Volume 54 Issue 15   28 October 2010   ISSN 1389-1286

ELSEVIER

Computer Networks

Blind maximum likelihood estimation of traffic matrices under long-range dependent traffic

P.L. Conti [a], L. De Giovanni [b], M. Naldi [c,*]

[a] Universitá di Roma "La Sapienza", Dipartimento di Statistica, Probabilitá e Statistiche Applicate, Piazzale Aldo Moro, Rome, Italy
[b] Universitá LUMSA, Piazza delle Vaschette 101, 00193 Rome, Italy
[c] Universitá di Roma "Tor Vergata", Dipartimento di Informatica, Sistemi e Produzione, Via del Politecnico 1, Rome, Italy

ABSTRACT

A new method, based on the maximum likelihood principle, through the numerical Expectation–Maximization algorithm, is proposed to estimate traffic matrices when traffic exhibits long-range dependence. The methods proposed so far in the literature do not account for long-range dependence. The method proposed in the present paper also provides an estimate of the Hurst parameter. Simulation results show that: (i) the estimate of the traffic matrix is more efficient than those obtained via existing techniques; (ii) the estimation error of the traffic matrix is lower for larger values of the true traffic intensity; (iii) the estimate of the Hurst parameter is slightly negatively biased.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Traffic matrices play a crucial role in network management and provisioning. They describe the amount of bits (packets) transmitted between every Origin–Destination (OD) pair. Their foremost application is in network dimensioning: the traffic matrix provides the load to be carried on the network, i.e., an essential input for the algorithm in charge of sizing transmission links. If compared to other forms of network traffic representation (such as path matrices or measurements on links), traffic matrices have an important advantage, when the nodes considered in the matrix are the original sources and final destinations: they are invariant under changes of either the network topology (excepting node additions or deletions) or routing (see [1]). Traffic matrices can therefore be used to predict link loads after changing routing or network topology. Finally, since they provide a reference picture of the typical traffic on the network, they can be employed in anomaly detection procedures [2].

Direct measurement of the traffic matrix elements is usually not feasible for several reasons. In fact, it would require: (a) the collection of flow statistics on all the routers on the edge of the network, with a direct inspection of each packet header; (b) the shipping of such data to a central processing entity; (c) a processing activity to derive the final destination router for each flow through the analysis of the routing tables [3]. It is customary to overcome this inconvenience by resorting to indirect estimates of the traffic matrix elements via measurements of the traffic on the links. Of course, this requires the knowledge of the routing configuration.

In a network with V nodes there are $N = V(V − 1)$ OD pairs, but only M links, with M considerably smaller than N. Hence, there is a one-to-many mapping relating the (expected) traffic on links to the (expected) OD traffic. In a sense, the information produced by observations on links is not enough to identify OD traffic. This means we are facing an incomplete information estimation problem (or, equivalently, an under-constrained problem).

The approaches to OD traffic matrix estimation under incomplete information are based either on optimization techniques or on statistical inference techniques.

* Corresponding author. Tel.: +39 06 72597269.
  E-mail address: naldi@disp.uniroma2.it (M. Naldi).

Proposals based on optimization techniques rely on the idea of reducing the space of solutions by appropriate constraints on OD traffic (see, for instance, [4,5]).

Proposals based on statistical inference (known as *network tomography* techniques) are based on probabilistic models for OD traffic, and aim at estimating appropriate parameters via either the maximum likelihood method [6–8] or Bayesian methods [9,10]. In [7] a functional mean–variance relationship for OD traffic guarantees identifiability under special assumptions on the network topology.

All the above mentioned papers assume that OD counts are independent Gaussian random variables over OD pairs, and independent and identically distributed (i.i.d.) within an OD pair over successive measurements periods. More formally they are based on the following assumptions:

(i) OD pairs are independent;
(ii) The traffic produced by a single OD pair is a stationary Gaussian process;
(iii) The traffic produced in different time intervals by an OD pair is uncorrelated.

The assumption of independence between OD pairs has been studied for real data in [11], via the correlation between the standardized residuals of the bits (packets) arrival process (bit/packet network traffic) of two different OD pairs at various time aggregation levels. The data at hand are measurements observed on one link in the Finnish university network (Funet), and partitioned into OD traffic on the basis of source and destination IP addresses; time aggregation varies from 1 s to 300 s. The main conclusion is that there is no particular evidence against the assumption of independence among OD pairs.

The assumption of Gaussianity of the OD bits (packets) arrival process is justified by (space/time) aggregation obtained by superimposing independent traffic processes (independent sources) satisfying the usual conditions of the functional central limit theorem (see [12,13]). Such an assumption has been considered in [7,14,15,11,16], and validated via QQ-plots and related correlation tests comparing the empirical distribution with a fitted Gaussian distribution. The data at hand are obtained from the traffic observed on one link, which is partitioned into OD traffic based on source and destination IP addresses (time aggregation from 1 s to 300 s), except in [7] where the data are the traffic observed on the links and OD pairs of a one-router network with 5 min time aggregation, as provided by SNMP. The data are consistent with a Gaussian based modeling approach under suitable (space/time) aggregation.

The assumption of stationarity is studied in [7,14]. In [7] the time-varying nature of network traffic is visually observed in Lucent data (traffic observed on the links and OD pairs of a one-router network with 5 min time aggregation as provided by SNMP). Stationarity of OD traffic (with respect to mean and variance) is assumed to hold in a window lasting up to 21 five minutes time-intervals. In order to take into account the time-varying nature of OD traffic, in [7] it is proposed to estimate the OD traffic (via the Expectation–Maximization algorithm) using a local i.i.d.

model within a moving data window. In [14] the assumption of stationarity is considered via the correlation coefficient between packet network traffic in adjacent periods (time aggregation from milliseconds to seconds). The main conclusion is that stationarity can be reasonably assumed to hold within periods of 30–90 min. The empirical studies performed in [15,11,16] for Funet data essentially confirm the stationarity assumption in [7] (again, a time aggregation from 1 s to 300 s is used).

Assumption (iii) (independence of traffic produced in different time intervals by an OD pair) is considerably more delicate and criticizable. In fact, since the seminal paper [17], a number of studies based on measurements from packet networks have shown that the arrival process of (bits) packets is self-similar, with increments exhibiting time correlation (long-range dependence or long memory).

In this paper we propose a new method to estimate traffic matrices through link measurements. This method explicitly considers the long-range dependent nature of Internet traffic, so far neglected in all the works on traffic matrix estimation. Our method also provides an estimate of the Hurst parameter. We propose a *blind* estimate, i.e., our estimation procedure does not exploit any model either for the traffic intensity values (e.g., the gravity model) or for the mean–variance relationship (e.g., the power-law model), unlike most of the so far proposed algorithms in the short-range dependence context. We review the statistical characteristics of traffic (both the OD traffic and that on the transmission links) in Sections 2 and 3. In the estimation procedure we employ the maximum likelihood principle, applied through the numerical EM (Expectation–Maximization) algorithm, as detailed in Section 3.2. The validation of the method is accomplished through simulation on a toy network, as reported in Sections 4 and 5.

2. The traffic model

As mentioned in Section 1, long-range dependence (LRD, for short) plays a fundamental role in the statistical modelling of teletraffic data. In fact, empirical evidence shows that there is a strong correlation (slowly decaying over time) among such data. A physical explanation for the occurrence of self-similarity, based on weak convergence results for processes that exhibit high variability, is reported in [13]. The key result is that the superposition of many (strictly alternating) i.i.d. (independent and identically distributed) ON/OFF sources with heavy-tailed ON and/or OFF periods produces a Gaussian, self-similar, aggregate traffic with long memory increments. An alternative model to the superposition of ON/OFF sources is the infinite-source Poisson model. Again, the key result in this direction reads as follows: when connections with a single heavy-tailed ON period arrive according to a Poisson distribution, the resulting aggregate traffic exhibits long-range dependence [18–20]. Due to the difficulty of identifying the tail of distributions from limited data, in [21] a generative model for network traffic (MHOP, i.e., Markovian Hierarchical ON–OFF Process) is proposed; it produces long memory traffic without relying on heavy tails. Self-similarity of the simulated traffic fits Bellcore

real data (by means of many statistical techniques), and it is argued that the multiple time-scale nature of traffic coupled with transport protocol aspects makes the presence of long-range dependence unavoidable for the Internet traffic.

In order to explore the self-similarity of the aggregate cumulative arrival process, some studies have analyzed traffic at the level of individual origin (or OD pair) in Local Area Networks (host-to-host) to validate the assumption that ON and/or OFF periods have a heavy-tailed distribution [22,23,19]. Similar studies have been performed for Wide Area Networks, to validate the assumption that session durations possess a heavy-tailed probability distribution [24,19]. Other studies have analyzed directly the cumulative arrival process in Wide Area Networks, to validate the assumption that the corresponding increment process is Gaussian with long-range dependence, through a variety of estimates of the long memory parameter (see [19,25]). The relevant time scales in these works are of the order of a few seconds or less. The autocorrelation of the cumulative aggregate (bits/packet) arrival process at the level of OD pair has been recently analyzed at a time scale either relevant for traffic matrix estimation or suitable to detect long memory (time aggregation varying from 1 s to 300 s). In particular, the presence of long-range dependence of traffic data is shown in [15,11] through the analysis of the correlation between the standardized residuals of the bits (packets) arrival process (bit/packet network traffic) at lag k of an OD pair at various time aggregation levels; in [26], again with a time aggregation for 10 ms to 60 s, through wavelet analysis (as developed in [27]); and in [14] through a visual analysis. All the above mentioned papers show that OD traffic is characterized by the presence of long-range dependence (with values of the Hurst parameter H ranging between 0.65 and 0.9 for aggregation time 1–300 s).

The above remarks justify the introduction of new models based on LRD in the statistical analysis of OD traffic. We consider a network, with N Origin–Destination (OD) pairs. The ends of each OD pair are connected by one of M transmission links. In general, the topology is not a full mesh so that $M < N$.

In the present paper we assume that the traffic generated by a single source can be modelled by a stationary, Gaussian process with LRD. Furthermore, we also assume that sources are independent. As a consequence, the traffic generated by N sources is modelled by a stationary multivariate Gaussian process, with independent long-memory components. Let X_i^t be the traffic for the OD pair i at time t and let $X^t = (X_1^t, X_2^t, \ldots, X_N^t)'$ be the column vector of traffic for all the OD pairs at time t. If we sample the network at T time instants, we have NT realizations of the traffic process. Our assumptions are listed below:

(i) The stochastic process X^t, $t \in \mathbb{N}$ is a stationary Gaussian process, with

$$\mathbb{E}[X_i^t] = \mu_{X_i}, \quad i = 1, 2, \ldots, N,$$
$$\mathbb{V}[X_i^t] = \sigma_{X_i}^2, \quad i = 1, 2, \ldots, N \tag{1}$$

(ii) Different OD pairs generate independent traffic

$$\mathbb{C}[X_i^t, X_j^{t+k}] = 0, \quad i \neq j, \ t \geqslant 1, \ k \geqslant 0 \tag{2}$$

(iii) The auto-covariance function at lag k for any OD pair has the form

$$\mathbb{C}[X_i^t, X_i^{t+k}] = \sigma_{X_i}^2 \rho_X(k), \quad i = 1, \ldots, N, \ t \geqslant 1, \ k \geqslant 0,$$
$$\rho_X(k) = \frac{1}{2}\left[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\right], \tag{3}$$

where $1/2 \leqslant H < 1$ is the Hurst parameter. The basic dichotomy is between $H = 1/2$ (short-range dependence) and $H > 1/2$ (long-range dependence).

(iv) The autocorrelation matrix over T time instants is assumed to be the same for all OD pairs. It will be denoted by

$$R = \begin{pmatrix} 1 & \rho_X(1) & \rho_X(2) & \cdots & \rho_X(T-1) \\ \rho_X(1) & 1 & \rho_X(1) & \cdots & \rho_X(T-2) \\ \rho_X(2) & \rho_X(1) & 1 & \cdots & \rho_X(T-3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_X(T-1) & \rho_X(T-2) & \rho_X(T-3) & \cdots & 1 \end{pmatrix}, \tag{4}$$

(v) The auto-covariance matrix for the ith OD pair is

$$\Sigma_{X_i} = \sigma_{X_i}^2 R. \tag{5}$$

As already mentioned in Section 1, the two assumptions of independence and stationarity are well supported in the literature. We have also assumed that H takes the same value for all the N OD pairs under consideration (*homogeneity hypothesis*). Such an assumption relies on the consideration that the long memory of network traffic is due to the multiple time-scale nature of traffic coupled with transport protocol aspects.

The Gaussian model introduced in the present section is essentially supported by the results in [13], where a special functional central limit theorem is proved. It provides the rational theoretical motivations to use both Gaussianity, long-range dependence, and homogeneity of H, as well. In particular, Theorems 1 and 2 in the above mentioned paper can be interpreted as follows:

(i) Under appropriate assumptions, the mean traffic level provides the main contribution to the observed traffic, and the fluctuations from such level approximately behave like a rescaled fractional Brownian motion.

(ii) If a finite number of independent heterogeneous sources, possibly with different values of the Hurst parameter H, are superimposed, then the term with the highest value of H tends to be dominant. In other words, the application with the highest value of the Hurst parameter determines the value of the Hurst parameter for the aggregate traffic. This point is also raised in [28]. Empirical analyses supporting the homogeneity assumption are in [11,26].

In this context the origin–destination traffic is supposed not to be observable. The only observable quantities are the traffic intensity values measured on the transmission

links. Similarly as above, the (M-dimensional) vector of traffic observed on transmission links is defined as

$$Y^t = (Y_1^t, Y_2^t, \ldots, Y_M^t)'.$$

This vector is related to the OD traffic by the relationship

$$Y^t = A \cdot X^t, \tag{6}$$

where the matrix A, of size $M \times N$, is the routing matrix, whose element a_{kl} is 1 if the kth transmission link carries the traffic pertaining to the lth OD pair (we assume that a single path is used for each OD pair). Being a linear combination of OD traffic components, each link traffic component has an expected value constant over time

$$\mu_Y = E[Y^t] = AE[X^t] = A \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \ldots \\ \mu_{X_N} \end{pmatrix} = A\mu_X. \tag{7}$$

Relationship (6) highlights that each element of Y^t is the sum of some elements of X^t as determined by the routing scheme. Each column of A corresponds to an OD pair and indicates which links are used to carry traffic between that OD pair.

Since the matrix A does have a rank smaller than N (usually M), the model introduced so far is unidentifiable. In order to make it identifiable, a mean–variance relationship is usually assumed in the literature (where only short-range dependence models are considered). In the present paper we do not use such an assumption in the estimation procedure, so that we aim at a *blind* estimate.

The model used in the present paper is essentially based on fractional Brownian motion. As a consequence, the random variables X_i^ts can take negative values. In order to overcome this drawback, a multifractal wavelet model has been proposed in the literature [29]. The basic idea is to take a (discrete) wavelet decomposition of the process $(X_i^t; t \geqslant 0)$, and to directly model the corresponding wavelet coefficients. Among several merits of such an approach, one has to be stressed: it produces random variables X_i^ts that only take positive values. Although interesting, and important as well, such a multifractal model is hardly applicable to the present case. In fact, the statistical estimation of its characteristics requires the availability of a large amount of data. Unfortunately, in realistic applications we only have one measurement per minute, and a total measurement period of one hour and a half (in case of longer measurement periods, stationarity cannot be reasonably assumed). As a consequence, for each link we can only have at most $T = 90$ observations. Such a sample size is insufficient to ensure reliable statistical estimates for complex multifractal models. We also remark that in concrete cases the model adopted in our paper produces negative values with small probability. To justify this statement, we first remark that the expectations μ_{X_i}s and the variances $\sigma_{X_i}^2$s are related by a monotone relationship: the larger the mean μ_{X_i}, the larger the variance $\sigma_{X_i}^2$. More precisely, a power law $\sigma_{X_i}^2 = \gamma\mu_{X_i}^{1/q}$ (with γ a suitable constant) is usually considered, with $q = 1$ as in [6], or $q = 2/3$, as in [16]. As a consequence, denoting by $\mathbb{N}(0, 1)$ a standard normal distribution, we have

$$Pr(X_i^t > 0) = Pr\left(\mathbb{N}(0,1) > -\mu_{X_i}/\sigma_{X_i}\right)$$
$$= Pr\left(\mathbb{N}(0,1) > -\gamma\mu_{X_i}^{(2q-1)/2q}\right) \to 1$$

as μ_{X_i} increases, provided that $q > 1/2$. Since the values of μ_{X_i} occurring in applications are usually large, we may conclude that in concrete cases the probability of negative values on X_i^t is usually negligible.

3. Statistical estimation of traffic parameters

In Section 2 we have shown how LRD traffic is described by the Hurst parameter and by the expected values and variances of the OD pairs. In this Section we concentrate on the estimation of those parameters.

3.1. The likelihood function

As a step towards the definition of the maximum likelihood estimator, we now compute the joint distributions of the statistics of interest. Though the traffic originated by each OD pair is independent of all other OD pairs, it is correlated with the traffic observed on transmission links, namely with those transmission links that carry it. In order to determine such a correlation, we define for each OD pair a column vector Z_i, which gathers both the traffic pertaining to the ith OD pair and the traffic carried by the transmission links:

$$Z_i = \begin{pmatrix} X_i \\ Y^1 \\ \vdots \\ Y^T \end{pmatrix} = \begin{pmatrix} X_i \\ Y \end{pmatrix}. \tag{8}$$

The random vector Z_i possesses a multinormal distribution with $(M + 1)T$ components. Each component of Z_i is correlated with other components. The resulting covariance matrix can be written as

$$M_{Z_i} = \begin{pmatrix} \Sigma_{X_i} & \Sigma_{X_iY} \\ \Sigma_{YX_i} & \Sigma_{YY} \end{pmatrix}, \tag{9}$$

where, using the form $\mathbb{C}(A,B)$ to denote the covariance matrix of vectors A and B, the submatrices in (9) are defined as

$$\begin{aligned} \Sigma_{X_i} &= \mathbb{C}(X_i, X_i), \\ \Sigma_{X_iY} &= \mathbb{C}(X_i, Y), \\ \Sigma_{YX_i} &= \mathbb{C}(Y, X_i) = \Sigma_{X_iY}', \\ \Sigma_{YY} &= \mathbb{C}(Y, Y). \end{aligned} \tag{10}$$

The expressions introduced so far may be reformulated, in order to obtain simpler expressions. Consider first the mixed covariance submatrix, which is equal to

$$\begin{aligned} \Sigma_{X_iY} &= E\left[\left(X_i - \mu_{X_i}\right)(Y - \mu_Y)'\right] \\ &= [P_i^{(1)}A'P_i^{(2)}A' \cdots P_i^TA'], \end{aligned} \tag{11}$$

where $\mu_{X_i} = \mu_{X_i}\mathbf{1}$, $\mathbf{1}$ is a vector composed by 1s, and $P_i^{(s)}$ ($s = 1, 2, \ldots, T$) are the $T \times N$ matrices having columns composed by 0s, except the ith one:

$$P_i^{(s)} = E\left[\left(X_i - \mu_{X_i}\right)\left(X^s - \mu_X\right)'\right]$$

$$= \begin{bmatrix} 0 & \cdots & 0 & \sigma_{X_i}^2 \rho_X(s-1) & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma_{X_i}^2 \rho_X(s-2) & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \sigma_{X_i}^2 \rho_X(s-T) & 0 & \cdots & 0 \end{bmatrix}. \quad (12)$$

The matrix Σ_{X_i} is given by (5). Finally, the covariance matrix of the observables is

$$\Sigma_{YY} = E\left[\left(Y - \mu_Y\right)\left(Y - \mu_Y\right)'\right]$$

$$= \begin{pmatrix} A\Sigma_0 A' & A\Sigma_1 A' & \cdots & A\Sigma_{T-1} A' \\ A\Sigma_1 A' & A\Sigma_0 A' & \cdots & A\Sigma_{T-2} A' \\ \cdots & \cdots & \cdots & \cdots \\ A\Sigma_{T-1} A' & A\Sigma_{T-2} A' & \cdots & A\Sigma_0 A' \end{pmatrix}, \quad (13)$$

where $\Sigma_u = E[(X^k - \mu_X)(X^l - \mu_X)']$ when $|k - l| = u$.

After having shown that Z_i does follow a multinormal distribution with covariance matrix M_{Z_i}, we now turn to the OD traffic statistics. Although they are not directly observable we can use the traffic as measured on the transmission links to update their estimates. In particular, we are interested in the expected value of the OD traffic conditionally on the observables,

$$m_{X_i} \equiv E\left[X_i | Y^1, \ldots, Y^T, \theta\right], \quad (14)$$

and in the covariance matrix,

$$\Sigma_{X_i}^* \equiv E\left[\left(X_i - m_{X_i}\right)\left(X_i - m_{X_i}\right)' | Y^1, \ldots, Y^T, \theta\right], \quad (15)$$

where $\theta = (\mu_{X_1}, \ldots, \mu_{X_N}, \sigma_{X_1}^2, \ldots, \sigma_{X_N}^2, H)$ is the parameter vector, supposed to be known, of the probability density function of the traffic vector. The distribution of X_i, conditionally on the observables, is again multinormal, with

$$m_{X_i} = \mu_{X_i} + \Sigma_{X_i Y} \Sigma_{YY}^{-1} \left(Y - \mu_Y\right), \quad (16)$$

$$\Sigma_{X_i}^* = \Sigma_{X_i} - \Sigma_{X_i Y} \Sigma_{YY}^{-1} \Sigma_{YX_i}. \quad (17)$$

Formulas (16) and (17) are obtained on the basis of the well known results on the conditional distributions of multivariate Gaussian distributions [30].

Our main goal is to simultaneously estimate the (expected) traffic intensity μ_{X_i} for all their OD pairs, as well as the (common) value of the Hurst parameter. Such quantities, as well as the standard deviations σ_{X_i}, are assumed unknown. They are gathered in the (already defined) vector $\theta = (\mu_{X_1}, \ldots, \mu_{X_N}, \sigma_{X_1}^2, \ldots, \sigma_{X_N}^2, H)$. Our aim is to estimate this vector by the maximum likelihood method. Hereafter we provide the expression of the likelihood function.

Since the traffic generated by each OD pair at any given time is independent of the traffic generated by all other OD pairs, we can write the likelihood function in a product form, where each term is the likelihood corresponding to a different OD pair:

$$L_X(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^T \sigma_{X_i}^{2T} |R|}}$$

$$\times \exp\left[-\frac{1}{2\sigma_{X_i}^2} \left(X_i - \mu_{X_i}\right)' R^{-1} \left(X_i - \mu_{X_i}\right)\right]. \quad (18)$$

Next, we can evaluate the likelihood function for the observables, i.e.,

$$L_Y(\theta) = f\left(Y^1, Y^2, \ldots, Y^T; \theta\right).$$

On the basis of the results on linear combinations of Gaussian distributions [30] the likelihood function of Y^t can be obtained by integrating out the likelihood function of X^t w.r.t. the X^t values that satisfy the relationship $Y^t = A \cdot X^t$, in symbols

$$L_Y(\theta) = \int_{X^t : Y^t = AX^t} L_X(\theta). \quad (19)$$

3.2. The expectation–maximization estimation

The likelihood function obtained in Section 3.1 cannot be maximized analytically. We therefore resort to an iterative numerical procedure: the Expectation–Maximization (EM) method. In this section we provide the details of this procedure.

As remarked in Section 3.1, since the traffic generated by each OD pair is independent of the traffic generated by all other OD pairs, we can write the likelihood function in the product form (18). This function can be maximized by the EM algorithm [31,32]. This iterative algorithm provides a refined estimate of θ at each iteration step; namely at the kth step it provides the new estimate

$$\theta_k = \arg\max_\theta \mathcal{L}(\theta_{k-1}, \theta), \quad (20)$$

where we use the logarithmic likelihood function

$$\mathcal{L}(\theta_{k-1}, \theta) = E\left\{\ln[L_X(\theta)] | Y^1, \ldots, Y^T; \theta_{k-1}\right\}. \quad (21)$$

Although the model itself is not identifiable, the EM algorithm (without any mean–variance relationship assumption) still produces maximum likelihood estimates of the parameters of interest [33]. The EM algorithm consists of the following two steps:

(i) Computation of the expected value of the likelihood function conditionally on the observables (expectation step).
(ii) Maximization of the likelihood function w.r.t. the unknown parameters (maximization step).

In what follows, these two steps are described in detail.

3.2.1. The expectation step

From (21) we first obtain the following expression:

$$\mathcal{L}_X(\theta_{k-1}, \theta) = E\left\{\ln[L_X(\theta)] | Y^1, \ldots, Y^T; \theta_{k-1}\right\}$$

$$= \sum_{i=1}^{N} \left[-\frac{T}{2}\ln(2\pi) - T\ln\sigma_{X_i} - \frac{1}{2}\ln|R|\right]$$

$$+ \sum_{i=1}^{N} E\left\{-\frac{1}{2\sigma X_i^2}\left(X_i - \mu_{X_i}\right)' R^{-1}\right.$$

$$\left. \times \left(X_i - \mu_{X_i}\right) | Y^1, \ldots, Y^T; \theta_{k-1}\right\}. \quad (22)$$

Now, if W is a scalar r.v., A, B are two matrices, and $\mathrm{Tr}(\cdot)$ denotes the trace operator, then the two relationships

$$E(W) = E[\mathrm{Tr}(W)] = \mathrm{Tr}[E(W)], \quad (23)$$

$$\mathrm{Tr}(A \cdot B) = \mathrm{Tr}(B \cdot A), \quad (24)$$

hold true. Using (23),(24) in (22), the expression

$$\mathcal{L}_X(\theta_{k-1},\theta) = \sum_{i=1}^{N}\left[-\frac{T}{2}\ln(2\pi) - T\ln\sigma_{X_i} - \frac{1}{2}\ln|R|\right]$$
$$+ \sum_{i=1}^{N} -\frac{1}{2\sigma_{X_i}^2}\mathrm{Tr}\left\{R^{-1}E\left[\left(X_i - \mu_{X_i}\right)\right.\right.$$
$$\left.\left.\times\left(X_i - \mu_{X_i}\right)'|Y^1,Y^2,\ldots,Y^T;\theta_{k-1}\right]\right\} \quad (25)$$

is obtained. Finally, by adding and subtracting $m_i = E[X_i|Y^1,Y^2,\ldots,Y^T,\theta_{k-1}]$ (i.e., the expected value of X_i conditionally on data $(Y^1,Y^2,\ldots,Y^T,\theta_{k-1})$) in the term $(X_i - \mu_{X_i})(X_i - \mu_{X_i})'$ we obtain the relationship

$$\mathcal{L}_X(\theta_{k-1},\theta) = \sum_{i=1}^{N}\left[-\frac{T}{2}\ln(2\pi) - T\ln\sigma_{X_i} - \frac{1}{2}\ln|R|\right]$$
$$- \sum_{i=1}^{N}\frac{1}{2\sigma_{X_i}^2}\mathrm{Tr}\left\{R^{-1}\Sigma_{X_i}^*\right\}$$
$$- \sum_{i=1}^{N}\frac{1}{2\sigma_{X_i}^2}\mathrm{Tr}\left\{\left(m_{X_i} - \mu_{X_i}\right)'R^{-1}\left(m_{X_i} - \mu_{X_i}\right)\right\}. \quad (26)$$

3.2.2. The maximization step

At the generic kth step, the likelihood function $\mathcal{L}_X(\theta_{k-1},\theta)$ has to be maximized w.r.t. the unknown parameter vector θ. The maximization condition is equivalent to minimizing the following sum:

$$\sum_{i=1}^{N}\left\{-\frac{1}{2}\left(m_{X_i} - \mu_{X_i}\right)'\Sigma_{X_i}^{-1}\left(m_{X_i} - \mu_{X_i}\right)\right\}.$$

This quantity is a sum of scalar terms that can be maximized separately. Since X_i is (a segment of) a stationary process, the vector of expected values μ_{X_i}s is composed by identical elements. Then, we can differentiate w.r.t. any component and equate to zero the derivative to obtain the desired updated estimate of traffic intensity:

$$\frac{d}{d\mu_{X_i}}\left[-\frac{1}{2}\left(m_{X_i} - \mu_{X_i}\right)'\Sigma_{X_i}^{-1}\left(m_{X_i} - \mu_{X_i}\right)\right] = 0$$
$$\rightarrow \frac{d}{d\mu_{X_i}}\left[m_{X_i}'\Sigma_{X_i}^{-1}m_{X_i} - 2m_{X_i}'\Sigma_{X_i}^{-1}\mu_{X_i} + \mu_{X_i}'\Sigma_{X_i}^{-1}\mu_{X_i}\right] = 0$$
$$\rightarrow 2m_{X_i}'\Sigma_{X_i}^{-1}1 + 21'\Sigma_{X_i}^{-1}1\mu_{X_i} = 0$$
$$\rightarrow \mu_{X_i} = \frac{m_{X_i}'\Sigma_{X_i}^{-1}1}{1'\Sigma_{X_i}^{-1}1} = \frac{m_{X_i}'R^{-1}1}{1'R^{-1}1} \quad \forall i = 1,\ldots,N. \quad (27)$$

As far as the traffic variance is concerned, recall that the determinant of $\Sigma_{X_i} = \sigma_{X_i}^2 R$ is $|\Sigma_{X_i}| = (\sigma_{X_i}^2)^T|R|$. Since the log likelihood function (26) can be written down as

$$\mathcal{L}_X(\theta_{k-1},\theta) = \sum_{i=1}^{N}\left[-\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma_{X_i}^2 - \frac{1}{2}\ln|R|\right]$$
$$- \sum_{i=1}^{N}\frac{1}{2}\mathrm{Tr}\left\{\Sigma_{X_i}^{-1}\Sigma_{X_i}^*\right\}$$
$$- \sum_{i=1}^{N}\frac{1}{2}\mathrm{Tr}\left\{\left(m_{X_i} - \mu_{X_i}\right)'\Sigma_{X_i}^{-1}\left(m_{X_i} - \mu_{X_i}\right)\right\}$$
$$= \sum_{i=1}^{N}\left[-\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma_{X_i}^2 - \frac{1}{2}\ln|R|\right]$$
$$- \frac{1}{2\sigma_{X_i}^2}\sum_{i=1}^{N}\left[\mathrm{Tr}\left(R^{-1}\Sigma_{X_i}^*\right) - \frac{1}{2}\left(m_{X_i} - \mu_{X_i}\right)'R^{-1}\left(m_{X_i} - \mu_{X_i}\right)\right], \quad (28)$$

by differentiating w.r.t. the traffic variance of the ith OD pair and letting the derivatives be equal to zero, we obtain the updated estimate

$$\frac{\partial\mathcal{L}}{\partial\sigma_{X_i}^2} = -\frac{T}{2\sigma_{X_i}^2} + \frac{1}{2\sigma_{X_i}^4}\left[\mathrm{Tr}\left(R^{-1}\Sigma_{X_i}^*\right)\right.$$
$$\left.+\left(m_{X_i} - \mu_{X_i}\right)'R^{-1}\left(m_{X_i} - \mu_{X_i}\right)\right] = 0$$
$$\rightarrow \sigma_{X_i}^2 = \frac{1}{T}\left[\mathrm{Tr}\left(R^{-1}\Sigma_{X_i}^*\right)\right.$$
$$\left.+\left(m_{X_i} - \mu_{X_i}\right)'R^{-1}\left(m_{X_i} - \mu_{X_i}\right)\right]. \quad (29)$$

Hence, taking into account (27) and (29), the parameters μ_{X_i} and $\sigma_{X_i}^2$ are expressed as a function of H, and, when replaced in the likelihood function (28), they allow us to obtain the optimal value of the Hurst parameter by a univariate maximization procedure. Although we do not provide a formal proof, in all the cases examined the likelihood function is a smooth quasi-parabolic function of the Hurst parameter, with a global maximum (i.e., a unique solution to the maximization problem). In order to perform the maximization task we have employed Brent's method (inverse parabolic interpolation) [34], since it does not require the knowledge of the derivatives of the function to be maximized. Using the expression provided in Section 5.5 in [35], we see that this algorithm cannot require more than $\frac{2}{\log_2[(1+\sqrt{5})/2]}[\log_2(0.5/\varepsilon)]^2$ function evaluations, where ε is the accuracy required, set equal to 10^{-4} in our case. The complete estimation algorithm develops as reported in Algorithm 1.

Algorithm 1. Estimation of traffic matrix

Set an initial estimate for the expected value and the variance of traffic X
Set an initial estimate for the Hurst parameter
repeat
  Compute for each OD pair the autocorrelation matrix R for the current estimate of H, through expression (4)
  Compute the mean value of the OD traffic conditioned on the traffic on links through expression (16)
  Compute the covariance matrix of the OD traffic conditioned on the traffic on links through expression (17)
  Compute the expected value of traffic as a function of H by expression (27)
  Compute the variance of traffic as a function of H by expression (29)
  Find the value of H maximizing the likelihood function (28) by Brent's method → new estimate of H
  Update estimate of expected value of traffic by expression (27)
  Update estimate of variance of traffic by expression (29)
  Update estimate of traffic on links by expression (6)
until Convergence

3.2.3. Convergence of the EM algorithm

A fundamental property of the EM algorithm is that the log likelihood function is a monotone non decreasing function of the iteration step k. The following result holds true [36].

If $\mathcal{L}(\theta_{k+1}, \theta_k)$ is a continuous function of (θ_{k+1}, θ_k), then the limit points of any instance of the EM algorithm are stationary points of $\log(L_Y(\theta))$ and $\log(L_Y(\theta))$ converges monotonically to $\log(L_Y(\theta^*))$ for some θ^* belonging to the set of stationary points of $\log(L_Y(\theta))$. Additionally, if each saddlepoint in the set of stationary points of $\log(L_Y(\theta))$ is not a global maximum of $\mathcal{L}(\theta_k, \theta_{k+1})$ given θ_k, then $\log(L_Y(\theta))$ converges to a local maximum.

The convergence of $\log(L_Y(\theta))$ to a maximum $\log(L_Y(\theta^*))$ does not guarantee the convergence of θ to θ^*. Actually, there could be many values of θ^* providing a maximum.

The continuity of $\mathcal{L}(\theta_{k+1}, \theta_k)$ is a consequence of the multinormal nature of the likelihood function. Further details about the convergence properties of the EM algorithm can be found in [37] (Chapters 3.4 and 3.5 in particular).

4. Simulation model

In order to evaluate the performance of the traffic estimation algorithm presented in the previous sections, we consider a toy network, already used by Vardi in his seminal paper [6]. This network, reported in Fig. 1 where the nodes are labelled by letters and the links by numbers, is made of four nodes (hence 12 origin–destination relationships), and seven unidirectional links. Although the network is of small size, it is a significant reference since it was adopted in the already mentioned seminal paper by Vardi, and can be used for a direct comparison with the performance reported in that paper. The routing matrix (where the OD pairs listed on the columns are sorted in lexicographic order, i.e., $aa, ab, ac, ad, ba, \ldots$) is

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}. \tag{30}$$

For this network we evaluate the performance of our traffic estimation algorithm by simulation. We resort to simulation for two fundamental reasons. First of all, as far as we know, there are no publicly available measurement sets with the time resolution considered in this paper (below 5 min): most data use the typical 5 min interval of SNMP (Simple Network Management Protocol) which would require a long stationarity period to get enough useful measurements. Furthermore, the use of simulated traces allows us to have a tight control on the traffic characteristics, i.e., the true traffic intensity and the Hurst parameter. In a set of real traffic traces the value of the Hurst parameter H is unknown, and must be estimated on the basis of the observed data. If different estimation methods are applied to the same data set, we can only

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$
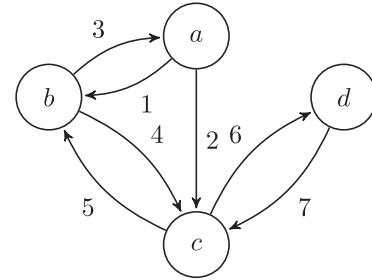

Fig. 1. Toy network.

compute the difference in the obtained estimates, without any information about how close they are to the *true value* of H. Furthermore, we would know the actual traffic intensity only by a separate address analysis conducted on each packet entering the network (an extremely demanding task).

In this section we review the assumptions on the traffic generated as adopted in the simulation analysis.

1. The traffic generated by each node is assumed to follow a long-range dependence process, and the autocorrelation functions obeys the relationship (3).
2. The stochastic processes associated to any two OD pairs are stochastically independent of each other.
3. All the OD pairs have the same value of the Hurst parameter.
4. The expected value of traffic follows a Zipf rank–size relationship.
5. The mean–variance relationship for the traffic intensity follows a power law.
6. The coefficient of preference, which determines how the traffic generated by a given origin node distributes among all the destinations, is proportional to the traffic generated by the destination node.

It is to be noticed that only Assumptions (1)–(3) are actually used in the proposed estimation method. Assumptions (4)–(6) are just used for the purpose of generating the synthetic traffic data employed to feed the estimation algorithm and could be replaced by different ones (although they have been chosen because they are well supported by the literature); they are not exploited in the estimation algorithm.

We now proceed to explain and justify each of the above assumptions.

Assumptions (1) and (2) have been shown in Section 2 to be well supported in the literature.

Assumption (3) is due to the fact that we do not expect the set of applications run by customers connected to a given node to differ significantly from those used by the customers of a different node, or, in other words, nodes (representing routers) are not specialized by service. Though we do not know any thorough work devoted to the relation between the Hurst parameter and the network application, it is known in practice that the feedback behaviour built in the TCP transport protocol (and hence in all the network applications relying on it) is a determinant of long-range dependence. More recently, it has been observed that the appearance of a particular P2P protocol (namely Blubster, a.k.a. Piolet) has been related to a variation of the measured Hurst parameter [26]. Hence, we can expect the Hurst parameter being somewhat determined by the mix of applications generating the observed traffic (actually by the application exhibiting the largest H in the mix, as reported in Section 2). Therefore it should not be different among nodes generating roughly the same application mix.

Assumption (4) means that, if we sort the nodes by their average generated traffic intensity in decreasing order, we expect that the intensity is related to the node rank (rank 1 corresponding the maximum traffic node) by the Zipf law

$$\mu_{O_{(i)}} \propto \frac{1}{i^\alpha}, \quad i = 1 \ldots, V. \tag{31}$$

This law, originally formulated in the context of linguistics [38], is found in many different contexts to describe rank–frequency relationships. The parameter α governs the imbalance of the traffic distribution: the larger α, the larger the differences in the traffic intensity between the highest and the lowest ranked nodes. As far as its applications in a telecommunications environment are concerned, Zipf law is supported by measurements conducted on the telephone network and on Internet users [39].

Assumption (5) is expressed by the relationship

$$\sigma_{X_i}^2 = \lambda \cdot \mu_{X_i}^c, \quad i = 1, \ldots, N. \tag{32}$$

It was put forward in the seminal paper [7] and is supported by several measurements campaigns [40,11,41].

Assumption (6) is common in teletraffic studies, e.g., in Chapter 13 in the reference book [1]. In the context of traffic matrix estimation, it was proposed in a general form, known as the Gravity model, in [42], and adopted in a number of traffic matrix estimation studies, see, e.g., [43,44]. If we indicate by μ_{O_k} the traffic intensity generated by the kth node, we can then associate each OD pair to the endnodes forming that pair. For the generic OD pair with traffic intensity X_i, we can denote its origin node as O_l and its destination node as O_m. The expected traffic intensity for that OD pair is then

$$\mu_{X_i} = \mu_{O_l} \frac{\mu_{O_m}}{\sum_{k=1}^{V} \mu_{O_k}}, \quad i = 1, \ldots, N. \tag{33}$$

As a consequence of Assumptions (4) and (6), the resulting matrix of expected traffic intensities is asymmetric.

Finally, the long-range dependent traffic traces are generated by using the Choleski method [45]. This method, though computationally heavy, is exact and has become the reference method for such a task [46].

5. Simulation results

In this section we provide the results of the simulation-based evaluation. Namely, we report the errors obtained in the estimate of the Hurst parameter and the expected traffic intensity for the OD pairs.

5.1. Simulation parameters

In our simulation we have employed the following set of parameter values:

- H = 0.6, 0.8.
- Zipf parameter α = 1.
- Traffic intensity of the heaviest OD pair = 100 (the unit of measure here is not relevant).
- Number of sample observations (traffic traces length) T = 30, 50, 70, 90.
- Parameter of the power-law relationship between mean and variance λ = 1 and c = 1.5.

The stopping rule for the search of the optimal Hurst parameter has been based on the relative updating step: the search is stopped as soon as the new estimate differs less than 2% from the previous estimate. Although this may somewhat reduce the accuracy of the results, it also reduces the simulation time to reasonable values.

5.2. Hurst parameter estimation

The first quantity of interest is the Hurst parameter. Though this parameter is not the main goal of our estimating task (we are trying to estimate the traffic matrix), the capability to provide reliable estimates of H is an important measure of how well our estimating procedure captures the long-range dependent nature of the process at hand. The results obtained for the Hurst parameter are shown in Table 1. The values reported are the mean values of the estimates as gathered over a block of 500 simulation runs. We can observe a slight underestimation of that parameter in all cases. Estimates are better for small values of the Hurst parameter. However, the bias is progressively reduced when longer traffic traces are used. Trebling the trace length reduces the relative error roughly by a factor of three both when H = 0.6 (from 9.7% to 3.7%) and when H = 0.8 (from 13.1% to 4.4%), as can be seen in Fig. 2.

Table 1
Estimated Hurst parameter.

T	H = 0.6	H = 0.8
30	0.542	0.695
50	0.564	0.738
70	0.576	0.755
90	0.578	0.765
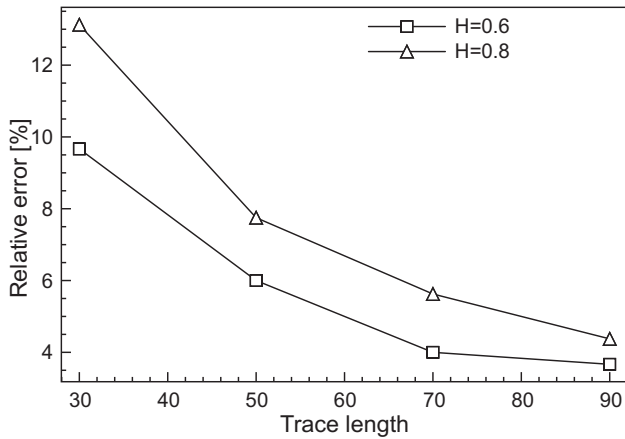
Fig. 2. Estimation error on the Hurst parameter.

Fig. 4. Estimation error on traffic intensity (T = 50).

5.3. Traffic intensity estimation

We now turn to our main goal, i.e., the estimation of the traffic matrix itself. Here we report the relative estimation error for the cases listed in Section 5.1 as resulting from the average of the errors occurring over a block of 500 simulation runs. Rather than reporting the average error over all the OD pairs we perform the analysis for each OD pair separately. In fact, our aim is to obtain small relative errors especially for the largest traffic OD pairs. As a consequence, we focus on the largest traffic OD pairs, e.g., on those making up 90% of the whole traffic [43]. In Figs. 3–6 we report the error as a function of the actual traffic intensity value. The error is first erratic and then decays with the traffic intensity, achieving values lower than 15%. The average error over all the OD pairs is in the range 13.9–21.7%. The OD pairs making up 90% of the overall traffic volume are those whose traffic intensity is above 11.12 (when the traffic intensity of the heaviest OD pair is 100). For those OD pairs the average error is roughly in the same range. Better estimates are in both cases achieved for the smallest values of the Hurst parameter. The trace length seems to have a negligible effect on the quality of estimates. As previously remarked, we do not have a benchmark to compare our results, since no other traffic matrix estimation methods
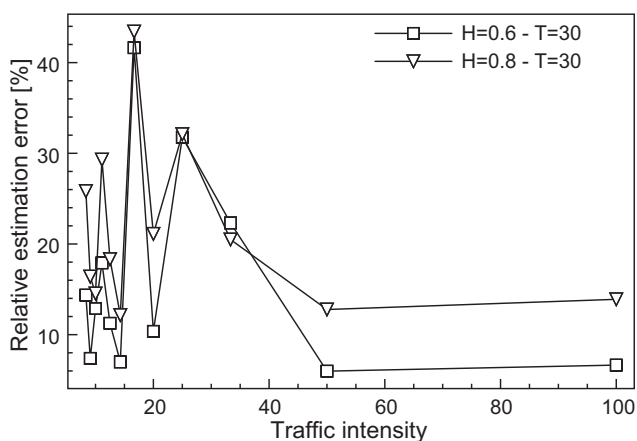

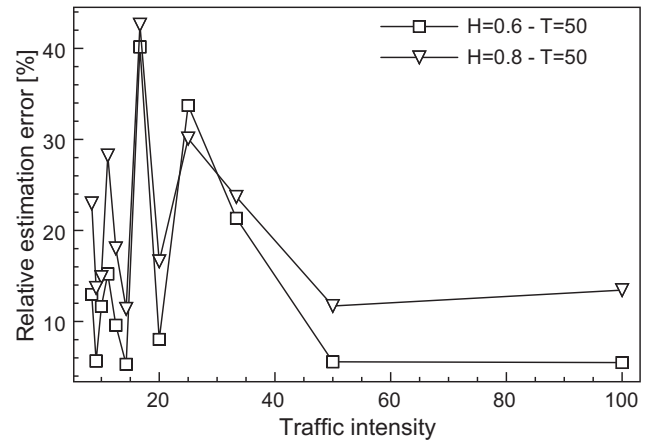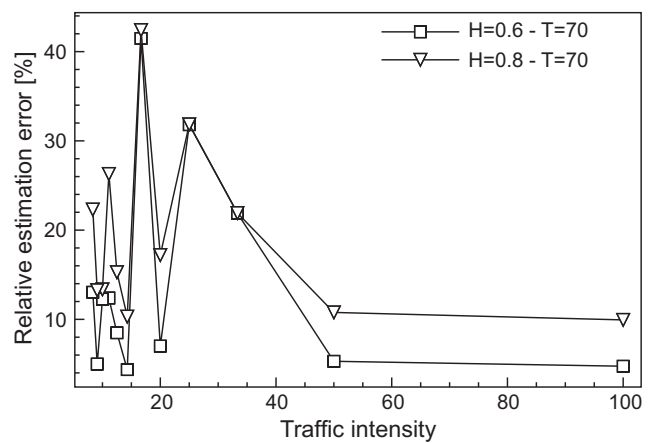**Fig. 5.** Estimation error on traffic intensity (T = 70).

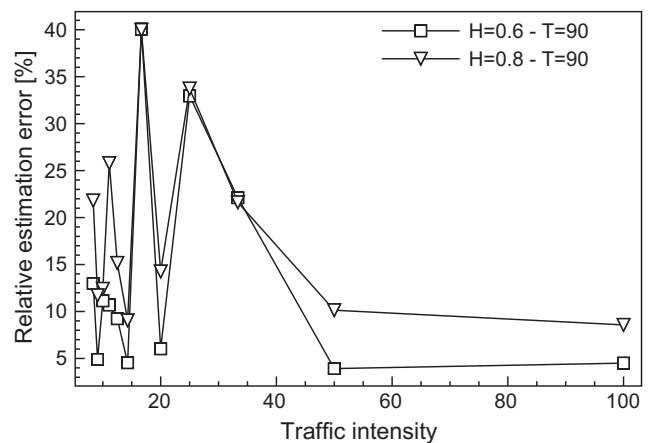**Fig. 6.** Estimation error on traffic intensity (T = 90).

Fig. 3. Estimation error on traffic intensity (T = 30).

have been proposed for LRD traffic. However, the figures reported in [43] for the overall average error in short-range dependent traffic for the three methods considered there (projection method, constrained optimization, and maximum likelihood estimation) lie in the 29–110% range.

In order to perform a comparison with methods that ignore the presence of long-range dependence in the

observed traffic, we have also applied EM algorithm not including the Hurst parameter H to estimate the expected OD traffic for LRD data. In other words, although traffic data are generated by a LRD process, the parameter H (and hence the time-dependence structure) is neglected. The EM algorithm used for this purpose is detailed in Algorithm 2.

Algorithm 2. LRD-unaware estimation of traffic matrix

Set an initial estimate for the expected value and the variance of traffic X
repeat
  Compute for each OD pair the (identity) autocorrelation matrix R (expressions (4), with $H = 0.5$)
  Compute the mean value of the OD traffic conditioned on the traffic on links through expression (16)
  Compute the covariance matrix of the OD traffic conditioned on the traffic on links through expression (17)
  Compute the expected value of traffic by expression (27)
  Find the values of the variances of the OD pairs maximizing the likelihood function (by Nelder and Mead's method [47]) → new estimate of the variances of the OD pairs
  Update estimate of traffic on links by expression (6)
until Convergence

We report the results for the values $H = 0.6$ and $H = 0.8$; in both cases the trace length $T = 30$ is used. On the average our LRD-aware algorithm performs better in both cases, as can be seen in Table 2. Those results also confirm that the estimation is better for small values of the Hurst parameter, roughly by 6 percentage points.

We can have a deeper look at the error, by examining the difference in the estimation error as a function of the true traffic intensity, as shown in Figs. 7 and 8 (the difference is positive when the error of the LRD-unaware algorithm is larger than the error of the LRD-aware algorithm). The difference between the performance of the two algorithms is larger than it could appear through averages, which are not very different from each other. In fact, the LRD-unaware algorithm may perform better for some traffic values but exhibits larger errors on the more relevant OD pairs characterized by large traffic values. Actually, the difference, in favour of the LRD-aware algorithm, may be larger than 2 percentage points when $H = 0.6$ and larger than 6 percentage points when $H = 0.8$.

Finally, we consider the performance of our algorithm, specifically designed to handle the presence of LRD, when
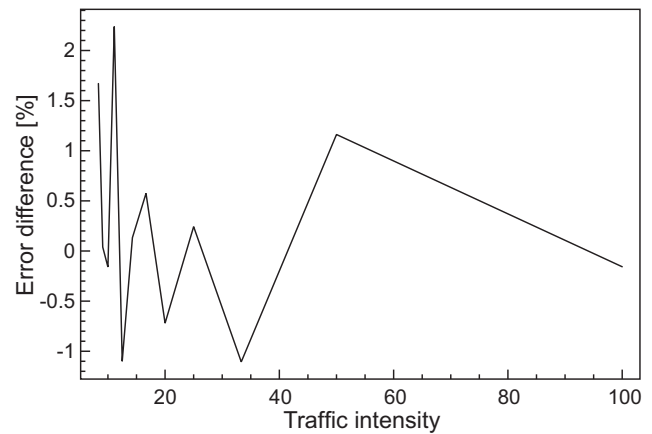
Table 2
Average error on traffic intensity (%).

	$H = 0.6$	$H = 0.8$
LRD-aware algorithm	15.69	21.70
LRD-unaware algorithm	16.03	22.30

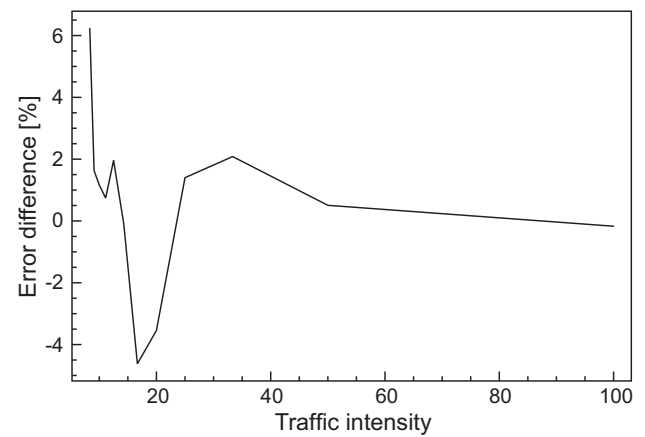**Fig. 7.** Difference in the estimation error on traffic intensity ($H = 0.6$).

Fig. 8. Difference in the estimation error on traffic intensity ($H = 0.8$).
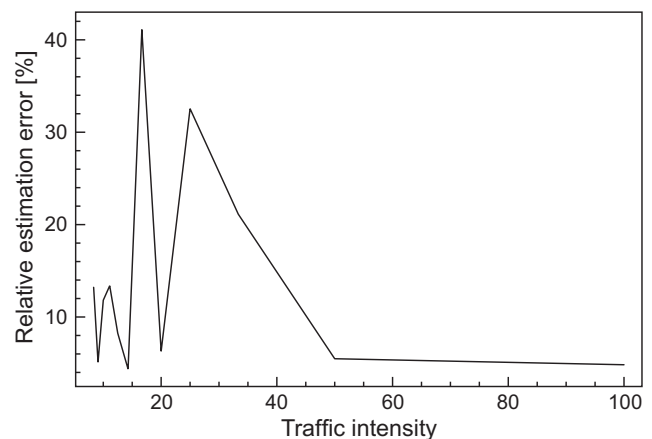
Fig. 9. Estimation error on traffic intensity in the absence of LRD ($T = 30$)

the traffic is instead short range dependent. Namely, we submitted to our algorithm a set of synthetically generated traffic traces with $H = 0.5$. The resulting estimation error is reported in Fig. 9. The average error is 13.96%. Both the average value and the dependence on the OD pair true traffic intensity are similar to the performances observed in the presence of LRD. In addition, the average error on the

estimate of H is quite smaller than the case of actually present LRD, since the variance of the estimates is reduced. We can then conclude that the LRD-aware algorithm performs equally well in the absence of LRD and can therefore be considered as a traffic matrix estimation algorithm of general application.

6. Conclusions

We have proposed a method for the estimation of the traffic matrix in the case of long-range dependent traffic, based on the maximum likelihood principle and on the numerical EM procedure. As a by-product the method also provides an estimate of the Hurst parameter. The method has been applied to the same toy network considered in previous seminal papers. The estimate of the Hurst parameter is slightly negatively biased, but such bias can be reduced to a few percentage points by using longer traffic traces. The estimates of traffic intensity are better for the largest traffic OD pairs, for which the relative error gets below 15%. Both the estimates (of the Hurst parameter and of the traffic intensity) are better for the lowest values of the Hurst parameter. Though no direct benchmarks exist, the errors lie below the range obtained with previous techniques for short-range dependent traffic. The comparison of the proposed LRD-aware estimation algorithm with the corresponding EM-based LRD-unaware algorithm when fed by LRD traffic shows that the LRD-aware algorithm exhibits lower estimation errors, especially on the more relevant OD pairs (those with larger traffic values). In addition, the method here proposed is completely blind, i.e., does not exploit any model either for the traffic intensity values (e.g., the gravity model) or for the mean–variance relationship (e.g., the power-law model), contrary to most of the so far proposed algorithms in the short-range dependence context.

Appendix A. Computational cost

An important issue in the practical application of the EM algorithm to the estimation of the traffic matrix is its computational cost. In this appendix we report an evaluation of the cost of each EM cycle in Algorithm 1. We employ the big O notation, which we recall briefly hereafter. The cost $f(x)$ of an algorithm, expressed as a function of a parameter x is $f(x) = O(g(x))$ ($g(x)$ being a reference, simpler function) if and only if there exists a positive real number L and a real number x_0 such that $|f(x)| \leqslant L|g(x)|$ for all $x > x_0$. In the following, when using results from algorithmics, we will use well established results, though they may provide conservative bounds for the computational cost. Since some matrices may be reused during the course of the algorithm, we assume that the useful matrices are stored for later reuse. This saves

computational resources but employs a trade-off between computational and memory resources. After the evaluation of the computational cost we include an evaluation of the memory resources needed to store the matrices to be reused.

We neglect the initialization steps, since they are performed just once, and consider simply the steps that are iterated in the EM algorithm. Each EM cycle consists of the following three main computational tasks:

1. Maximization of the likelihood function through Brent's algorithm.
2. Updating of the estimate of expected value of traffic.
3. Updating of the estimate of the variance of traffic.

Now we consider each main phase and decompose it down to its most elementary steps. Then we proceed to evaluate the computational cost of each elementary step and to recombine all of them so to arrive at the computational cost of the main phases.

The maximization of the likelihood function, performed through Brent's algorithm, requires at most $O(\log(1/\varepsilon)^2)$ evaluations of the likelihood function, as recalled in Section 3.2.2. We now analyse the cost of each evaluation of the likelihood function. By recalling expression (26), this task consists in turn of the following computational subtasks (we retain just those relevant for the purpose of evaluating the computational cost):

1.1 Computation of the determinant of correlation matrix $|R|$.
    1.1.1 Computation of correlation matrix R as per expression (4).
1.2 Computation of $\mathrm{Tr}\{R^{-1}\Sigma_{X_i}^*\}$.
    1.2.1 Computation of the inverse of correlation matrix R^{-1}.
    1.2.2 Computation of the covariance matrix $\Sigma_{X_i}^*$ conditioned on the observables as per expression (17).
1.3 Computation of $\mathrm{Tr}\{(m_{X_i} - \mu_{X_i})' R^{-1} (m_{X_i} - \mu_{X_i})\}$.
    1.3.1 Computation of $m_{X_i} - \mu_{X_i} = \Sigma_{X_i Y}\Sigma_{YY}^{-1}(Y - \mu_Y)$.
    1.3.2 Transposition of $m_{X_i} - \mu_{X_i} = \Sigma_{X_i Y}\Sigma_{YY}^{-1}(Y - \mu_Y)$.

We start the analysis of the computational cost from the bottom, i.e., from the computational tasks that lie at the lowest level in the decomposition just sketched.

A.1. Task 1.1: Computation of the determinant of the correlation matrix R

The correlation matrix is a square matrix of size T, so that it has T^2 elements, but is a Toeplitz matrix, so that its unique elements are T and its computational cost is $O(T)$. After we have the correlation matrix, we can extract its determinant. By resorting to LU decomposition (the decomposition of a matrix in the product of a lower triangular matrix and an upper triangular one [48]), the pertaining computational cost is $O(T^3)$. Hence, the overall cost of Task 1.1 is $O(T^3)$.

A.2. Task 1.2: Computation of $\text{Tr}\{R^{-1}\Sigma_{X_i}^*\}$

This task involves in turn inverting the correlation matrix first and then obtaining $\Sigma_{X_i}^*$.

Leiserson et al. [49] provide both the naive algorithm and Strassen's algorithm for matrix inversion, whose costs are respectively $O(T^3)$ and $O(T^{2.81})$. In the spirit of considering conservative bounds we assume in the following that the cost of inverting a matrix is represented by the third power of its size.

The computation of the conditional covariance matrix $\Sigma_{X_i}^* = \Sigma_{X_i} - \Sigma_{X_iY}\Sigma_{YY}^{-1}\Sigma_{YX_i}$ is much more complex, since it requires the addition of two terms. The former is $\Sigma_{X_i} = \sigma_{X_i}^2 R$ and, being simply the multiplication of a Toeplitz matrix of size T by a scalar, requires a computational cost $O(T)$. The latter term is the product of three matrices: Σ_{X_iY}, Σ_{YY}^{-1}, and Σ_{YX_i}.

The first of these is Σ_{X_iY}, which requires the computation of T matrices P_i^j (where $j = 1,\ldots,T$) and T multiplications by the matrix A', see expression (11). Each matrix P_i^j is obtained by extracting T values from the correlation matrix, hence its computational cost is $O(T)$. The size of matrix A' (obtained once and for all outside the EM cycle) is $N \times M$. The cost of the product $P_i^j A'$ is given by the three sizes involved, namely $O(TMN)$. Since we have T of these products (and the cost of obtaining the matrices P_i^j is largely dominated by the cost of the products), the cost for each Σ_{X_iY} is $O(T^2MN)$. If we consider the cost for all the N pairs, we end up with a cost $O(T^2MN^2)$.

Now we come to Σ_{YY}^{-1}. The matrix Σ_{YY} is a Toeplitz matrix, of size MT, which requires the computation of T elements of the $A\Sigma_i A'$ kind. The computation of each Σ_i has a cost $O(N)$, since it is a diagonal matrix of size N. Since we have to compute T of them, their overall cost is $O(TN)$. Each product $A\Sigma_i A'$ has a cost $O(MN^2)$, and, since we have to perform again T of them, their overall cost is $O(TMN^2)$. The computation of Σ_{YY} has then a cost $O(TMN^2)$. Finally, the inversion of Σ_{YY} has a cost $O(T^3M^3)$ (the third power of the matrix's size). Since in our case $N > M$, the overall cost of obtaining Σ_{YY}^{-1} is dominated by the cost of obtaining Σ_{YY} in the first place, i.e., $O(TMN^2)$.

The third matrix to be computed is Σ_{YX_i}, which is simply the transpose of the already computed Σ_{X_iY}. The cost of transposition is given by the number of elements of the matrix. Since the size of Σ_{X_iY} is $T \times TM$, the cost of obtaining each Σ_{YX_i} is $O(T^2M)$. The total cost for all the N pairs is then $O(T^2MN)$.

Having the three terms, we can now evaluate the computational cost of the product $\Sigma_{X_iY}\Sigma_{YY}^{-1}\Sigma_{YX_i}$. The multiplication of the first two terms involves a matrix of size $T \times TM$ and a matrix of size $TM \times TM$. The product cost is then $O(T^3M^2)$. Then we multiply a matrix of size $T \times TM$ ($\Sigma_{X_iY}\Sigma_{YY}^{-1}$) by a matrix of size $TM \times T$. The cost of this second product is $O(T^3M)$. The cost of the first product dominates that of the second, so that the overall cost of the three-term-product is $O(T^3M^2)$. Again, the total cost for all the N pairs is $O(T^3M^2N)$.

We have now all the elements needed to get $\Sigma_{X_i}^*$ through the sum $\Sigma_{X_i} - \Sigma_{X_iY}\Sigma_{YY}^{-1}\Sigma_{YX_i}$. This sum involves two square matrices of size T, so that its cost is $O(T^2)$. That task, performed over all the N pairs, has a cost $O(T^2N)$.

Finally, the product $R^{-1}\Sigma_{X_i}^*$ involves two square matrices of size T, and its computational cost is $O(T^2)$, which turns into $O(T^2N)$ when we consider all the N pairs. The trace extraction has a cost $O(T)$, which turns $O(TN)$ over all the N pairs.

The dominant cost for Task 1.2, i.e., over all the operations leading to $\text{Tr}\{R^{-1}\Sigma_{X_i}^*\}$ is then $O(T^3M^2N) + O(T^2MN^2)$, for which we can safely set the upper bound $O(T^3MN^2)$.

A.3. Task 1.3: Computation of $\text{Tr}\{(m_{X_i} - \mu_{X_i})'R^{-1}(m_{X_i} - \mu_{X_i})\}$

By recalling expression (16) we have $m_{X_i} - \mu_{X_i} = \Sigma_{X_iY}\Sigma_{YY}^{-1}(Y - \mu_Y)$. For this task we can assume that the product $\Sigma_{X_iY}\Sigma_{YY}^{-1}$ as well as the inverse R^{-1} are already available, since we used them in Task 1.2. We are left with computing $(Y - \mu_Y)$.

According to expression (7) we have $\mu_Y = A\mu_X$, which is a product of an $M \times N$ matrix and a composite NT-vector. The multiplication requires a cost $O(TMN)$. The sum $(Y - \mu_Y)$ requires instead a cost $O(TM)$.

The product $\Sigma_{X_iY}\Sigma_{YY}^{-1}(Y - \mu_Y)$ involves a matrix of size $T \times TM$ and a vector of size TM, so that its computational cost is $O(T^2M)$.

The transposition of $m_{X_i} - \mu_{X_i}$ involves just the rearrangement of indices for a cost $O(T)$.

Finally, the double product $(m_{X_i} - \mu_{X_i})'R^{-1}(m_{X_i} - \mu_{X_i})$ has a cost $O(T^2)$.

The overall cost of Task 1.3 is then $O(T^2M) + O(TMN)$.

Since the three subtasks of Task 1 have respectively the costs $O(T^3)$, $O(T^3MN^2)$, and $O(T^2M) + O(TMN)$, the overall cost of the maximization of the likelihood function is $O(T^3MN^2 \log(1/\varepsilon)^2)$.

A.4. Task 2: Updating of the estimate of expected value of traffic

We can use expression (27) to evaluate the computational cost of updating the expected value of traffic. The computational tasks associated to such evaluation are the following:

2.1 Transposition of the vector m_{X_i}.
2.2 Multiplication $m_{X_i}R^{-1}$.
2.3 Sum of the elements of $m_{X_i}R^{-1}$.
2.4 Sum of the elements of R^{-1}.

Since m_{X_i} is a vector of T elements, its transposition requires a cost $O(T)$. Its multiplication by a square matrix of size T has a cost $O(T^2)$ and the sum of the resulting product has again a cost $O(T)$. As to Task 2.4, we are summing a matrix of T^2 elements, so that we have a cost $O(T^2)$. Since we have to perform such tasks for each OD pair, all the figures have to be multiplied by N.

The overall cost of updating the estimate of the expected value of traffic is then $O(T^2N)$.

A.5. Task 3: Updating of the estimate of the variance of traffic

In order to estimate $\sigma_{X_i}^2 = \frac{1}{T}[\text{Tr}(R^{-1}\Sigma_{X_i}^*) + (m_{X_i} - \mu_{X_i})'R^{-1}(m_{X_i} - \mu_{X_i})]$ (the traffic variance) we recall expression (27). Here all the basic elements are already available,

Table A.1
Memory requirements for reused matrices.

Matrix	Size	Number	Elements
R	$T \times T$	1	T^2
R^{-1}	$T \times T$	1	T^2
$\Sigma_{X_i Y}$	$T \times MT$	N	$T^2 MN$
Σ_{YY}	$MT \times MT$	1	$T^2 M^2$
$\Sigma_{X_i}^*$	$T \times T$	N	$T^2 N$
m_{X_i}	$T \times 1$	N	TN
μ_{X_i}	$T \times 1$	N	TN
$\sigma_{X_i^2}$	$T \times 1$	N	TN

since we computed them when searching for the optimal value of the Hurst parameter. The operations to be performed are (we neglect those operations that are not significant for the evaluation of the computational cost).

3.1 Computation of the product $R^{-1}\Sigma_{X_i}^*$.
3.2 Extraction of $\mathrm{Tr}(R^{-1}\Sigma_{X_i}^*)$.
3.3 Transposition of $m_{X_i} - \mu_{X_i}$.
3.4 Computation of the product $(m_{X_i} - \mu_{X_i})' R^{-1}(m_{X_i} - \mu_{X_i})$.

Task 3.1 consists in the product of two square matrices of size T, hence it has a cost $\mathrm{O}(T^3)$. The extraction of its trace (Task 3.2) has a cost $\mathrm{O}(T)$, since it consists in summing T elements. The transposition (Task 3.3) has a cost $\mathrm{O}(T)$, since that is the size of the vector to be transposed. The double product involved in Task 3.4 has a cost $\mathrm{O}(T^2)$, since now a vector of size T and a square matrix of size T are involved.

The overall cost of Task 3 is then $\mathrm{O}(T^3)$.

The overall computational cost is then $\mathrm{O}(T^3 M N^2 \log(1/\varepsilon)^2)$, i.e., that of Task 1, since it dominates the costs of the other tasks.

We now consider the memory resources needed to store matrices that may be reused during the execution of the algorithm. The most important example is given by the updating of the estimates of the expected value and the variance of traffic, which require the use of matrices already computed during the execution of the maximization of the likelihood function.

We provide for each of such matrices its size and its number (some matrices have to be computed for each OD pair). Such data are gathered in Table A.1, together with the overall number of elements needed for each matrix. By summing the number of elements over all the matrices we obtain the following number of elements $E = T^2 (2 + M^2 + MN + N) + 3TN$.

References

[1] D. Bear, Principles of telecommunication traffic engineering, IEE/Peter Peregrinus, London, 1988.
[2] P.C. Hernandez, S. Vaton, L. Fillatre, I. Nikiforov, Volume anomaly detection in data networks: an optimal volume detection vs the PCA approach, in: FitraMen'08: International Workshop on Traffic Management and Traffic Engineering for the Future Internet, December 11–12, Porto, Portugal, 2008.
[3] K. Papagiannaki, N. Taft, A. Lakhina, A distributed approach to measure ip traffic matrices, in: IMC '04: Proceedings of the 4th ACM SIGCOMM Conference on Internet measurement, ACM, New York, NY, USA, 2004, pp. 161–174.
[4] O. Goldschmidt, ISP backbone traffic inference methods to support traffic engineering, in: Internet Statistics and Metrics Analysis (ISMA) Workshop, San Diego, CA, December 2000.
[5] I. Juva, S. Vaton, J. Virtamo, Quick traffic matrix estimation based on link count covariances, in: Proceedings of the 2006 IEEE International Conference on Communications (ICC 2006), Istanbul, June 2006, vol. 2, 2006, pp. 603–608.
[6] Y. Vardi, Network tomography: estimating source–destination traffic intensities from link data, Journal of the American Statistical Association 91 (1996) 365–377.
[7] J. Cao, D. Davis, S. Vander Wiel, B. Yu, Time-varying network tomography: router link data, Journal of the American Statistical Association 95 (2000) 1063–1075.
[8] P. Bermolen, S. Vaton, I. Juva, Search for optimality in traffic matrix estimation: a rational approach by Cramer-Rao lower bounds, in: Proceedings of the 2nd Conference on Next Generation Internet Design and Engineering (NGI 2006), Valencia, April 2006.
[9] C. Tebaldi, M. West, Bayesian inference on network traffic using link count data, Journal of the American Statistical Association 83 (1988) 557–573.
[10] S. Vaton, A. Gravey, Iterative Bayesian estimation of network traffic matrices in the case of Bursty flows, in: Internet Measurement Workshop IMW2002, Marseille, 2002. Available from: <http://public.enst-bretagne.fr/vaton/extended-abstract.ps>.
[11] R. Susitaival, I. Juva, M. Peuhkuri, S. Aalto, Characteristics of origin–destination pair traffic in Funet, Telecommunication Systems 33 (2006) 67–88.
[12] I. Norros, P. Pruthi, On the applicability of Gaussian traffic models, in: P.J. Emstad, B.E. Helvik, M.A.H. (Eds.), The Thirteenth Nordic Teletraffic Seminar, Trondheim, August 1996.
[13] M.S. Taqqu, W. Willinger, R. Sherman, Proof of a fundamental result in selfsimilar traffic modeling, Computer Communications Review 27 (1997) 5–23.
[14] I. Norros, J. Kilpi, Testing the Gaussian approximation of aggregate traffic, in: Internet Measurement Workshop IMW2002, Marseille, 2002.
[15] I. Juva, R. Susitaival, M. Peuhkuri, S. Aalto, Traffic characterization for traffic engineering purposes: analysis of Funet data, in: Proceedings of the 1st EuroNGI Conference on Next Generation Internet Networks (NGI 2005), Rome, 2005, pp. 404–422.
[16] I. Juva, R. Susitaival, M. Peuhkuri, S. Aalto, Effects of spatial aggregation on the characteristics of origin–destination pair traffic in Funet, in: Proceedings of the 7th International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN 2007), St. Petersburg, September 2007, Lectures Notes In Computer Science, no. 4712, Springer, Berlin, 2007, pp. 1–12.
[17] W.E. Leland, M.S. Taqqu, W.W.D.V. Willinger, On the self-similar nature of Ethernet traffic (extended version), IEEE/ACM Transactions on Networking 2 (1994) 1–15.
[18] V. Paxson, S. Floyd, Wide area traffic: the failure of Poisson modeling, in: Proceedings of the ACM/SIGCOMM '94, 1994, pp. 257–268.
[19] K. Park, W. Willinger, Selfsimilar Network Traffic and Performance Evaluation, Wiley, New York, 2000.
[20] S. Resnick, G. Samorodnitsky, Steady-state distribution of the buffer content for M/G/∞ input fluid queues, Bernoulli 7 (2001) 191–210.
[21] W.B. Gong, Y. Liu, V. Misra, D. Towsley, Self-similarity and long range dependence on the internet:a second look at the evidence, origin and implications, Computer Networks 48 (2005) 377–399.
[22] W. Willinger, M.S. Taqqu, W.E. Leland, D.V. Wilson, Self similarity in high-speed traffic: analysis and modelling of Ethernet traffic measurements, Statistical Science 10 (1995) 67–85.
[23] W. Willinger, M.S. Taqqu, W.E. Leland, D.V. Wilson, Self similarity and heavy tails: structural modeling of network traffic, in: R.J. Adler, R. Feldman, M.S. Taqqu (Eds.), A Practical Guide to Heavy Tails: Statistical Techniques and Applications, Birkhäuser, Boston, 1998.
[24] M.E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, IEEE/ACM Transactions on Networking 5 (1997) 835–846.
[25] G. Fay, E. Moulines, M.S. François, R. Taqqu, Estimators of long-memory: Fourier versus wavelets, Tech. Rep., 2008. Eprint 0801.4329.
[26] C. Park, F. Hernandez-Campos, J.S. Marron, F. Donelson Smith, Long range dependence in a changing internet traffic mix, Computer Networks 48 (2005) 401–422.
[27] D. Veitch, P. Abry, A wavelet based joint estimator of the parameters of long range dependence traffic, in: IEEE Transactions on Information Theory, Special issue on Multiscale Statistical Signal Analysis and Its Applications 45 (1999) 878–897.

[28] N. Fonseca, G. Mayor, C. Neto, On the equivalent bandwidth of self-similar sources, ACM Transactions on Modeling and Computer Simulation 10 (2000) 104–124.

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

[30] T. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958.

[31] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B (Methodological) 39 (1) (1977) 1–38.

[32] T.K. Moon, The expectation–maximization algorithm, IEEE Signal Processing Magazine 13 (6) (1996) 47–60.

[33] A. Kuk, J. Chan, Three ways of implementing the EM algorithm when parameters are not identifiable, Biometrical Journal 43 (2001) 207–218.

[34] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes in C, second ed., Cambridge University Press, Cambridge, UK, 1992.

[35] R. Brent, Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[36] C. Wu, On the convergence properties of the EM algorithm, Annals of Statistics 11 (1) (1983) 95–103.

[37] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, 1997.

[38] G. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, New York, 1949.

[39] M. Naldi, C. Salaris, Rank-size distribution of teletraffic and customers over a wide area network, European Transactions on Telecommunications 17 (4) (2006) 415–421.

[40] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, C. Diot, Traffic matrix estimation: existing techniques and new directions, in: SIGCOMM02, Pittsburgh, Pennsylvania, USA, 2002.

[41] A. Gunnar, M. Johansson, T. Telkamp, Traffic matrix estimation on a large IP backbone: a comparison on real data, in: IMC '04: Proceedings of the 4th ACM SIGCOMM Conference on Internet measurement, ACM, New York, NY, USA, 2004, pp. 149–160.

[42] J. Kowalski, B. Warfield, Modelling traffic demand between nodes in a telecommunications network, in: Australian Telecommunications and Networks Conference, Sydney, Australia, 1995.

[43] I. Juva, Traffic matrix estimation in the Internet: measurement analysis, estimation methods and applications, Ph.D. in Technology, Helsinki University of Technology, 2007.

[44] M. Roughan, M. Thorup, Y. Zhang, Traffic engineering with estimated traffic matrices, in: IMC '03: Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement, ACM, New York, NY, USA, 2003, pp. 248–258.

[45] P. Hall, B. Jing, S. Lahiri, On the sampling window method for long-range dependent data, Statistica Sinica 8 (4) (1998) 1189–1204.

[46] L.D. Giovanni, M. Naldi, Tests of correlation among wavelet-based estimates for long memory processes, Communications in Statistics – Simulation and Computation 37 (2) (2008) 301–313.

[47] J.A. Nelder, R. Mead, A simplex algorithm for function minimization, Computer Journal 7 (4) (1965) 308–313.

[48] D.C. Lay, Linear Algebra and its Applications, Addison-Wesley/Longman, New York/London, 2000.

[49] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, MIT Press and McGraw-Hill, Cambridge, MA/New York, 2001.

Pier Luigi Conti is a professor of Statistics. After obtaining his Ph.D. in Statistics in 1992, he joined the University of Rome "La Sapienza" as a Researcher. In 1998 he moved to the University of Bologna as Associate Professor, and in 2000 he became Full Professor of Statistics at the University of Rome "La Sapienza". Prof. Conti is author of more than 40 papers published in international journals. His main research interests are related to non-parametric statistics, structures of dependence, and missing data.

Livia De Giovanni received her degree in Statistics from the University of Rome "La Sapienza" in 1988, Italy. From 1988 to 1997 she worked at Telecom Italia, Network Division, Research and Development Department. Since 1997 she joined the University and she is currently Associate Professor in Statistics at the University of Rome LUMSA. Her main research interests include statistical inference in stochastic processes, long memory processes, non-parametric statistics, neural networks and statistical models in telecommunications.

Maurizio Naldi graduated cum laude in 1988 in Electronic Engineering at the University of Palermo and then received his Ph.D. in Telecommunications Engineering from the University of Rome "Tor Vergata". After graduation he pursued an industrial career, first at Selenia as a radar designer (1989–1991), and then in the Network Planning Departments of Italcable (1991–1994), Telecom Italia (1995–1998), and WIND (1998–2000) where he was appointed Head, Traffic Forecasting & Network Cost Evaluation Group. In the 1992–2000 period he was active in the standardization bodies (ETSI and ITU), in particular as Associate Rapporteur for Broadband Traffic Measurements and Models at ITU Study Group 2. Since 2000 he is with the University of Rome at Tor Vergata, where he is now Aggregate Professor.