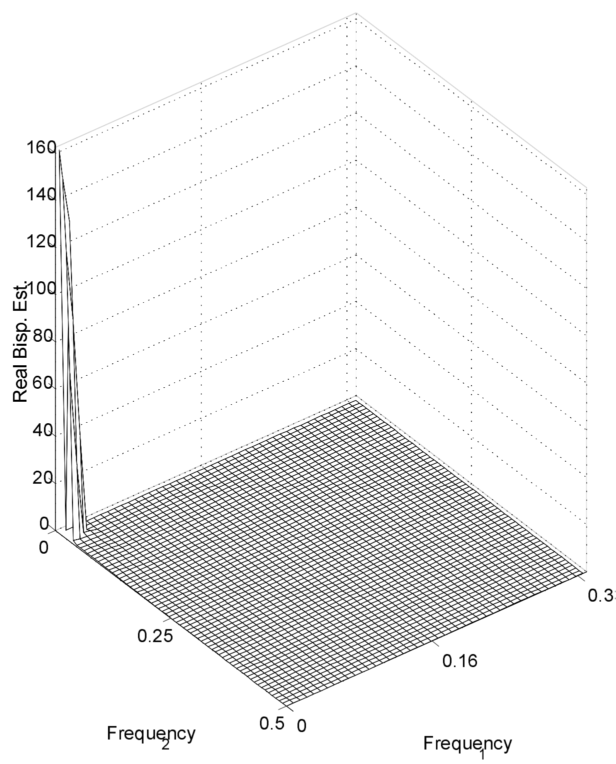
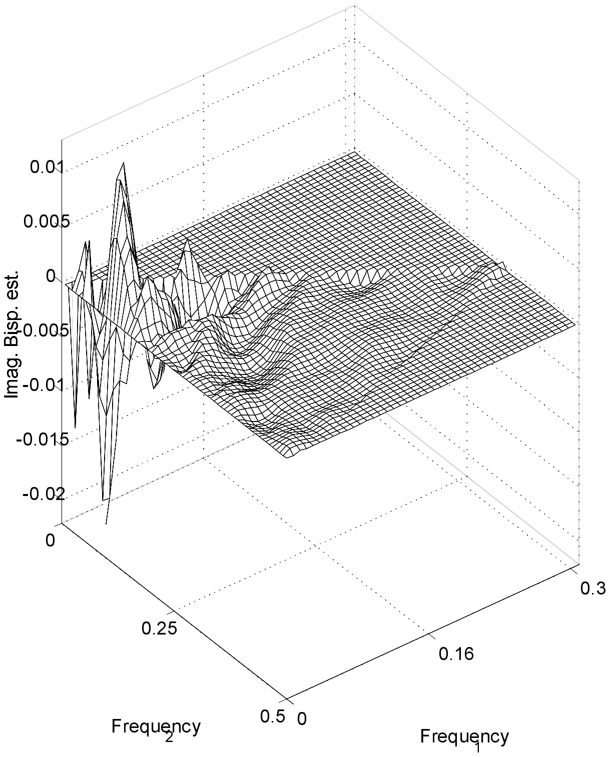


Figure 3: Theoretical and estimated spectra for the Ethernet frame size of UDMAN-Internet traffic, see formula (3).



Estimated Bispectrum of Ethernet frame size for UDMAN-Internet traffic, Real part



Estimated Bispectrum of Ethernet frame size for UDMAN-Internet traffic, Imaginary part

4 Conclusions

This paper gives response to the undecided question regarding self-similar or multifractal nature of the Internet traffic. A new method was applied for different data traces between Bellcore Laboratories and Internet and University of Debrecen MAN and Internet respectively. The results are the same for both data sets which implies clearly the presence of multifractal property against self-similarity of the Internet traffic. At the same time it can be concluded that during more than ten years time the new Internet applications have no radical influence on multifractal nature of the networks.

ACKNOWLEDGMENTS
This research is partially supported by the Hungarian National Science Foundation OTKA No. T 032658.

References

[1] S. Geweke, J.and Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4:221–237, 1983.

[2] E. Iglói and G. Terdik. Superposition of diffusions with linear generator and its multifractal limit process. Technical Report 2000/18(258), Kossuth University of Debrecen, Department of Mathematics and Informatics., KLTE, HU-4010 Debrecen, PF.18, 2000.

[3] H. R. Riedi, S. M. Crouse, J. V. Ribeiro, and G. R. Baraniuk. A multifractal wavelet model with application to network traffic. *IEEE Transactions on Information Theory*, 45(3):992–1018, 1999.

[4] M. S. Taqqu, V. Teverovsky, and W. Willinger. Is network traffic self-similar or multifractal? *Fractals*, 5(1):63–73, 1997.

[5] G. Terdik, Z. Gál, S. Molnár, and E. Iglói. Bispectral analysis of traffic in high speed networks. Technical Report 99/5, Kossuth University of Debrecen, Department of Mathematics and Informatics., KLTE, HU-4010 Debrecen, PF.18, 1999. To appear in SAMS.

[6] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high variability: Statistical analysis of Ethernet LAN traffic at the source level. *Computer Communication Review*, 25:100–113, 1995.