NOTE: This is a version of an unedited manuscript that was accepted for publication. Please, cite as:
J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. Lopez de Vergara, and S. Lopez-Buedo. "Characterization of the busy-hour traffic of IP networks based on their intrinsic features". COMPUTER NETWORKS, Elsevier, Vol. 55, No. 9, pp. 2111-2125, June 2011.

Characterization of the busy-hour traffic of IP networks based on their intrinsic features

José Luis García-Dorado*, José Alberto Hernández[1], Javier Aracil, Jorge E. López de Vergara, Sergio Lopez-Buedo

High Performance Computing and Networking research group, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Francisco Tomás y Valiente 11, 28007 - Madrid (Spain)

Abstract

Internet Traffic measurements collected during the busy hour constitute a key tool to evaluate the operation of networks under the heaviest-load case scenarios, and further provide a means to network dimensioning and capacity planning. In this light, this study provides a throughout analysis of the busy-hour traffic measurements of an extensive set of universities, regional networks and Internet exchange points collected from the Spanish Research and Education Network, RedIRIS. After showing that the traffic volumes observed in the busy hour over time can be modeled by a white Gaussian process, this work takes one step further and examines the influence of the networks' intrinsic features, mainly population size and access link capacity, on the busy-hour traffic. Well-known statistical methodologies, such as ANOVA and ANCOVA, show that the network size in terms of number of users justifies most of the busy-hour traffic information. We further provide a linear-regression model that adjusts the amount of traffic that each network user contributes to the busy-hour traffic mean values, with a direct application to the problem of link capacity planning of IP networks.

Key words: Internet traffic busy hour, Capacity planning, Bandwidth demands, University access link, ANCOVA, Network intrinsic features.

*Corresponding author

Email address: jl.garcia@uam.es (José Luis García-Dorado)

[1]José Alberto Hernández is at present with the Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid, avda. Universidad 30, E-28911 - Leganés (Madrid, Spain)

1. Introduction

The characterization of Internet Traffic has received much attention from both network operators and the research community over years [1]. Indeed, there has been an intensive research effort in the characterization of the packet and byte counting process at small time-scales (say milliseconds and smaller), giving raise to a number of long-range dependence models [2, 3]. Also, the estimation of Internet bandwidth demands has been a subject of study, either from a long-term [4] or a short-term [5, 6] point of view. While these analyses and models serve to better understand the dynamics of Internet traffic, it turns out that network operators often use a different metric for capacity planning purposes: the total traffic volume observed in a given link during its busiest hour [7, 8]. Obviously, network operators base their capacity planning strategies on worst-case scenarios, that is, on measurements collected when the network is most heavily loaded. This justifies the interest by the research community on studying and characterizing the busy-hour traffic, and its evolution over time. Typically, as noted by the authors in [7], network operators use the following rule of thumb:

$$C = d \cdot M \tag{1}$$

where C is the target link capacity, M represents the bandwidth demand over the link under study, and d is some constant. Clearly $d \geq 1$, and is often much greater than one to provide sufficient capacity C to satisfy the burstiness of the bandwidth demand M.

The goal of this study is to characterize such bandwidth demand M for university access links during the busy hour, and further study the impact of intrinsic features [9] of the universities on such bandwidth demands. Examples of intrinsic features of networks comprise their population (i.e., number of users) and access link capacity, among others. More specifically, this work studies the busy-hour traffic observed in the access links of a numerous set of large-size networks, focusing on its statistical properties and applicability to network dimensioning tasks. To this end, RedIRIS, which is the Spanish National Research and Education Network (NREN), has kindly donated the traffic measurements of the access links to a large number of universities, regional networks and Internet exchange points over a four-month period.

The remainder of this work is organized as follows: Section 2 briefly reviews the state-of-the-art in the field, presents the goals of this work and provides a detailed description of the measurement set under study. Sections 3 and 4 present the core results of this work, which are finally discussed in Section 5. More specifically, Section 3 shows that the busy hour traffic

can be accurately characterized by a pure Gaussian process, i.e., the average traffic volume during the most loaded hour is independent from one day to another, and it is further Gaussian distributed over time. Section 4 goes one step further and examines the influence of intrinsic network features, such as its population size (number of users) and its actual access link capacity, on the mean and variance of such a Gaussian model. After performing an Analysis of Variance (ANOVA) test, it is found that the influence of the link capacity factor is limited, whereas by means of an Analysis of Covariance (ANCOVA) test it is shown that the population size exerts a significant effect. Consequently, in the set of networks under study (which show high capacity over-provisioning), it is only the population size that matters in the characterization of the busy-hour Gaussian process.

2. Preliminaries

First, we present the related work, followed by several definitions, and finally we detail the set of measurements used in this paper.

2.1. Related work and contributions

In spite of its paramount importance for capacity planning and network design purposes, the network research community has paid relatively little attention to the study of the busy-hour traffic observed in network links, on the contrary to Plain Old Telephone System (POTS) designers. In fact, the network research community has addressed the problem of capacity planning by modeling the whole traffic process at different aggregation scales: packet, flow, application and aggregated traffic volumes.

At the packet level, the classical queueing theory has provided a framework for capacity planning, considering Markovian arrivals and service times. However, such assumptions no longer apply in light of the observed self-similar features of Internet traffic [10, 11, 12].

A flow-based approach is proposed in [13] whereby the authors base their capacity planning models on flow metrics. In such work, the authors end up with a model that considers the bandwidth mean and distribution tail of simulated TCP flows. On the downside, such flow-based dimensioning models are hardly feasible in practice and very sensitive to changes in the profiles of flows. The use of aggregated busy-hour traffic values provides a more robust approach to the process of traffic characterization.

The authors in [5] take one step further and propose a hybrid model $\rho + \alpha\sqrt{\rho}$ which considers both aggregated (the network load ρ) and per-flow (by means of α) metrics. Such parameter α is related to some characteristics

of individual measured flows, for instance their size and peak rate. A further refinement to this approach is proposed in [7] where the burstiness of traffic is modeled from the variance of aggregated traffic, rather than following a flow-based approach (parameter α).

In both approaches [5, 7], the estimation of average demands is kept constant throughout the entire analysis, and the authors mostly focus on the burstiness of traffic. However, given the assumption of traffic stationarity (at small time-scales) these approaches are only valid for capacity planning over short periods of time (in the order of few hours or so), which makes them impractical for long term planning purposes (in the range of months, as defined in [14]). In fact, such a stationarity assumption breaks with the well-known fact that traffic patterns follow human behavior [15].

At the application level, the authors in [16] characterize the traffic demand of individual users as a combination of the typical application sessions started by them: web browsing, P2P, Instant Messaging and email. However, such an application-based model requires network operators to correctly identify each application (which is not straightforward [17]). Additionally, this model is extremely sensitive to changes in user request patterns and hardly viable for forecasting purposes.

Finally, the authors in [4] have addressed the problem of capacity planning and network dimensioning by modeling the whole traffic measurement plot. Essentially, they apply Auto-Regressive Integrated Moving Average (ARIMA) models to the measurements collected on attempts to infer future network load values. Such a model is further applied in [18] to characterize the end-to-end traffic demands between each pair of POPs in a backbone network. Interestingly, such work shows that the bandwidth overprovisioning could be lower than usually assumed for a given QoS requirements. For instance, only about 15% of extra bandwidth (that is, $d = 1.15$ parameter of Eq. (1)) is required to ensure less than 3 ms of queuing delay.

Given the large size of measurements involved in such studies (one measurement every five minutes), the authors in [4] firstly aggregate the data to 90-minute intervals and then, they apply wavelets and ANOVA to further reduce the data volume. In contrast, using the busy-hour traffic as an approach to *summarize* the traffic only requires one measurement per day (the throughput value during the most loaded hour) and data preprocessing is barely required. This first simplifies the process of data collection, storage, management, and analysis, and secondly considers the worst case scenario for capacity planning purposes.

The model proposed in this work tries to overcome the above limitations found in the literature. More precisely, our model studies only the traf-

fic volumes during the busy hour over a relatively long period of time (in the range of several months), and finds that such busy-hour traffic characterization is accurately modeled with a Gaussian process. Additionally, the model only requires one aggregated traffic-related value to be collected every day: the busy-hour traffic mean, which makes it more practical and robust (less sensitive to fine-grain measurements). Moreover, the model relates the bandwidth demand results to the intrinsic characteristics of networks, such as its population size, which is novel.

Concerning traffic Gaussianity, the authors in [19, 20] test whether or not aggregated traffic follows a Gaussian distribution at different aggregation levels in terms of number of users and time-scales. Both studies find Gaussian behavior from 5-ms to 5 seconds of time-granularity. It is worth noticing that we are facing a different problem: The Gaussian modeling of the busy-hour traffic over a number of consecutive days, i.e., as an stochastic process, rather than characterizing the aggregated traffic sample itself.

2.2. Definition of busy-hour traffic

Let $A(t)$ be the instantaneous network throughput measured (for instance, in units of Mb/s) on a given access link. Here, t spans a day of throughput measurements, that is, $t \in [0, 24)$ hours. Also, let $H_T(t)$ denote an average throughput metric computed over a given range $[t - \frac{T}{2}, t + \frac{T}{2}]$ of length T, typically one hour:

$$H_T(t) = \frac{1}{T} \int_{t-\frac{T}{2}}^{t+\frac{T}{2}} A(\tau) d\tau \tag{2}$$

According to this, the busy-hour traffic X is the value that maximizes the above equation, i.e.,:

$$\begin{aligned} X &= \max_t H_T(t), \quad t \in [0, 24) \text{ hours} \\ & \quad T = 1 \text{ hour} \end{aligned} \tag{3}$$

which gives the average throughput (in Mb/s) during the busiest hour of a given day, and such value occurs at time t that maximizes $H_T(t)$.

Additionally, let V be the variance of $A(t)$ during the busy hour $[t - \frac{T}{2}, t + \frac{T}{2}]$, that is

$$V = \frac{1}{T} \int_{t-\frac{T}{2}}^{t+\frac{T}{2}} (A(\tau) - X)^2 d\tau \tag{4}$$

Since the traffic is collected in intervals of five minutes the above equations are discretized accordingly.

Thus, for each data unit (either university, regional network or Internet exchange point as explained below), the above equations define the time-series $\{X_i, i = 1, \ldots, N\}$ and $\{V_i, i = 1, \ldots, N\}$ which comprise the average traffic volume observed during the busy hour and its variance for different days $i = 1, \ldots, N$. In addition, it is also interesting to study the time of day t_i at which the traffic busiest hour occurs. This is given by the time-series $\{t_i, i = 1, \ldots, N\}$. Fig. 1 illustrates how these metrics are computed for a given network over three days.

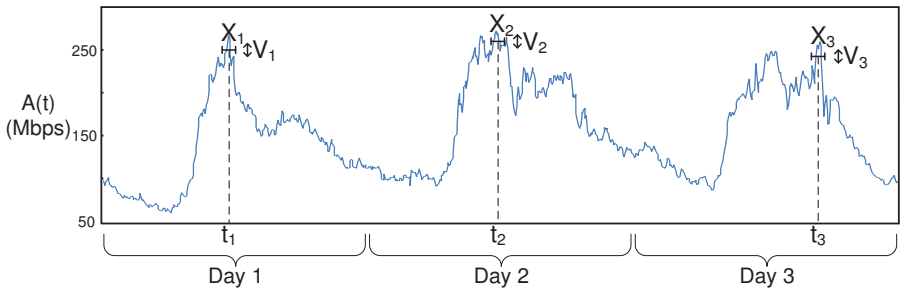

Figure 1: A three-day example of traffic measurements to illustrate X_i, V_i and t_i

Finally, with daily values of X_i and V_i, it is also possible to compute the coefficient of variation CV as defined by:

$$CV_i = \sqrt{\frac{V_i}{X_i^2}}, \quad i = 1, \ldots, N \tag{5}$$

which gives a measure of the variability of the busy-hour traffic volume with respect to its mean.

2.3. Measurement set description

The Spanish National Research and Education Network, RedIRIS, kindly provided the measurements to carry out this study. RedIRIS spans more than 350 institutions, mainly universities and research centers, and it has several Internet exchange points with the European Research and Education Network GEANT, and with other ISPs (Telia, Level3, Cogent, etc.). The traffic capture process lasted four months, ranging from January to April 2009, in which traffic was monitored in both directions of the access link,

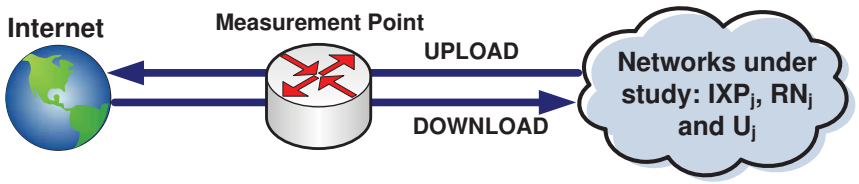

Figure 2: Upload and Download directions of network traffic

say download (from the Internet to the network under study) and upload (sourced on the network under study and destined to the Internet), as shown in Fig. 2.

The traffic trace collected is based on Multi Router Traffic Grapher (MRTG) logs [21] and Cisco's Netflow data [22]. The former comprises a measured sample per five minutes that represents the uploaded and downloaded traffic volume during such a five-minute period of time; the latter comprises the summaries of flow records traversing a given access link, which typically includes the values of bytes transferred, flow starting and finishing times, protocol, etc. As explained in [23, 24], such netflow data provides another means to obtain the uploaded and downloaded traffic volumes, which were shown to validate the traffic values given by the MRTG data. Both these data provide an approximation to the instantaneous network throughput $A(t)$ defined previously, and consequently have been used to calculate the daily busy-hour traffic volume mean X and variance V as stated in Eqs. (3) and (4) for each access link.

After the time-series $\{X_i, i = 1, \ldots, N\}$ for each access link was calculated over different days, it is worth mentioning that the values obtained on both local and bank holidays as well as other non-teaching periods were removed from all time-series. The reason is that only the working-day values are of interest since the network operators work with worst-case scenarios for capacity planning purposes. In addition to this, note that both network upgrades and configuration changes, for instance infrastructure improvements, new filtering policies, new killer applications appraisal, etc. may also have a negative effect on the long/mid term characterization of the busy-hour throughput, yielding to a non-stationary process. Hence, we have carefully removed those networks that have been upgraded or whose configuration have changed in 2009. Such a refined measurement set comprises data from four regional networks (MRTG data), in what follows RN$_j$, which aggregate traffic from several universities, hospitals, computing and research centers; five Internet exchange points (MRTG data), namely IXP$_j$; and 22 university

Networks	Capacity (Mb/s)	Population size (thousand)	max (X_i) (Mb/s)	Max. utilization (Mb/s)
U_1	1000	60	236 / 299	0.25 / 0.30
U_2	1000	57	220 / 244	0.22 / 0.12
U_3	1000	48	137 / 144	0.14 / 0.14
U_4	1000	30	131 / 128	0.13 / 0.13
U_5	1000	28	137 / 131	0.14 / 0.13
U_6	1000	28	129 / 140	0.13 / 0.14
U_7	1000	25	123 / 136	0.12 / 0.14
U_8	1000	14	44 / 75	0.04 / 0.08
U_9	1000	11	50 / 88	0.05 / 0.09
U_{10}	1000	8	38 / 66	0.04 / 0.07
U_{11}	1000	6	16 / 25	0.02 / 0.03
U_{12}	300	58	195 / 175	0.65 / 0.58
U_{13}	200	37	105 / 144	0.53 / 0.72
U_{14}	200	35	140 / 148	0.70 / 0.74
U_{15}	200	20	70 / 88	0.35 / 0.44
U_{16}	200	19	120 / 110	0.60 / 0.55
U_{17}	200	15	64 / 78	0.32 / 0.39
U_{18}	100	14	49 / 74	0.49 / 0.74
U_{19}	100	13	20 / 68	0.20 / 0.68
U_{20}	100	9	30 / 36	0.30 / 0.36
U_{21}	100	7	24 / 17	0.24 / 0.17
U_{22}	100	5	22 / 23	0.22 / 0.23

Table 1: Description of universities, their intrinsic features and maximum utilization in upload / download direction

networks (Netflow data), generically labeled as U_j for privacy reasons.

This data set was completed with the so-called network intrinsic features for the university networks (U_j), that is, the values of their population size and access link capacity. There exist well-documented central repositories which describe the university networks' user population, Internet access capacity and organization [25]. This information has allowed us to select a representative set of universities regarding such intrinsic features (see Table 1), and rank them by means of both population size and access link capacity.

Finally, the *capping effect*, as introduced in [26], states that the traffic demands may be affected (bounded) by a limiting bandwidth capacity. In

this light, Table 1 details the access link capacity C_{U_j} for each university network U_j (third column), along with the maximum average busy-hour traffic over the N days of measurement (last column). The reader should note that all links show low-levels of utilization even at highly-loaded days, typically below 40%. Indeed, such over-provisioning of access links is a common practice by network operators, as noted in [27, 28]. Actually, the average utilization during the busy hour in our set of measurements was typically under 25%, and the most loaded network, U_{14} showed a maximum utilization lower than 75%. Such low levels of utilization make the capping effect negligible in this data, since the maximum busy hour traffic volumes found in the measurements are far from reaching the maximum capacity of access links [29, Chapter 4]. Obviously, in other under-provisioned scenarios with higher levels of utilization, the capping effect cannot be ignored.

3. Characterization and dynamics of the busy-hour traffic process

The following experiments firstly study the marginal distribution of the busy-hour traffic volume or throughput, and then focus on its correlation structure. The results obtained for all IXPs, RNs and Us are summarized in Table 2.

3.1. Gaussian marginal distribution

The first two columns of Table 2 show the estimated mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of the busy-hour throughput distribution over time, measured at each monitored point in both upload and download directions. Essentially, for a given university U_j, whose busy-hour traffic over N days is defined in the set $\{X_1^{U_j}, \ldots, X_N^{U_j}\}$, such mean and standard deviation values are computed as:

$$\hat{\mu}_{U_j} = \frac{1}{N} \sum_{i=1}^{N} X_i^{U_j} \tag{6}$$

$$\hat{\sigma}_{U_j} = \frac{1}{N-1} \sqrt{\sum_{i=1}^{N} (X_i^{U_j} - \hat{\mu}_{U_j})^2} \tag{7}$$

The fourth column in the table shows the maximum coefficient of variation during each busy hour, calculated following Eq. (8). Essentially, for each university U_j, we compute its Coefficient of Variation for each day i and take its maximum value over all N days:

Networks (upload / download)	$\hat{\mu}_{U_j}$(Mb/s)	$\hat{\sigma}_{U_j}$(Mb/s)	$CV_{U_j}^{max}$	Lilliefors ($\alpha = 0.05$)	Shapiro-Wilk ($\alpha = 0.05$)	Anderson-Darling ($\alpha = 0.05$)	Correlation Test ($R > 0.9$)
IXP_1	1243 / 599	78 / 33	0.14 / 0.12	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
IXP_2	1841 / 1021	93 / 55	0.23 / 0.18	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
IXP_3	2335 / 3763	71 / 102	0.07 / 0.04	✓ / ✓	✓ /✓	✓ / ✓	✓ / ✓
IXP_4	32 / 30	2.7 / 2.9	0.18 / 0.21	✓ / ✓	× / ✓	× / ✓	✓ / ✓
IXP_5	1252 / 1166	99 / 86	0.24 / 0.33	✓ / ✓	✓ / ✓	× / ✓	✓ / ✓
RN_1	545 / 1107	32 / 65	0.13 / 0.16	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
RN_2	1365 / 1287	42 / 33	0.08 / 0.06	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
RN_3	101 / 355	10 / 19	0.23 / 0.18	✓ / ✓	✓ / ✓	× / ✓	✓ / ✓
RN_4	80 / 241	5.1 / 14	0.22 / 0.23	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_1	209 / 262	9.2 / 11	0.13 / 0.09	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_2	140 / 151	20 / 15	0.25 / 0.13	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_3	100 / 107	6.2 / 21	0.16 / 0.29	✓ / ✓	✓ / ✓	✓ / ×	✓ / ✓
U_4	65 / 88	14 / 10	0.26 / 0.15	× / ✓	× / ✓	× / ✓	✓ / ✓
U_5	79 / 112	12 / 6.4	0.13 / 0.13	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_6	64 / 107	14 / 9.7	0.20 / 0.09	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_7	85 / 83	8.6 / 9.5	0.19 / 0.11	× / ✓	× / ✓	✓ / ×	✓ / ✓
U_8	19 / 42	3.6 / 4.1	0.28 / 0.09	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_9	33 / 50	2.9 / 6.5	0.46 / 0.28	× / ✓	× / ✓	✓ / ×	✓ / ✓
U_{10}	13 / 26	3.7 / 5.9	0.34 / 0.21	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_{11}	5.0 / 9.1	1.7 / 1.6	0.31 / 0.26	✓ / ×	✓ / ×	✓ / ×	✓ / ✓
U_{12}	109 / 106	14 / 10	0.17 / 0.11	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_{13}	47 / 96	11 / 9.0	0.12 / 0.14	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_{14}	100 / 105	14 / 14	0.19 / 0.21	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_{15}	31 / 52	7.9 / 5.1	0.25 / 0.15	× / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_{16}	86 / 51	7.0 / 8.0	0.20 / 0.16	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_{17}	23 / 49	4.5 / 4.4	0.18 / 0.17	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
U_{18}	31 / 60	3.9 / 3.7	0.26 / 0.23	✓ / ✓	✓ / ✓	× / ✓	✓ / ✓
U_{19}	10 / 28	1.3 / 5.4	0.27 / 0.30	✓ / ×	✓ / ×	✓ / ✓	✓ / ✓
U_{20}	13 / 24	2.9 / 1.1	0.10 / 0.21	✓ / ✓	× / ✓	× / ✓	✓ / ✓
U_{21}	3.5 / 8.2	1.6 / 1.0	0.40 / 0.33	✓ / ✓	× / ✓	× / ✓	✓ / ✓
U_{22}	5.1 / 6.6	2.0 / 2.9	0.46 / 0.25	✓ / ✓	✓ / ✓	× / ×	✓ / ✓

Table 2: Gaussian characterization of busy-hour traffic $N(\hat{\mu}, \hat{\sigma})$ and goodness-of-fit test results in both upload / download directions of traffic

$$CV_{U_j}^{max} = \max_i \sqrt{\frac{V_i}{X_i^2}}, \quad i = 1, \ldots, N \tag{8}$$

This value represents the ratio of the variance V to the mean X, and it is very useful for comparing the degree of variation of the busy hour traffic over different days. This maximum considers the worst possible case: the day which shows highest variability ratio (highest bursty behavior). This result is discussed at the end of this section.

Finally, the last columns in Table 2 provide the results of different Gaussian goodness-of-fit tests applied to the measurements. Essentially, the null hypothesis assumes that the busy hour traffic follows a Gaussian distribution with parameters $\hat{\mu}_{U_j}$ and $\hat{\sigma}_{U_j}$ for university network U_j. The easiest way to visually assess on the validity of the null hypothesis is via the Quantile-Quantile plot [30, Chapter 2], which plots the pairs $x_{(i)}$ versus $Q(i/n))$, whereby $x_{(i)}$ is the order statistics of the empirical sample and $Q(\cdot)$ is the Quantile function (inverse of the cumulative distribution function). If indeed the measured data follows the Gaussian distribution $N(\hat{\mu}_{U_j}, \hat{\sigma}_{U_j})$, the points depicted overlap the angle bisector (line $y = x$). This is the case of Fig. 3, where the QQ-plot technique is applied to the busy-hour measurements of IXP_1, RN_1 and U_1, respectively in both upload and download directions. The same experiment has been applied to all other measurement sets, showing linear QQ-plots in all cases.

Besides visual matches, it is desirable to assess Gaussianity objectively following the most common goodness-of-fit tests found in the literature, say: Lilliefors [31], Shapiro-Wilk [30, Chapter 9], Anderson-Darling [30, Chapter 9] and correlation-based [19]. Basically, the correlation test consists in checking whether or not the linear correlation coefficient R computed between the pairs $x_{(i)}$ and $Q(i/n))$ in the QQ-Plot gives a relatively high value, say 0.9 [20].

Table 2 gives the results obtained after applying such tests. As shown, all empirical distributions pass the correlation test. Also, we observe that the goodness-of-fit tests support the null hypothesis (Gaussianity) for most of the cases and further show visual Gaussianity too. However, as noted in [20], conventional goodness-of-fit tests are usually excessively demanding with traffic measurements. Note that certain outliers may arise from events such as network misuse, power cuts, temporal malfunctioning, etc. instead of typical network behavior, hence making the tests fail. For this reason, we conclude that the busy-hour traffic measurements for the access links of all universities, regional networks and Internet exchange points can be

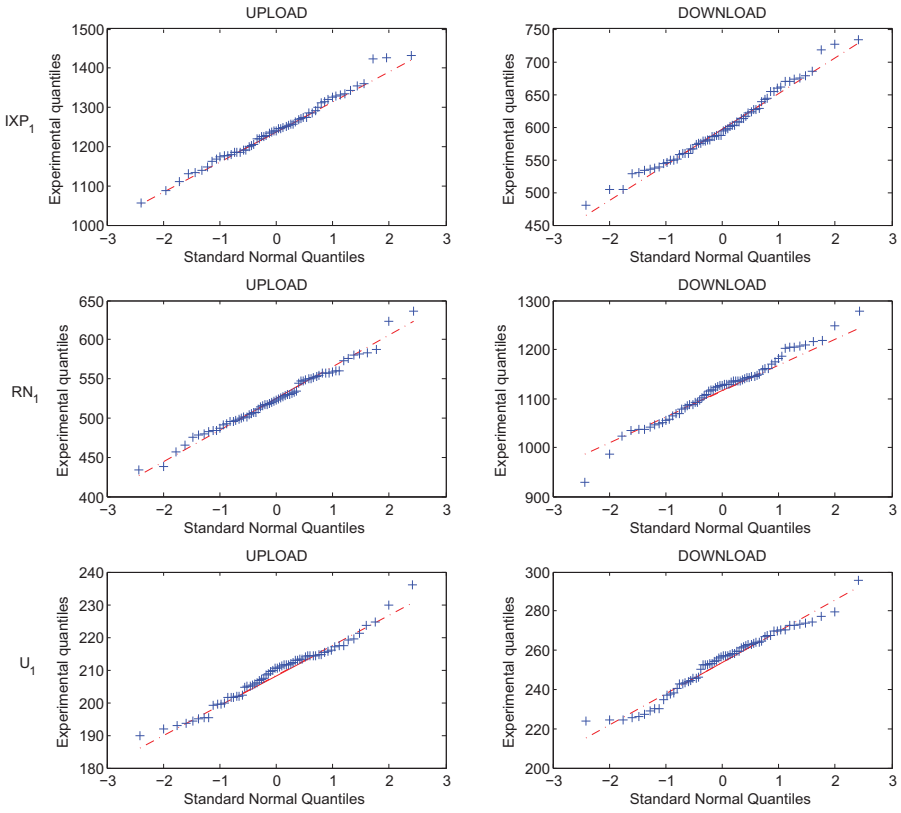

Figure 3: QQ-Plot for IXP_1, RN_1 and U_1

considered "fairly Gaussian", borrowing the term from [20].

3.2. Autocorrelation experiments

This section studies the correlation between consecutive busy-hour traffic measurements, that is, whether or not the busy-hour traffic experienced on one day depends on the values measured the previous days. To this end, the autocorrelation function was calculated for all data items (Us, RNs and IXPs) together with their confidence intervals (with significance level $\alpha = 0.05$) as described by the Bartlett test [32] for the autocorrelation of a pure *white* Gaussian process, i.e., a Dirac delta at the origin of the autocorrelation function. Fig. 4 shows the autocorrelation (solid line) and the confidence intervals (dashed lines) applied again to IXP_1, RN_1 and U_1, respectively. Interestingly, all networks pass this test, which proves the independence of the busy-hour traffic values from one day to another after removing both weekends and holidays.

3.3. Distribution of the busy-hour times

Fig. 5 shows the cumulative distribution function (CDF) of the time instants when the daily busy hour occurs, that is, the value of t in Eq. (3). For the sake of clarity, only the results for six universities are shown, although similar behaviors were observed for the rest of the networks under study. As shown, the CDFs for all six universities behave in a similar manner in each direction of traffic. In the download direction of traffic, the busiest hour typically occurs in the range from 10:00 a.m. to 2:00 p.m. However, the upload direction shows a bimodal behavior with its busiest hour typically found either around 12:00 p.m. or around 5:00 p.m.

These results are consistent with the "Daily traffic pattern" invariant defined in [15]. Essentially, the authors in [15] expose that some traffic patterns follow strictly the human behavior which, in the case of an academic network, this seems to show two peaks of traffic: one in the morning and another one in the afternoon.

As shown, the busiest hour t never occurs at night, which gives at least 12 hours between any two consecutive busy-hour traffic measurements X. Intuitively, this can be the reason that explains why the correlation structure in the busy-hour traffic time-series $\{X_i, i = 1, \ldots, N\}$ vanishes since there is a gap of at least 12 hours between any two consecutive busy-hour traffic measurements X (see Fig. 5).

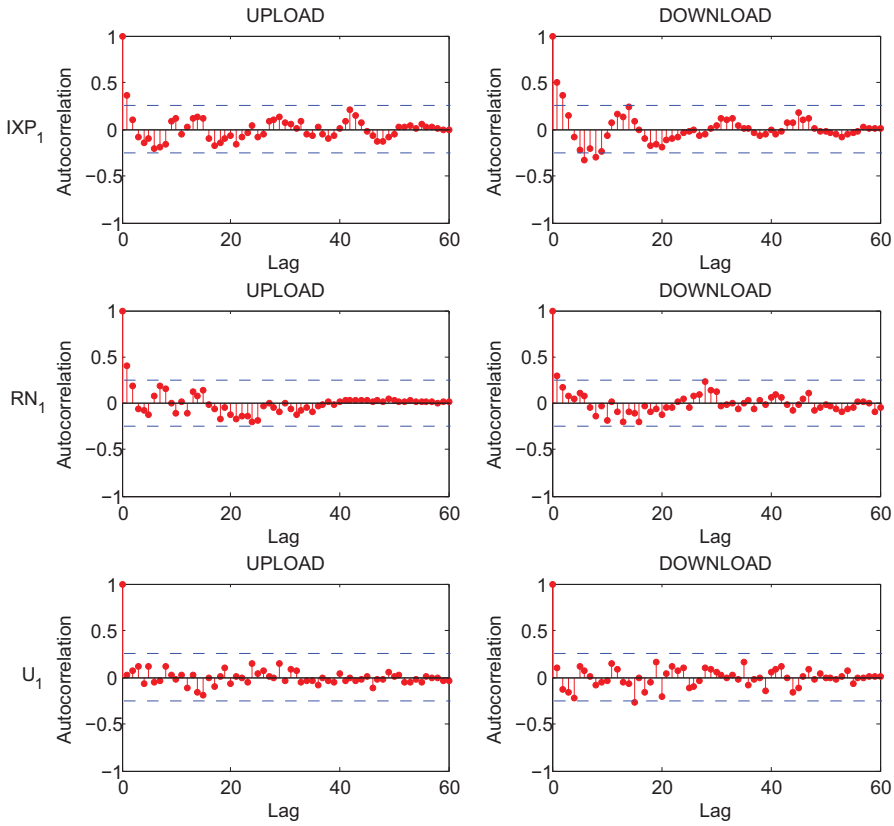

Figure 4: Autocorrelation function and Bartlett Test (dashed lines) for IXP_1, RN_1 and U_1

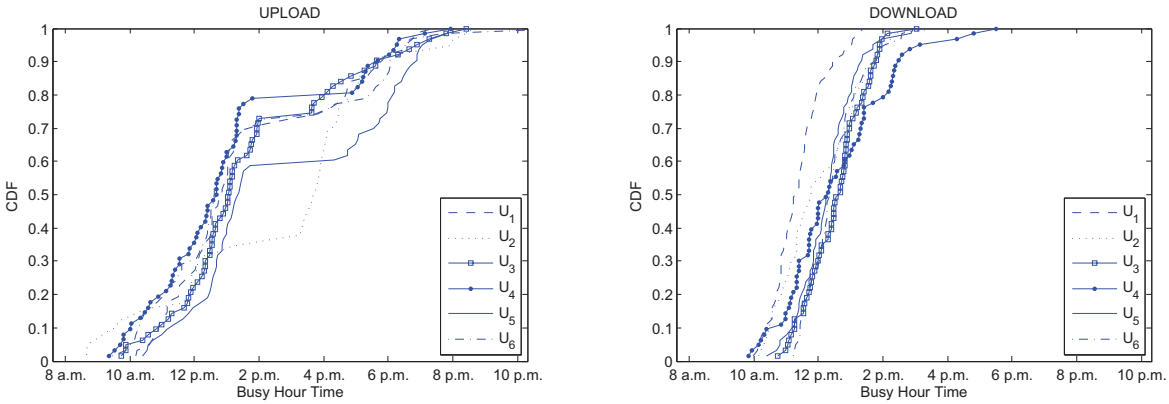

Figure 5: Busy-hour time CDF in both upload and download directions of traffic

3.4. Discussion

On the one hand, the above results show that the busy-hour traffic samples $\{X_1, \ldots, X_N\}$ are both uncorrelated and Gaussian distributed. Hence, the busy-hour traffic process can be modeled by a white Gaussian process.

Additionally, the maximum coefficient of variation, which gives the maximum ratio of the variance V to the mean X over different days i is always smaller than one, hence showing sub-exponential behavior in all cases. In other words, the traffic during the busy hour is close to the average value (small variation with respect to the mean), which is of clear importance for capacity planning purposes.

Having found that the process is white and Gaussian, network operators can apply the conventional sample mean and variance estimator to a measurement set. With such parameters at hand, operators can use the following formula to derive the access link capacity C required such that the busy-hour traffic volume is met with probability $1 - \varepsilon$ (typically $\varepsilon \leq 0.1$):

$$\begin{aligned} C_{U_j} \text{ such that} \quad & \text{Prob}(d \cdot X^{U_j} < C_{U_j}) \geq 1 - \varepsilon, \\ \text{with} \quad & X^{U_j} \sim N(\hat{\mu}_{U_j}, \hat{\sigma}_{U_j}) \end{aligned} \tag{9}$$

This constitutes a first refinement of Eq. (1).

Finally, it is worth noticing that this section's conclusions are derived based on measured busy-hour traffic volumes only. This capacity planning application is not valid for designing new networks over which no measurements have already been taken. For this reason, the next section is

devoted to extracting how much information of the busy-hour traffic is directly related to the number of users (population size) for a given network, on attempts to refine the dimensioning rule above Eq. (9).

4. Factor analysis of access link capacity and population size

The previous tests have shown that the busy-hour traffic distribution of university networks can be accurately characterized by a Gaussian distribution $N(\mu, \sigma)$, whereby its characteristic parameters μ and σ can be estimated from measurements. The next set of experiments aim to study whether or not the intrinsic features of the networks (population size and access link capacity), which are denoted as explanatory variables in what follows, have any influence on such parameters μ and σ, which are denoted as response variables. Note that we are able to compare the traffic of an extensive set of network by means of only two parameters.

To do so, the Analysis of Variance (ANOVA) and Covariance (ANCOVA) methodologies are first reviewed, and then applied to the measurement set. Before that, we remark that this section only takes into account the measurements collected at university access links U_j, with $j = 1, \ldots, 22$. The IXPs and RNs measurements are not considered in these experiments since their population size is unknown.

Both the ANOVA and ANCOVA methodologies require the data to meet a few requirements: First, the samples must be independent and Gaussian distributed; and second, they must share the same intra-group variance (exhibit *homoscedasticity*). However, the results of ANOVA and ANCOVA are generally accepted provided that the number of elements in each group are similar and there is a non-excessive deviation from the homoscedasticity assumption [33, 34]. This is the case for our measurements. Additionally, the ANCOVA model assumes a linear relationship between the response and the explanatory variables. For further details see, for instance, [35, 36].

Table 3 summarizes the factors (access link capacity C_{U_j} and population size P_{U_j}) for each university under study. As noted from the table, the universities have been split into two groups depending on the capacity of their access links: The universities with 1 Gb/s of capacity belong to group G_{high} (which stands for high-speed access link), thus leaving G_{low} to the universities with low access capacity (lower than 1 Gb/s). This classification is important to apply ANOVA, as shown in the following.

Networks	Capacity access		Population size (users)	
	Group	(Mb/s)	Group	(Thousands)
U_1		1000		60
U_2		1000	Large	57
U_3		1000		48
U_4		1000		30
U_5		1000	Medium	28
U_6	$Ghigh$	1000		28
U_7		1000		25
U_8		1000		14
U_9		1000	Small	11
U_{10}		1000		8
U_{11}		1000		6
U_{12}		300		58
U_{13}		200	Large	37
U_{14}		200		35
U_{15}		200		20
U_{16}		200	Medium	19
U_{17}	G_{low}	200		15
U_{18}		100		14
U_{19}		100		13
U_{20}		100	Small	9
U_{21}		100		7
U_{22}		100		5

Table 3: Set of universities grouped by access link capacity and population size

4.1. Effect of access link capacity: ANOVA

This section studies the effect of the access link capacity only on the busy-hour traffic volumes for each university characterized by $N(\mu_{U_j}, \sigma_{U_j})$. Remark that, for each university U_j, its access link capacity C_{U_j} and population size P_{U_j} are known but, for this former experiment, P_{U_j} is ignored.

ANOVA is a widely used statistical methodology whereby the observed variance of a given response variable is split into explanatory factors or categories and provides a means to determine if the factors have any importance in explaining such a response variable, and how much this accounts for.

In our example, ANOVA proceeds as follows: it first splits the response variable μ_{U_j} into two categories: G_{high} and G_{low}. Then, it computes the adjusted mean squares for each category and for the total. The difference between both is due to the experimental error.

Finally, ANOVA performs a contrast test using the ratio between the adjusted mean square of each factor and the total, which follows a Snedecor-F distribution under the null hypothesis; which considers that the total adjusted mean square is due to the experimental error, and not to differences in the population when grouped by categories. However, if the null hypothesis is not accepted, this means that the factor used to build the groups (access link capacity) is statistically significant according to the F-test. Moreover, the ANOVA test provides a p-value which determines if the null hypothesis should be accepted or not, according to a given pre-defined significance level α (typically $\alpha = 0.05$). Basically, if $p > \alpha$, then the null hypothesis is accepted (non-significant factor), and rejected otherwise. Furthermore, ANOVA also quantifies the amount of variance explained by the factors (explained variance) and the amount of variance that remains unexplained (non-explained or residual variance).

It is worth noticing that this test will be applied to both μ and σ in both upload and download directions of traffic. For now, let us refer to them as a generic response variable y. The ANOVA model for response variable y with the access link capacity as its only factor is given by:

$$y_{U_j}^{group} = k_y + \alpha^{group} + \epsilon_{U_j}^{group} \tag{10}$$

Here, k_y is the overall means response for the response variable under study (y), typically named as μ but, in this case, to avoid confusion with the response variable μ_{U_j} we have renamed this term as k. $y_{U_j}^{group}$ refers to the mean or variance (in upload or download direction) value of university U_j which belongs to a given *group* (either G_{high} or G_{low}). The value of α^{group} represents the effect because of a given network U_j belongs to a given

Response variable (direction)		Source of variation	Sum of squares	df	Adj. mean square	F	p-value
Upload	μ	Capacity	5664	1	5664	2.12	0.156
		Error	52042 (90%)	20	2602		
		Total	57706	21			
	σ	Capacity	30.9	1	30.9	1.03	0.323
		Error	604 (95%)	20	30.2		
		Total	634	21			
Download	μ	Capacity	9266	1	9266	2.99	0.099
		Error	62029 (87%)	20	3101		
		Total	71296	21			
	σ	Capacity	64.1	1	64.1	2.62	0.121
		Error	488 (88%)	20	24.439		
		Total	553	21			

Table 4: ANOVA table with access link capacity as factor and μ and σ parameters as response variables (in both directions)

group. Finally, the value of $\epsilon_{U_j}^{group}$ refers to the experimental error introduced above. Clearly, large values of $\epsilon_{U_j}^{group}$ means that the link access capacity factor explains little variance and, perhaps, other factors that explain more variance must be included in the model given by Eq. (10).

Table 4 shows the results after applying the ANOVA test to the busy-hour traffic mean μ and standard deviation σ in both upload and download directions of the university access routers under study. The third column gives the sum of squares for each source of variation: Capacity and Error ϵ. According to the results only the access link capacity factor shows moderate significance (that is, $p \approx \alpha = 0.05$) in the test for the mean in the download direction. On the other hand, the access link capacity factor has no influence for the case of mean and standard deviation in the upload direction of traffic nor for the standard deviation in the download direction.

Finally, the third column also shows the percentage that the error represents of the total variance (inside brackets). It can be noted that the amount of unexplained variance remains high after the test is applied. More specifically, these values are 90%, 95%, 87% and 88%, of the total variance for μ and σ in the upload and download directions, respectively. Indeed, such large values of error reinforce the conclusion that the access capacity barely influences the measurements, namely the measurements are not distorted by

capping effects. Following this, the next section checks whether or not the other intrinsic network parameter, population size, explains more variance than that of the access link capacity.

4.2. Combined effect of the access link capacity and population size: ANCOVA

This section aims to repeat the previous experiment but taking into account both intrinsic network factors: the access link capacity and population size. In this case, the model of Eq. (10) is extended to:

$$y_{U_j}^{group} = k_y + \alpha^{group} + \beta^{group} P_{U_j} + \epsilon_{U_j}^{group} \tag{11}$$

where the term $\beta^{group} P_{U_j}$ has been included with respect to Eq. (10) to deal with the population size factor.

In this case, such factor appears as a quantitative variable rather than a categorical group as it is the case for the access link capacity. When this occurs, it is recommended to use the Analysis of Covariance (ANCOVA) methodology instead of ANOVA. Additionally, ANCOVA is recommended when the two factors are strongly correlated since it helps to separate the amount of variance explained by each factor.

Basically, the Analysis of Covariance is the result of removing the variance for which some covariates or quantitative variables (in this case, the population size) account by means of a linear regression and, after this, applying a regular ANOVA with the access link capacity as unique factor. Note that such a linear regression does not assume that the value of the slopes β^{group} in Eq. (11) for groups G_{high} and G_{low} are equal.

Following this, Table 5 shows the results obtained after applying ANCOVA to the whole set of universities. The table shows a new row that quantifies the adjusted sum of squares of the explained variance by the population size as covariate. As shown, including the population size in the analysis brings two important conclusions: (i) the amount of total unexplained variance reduces very significantly with respect to the previous experiment; and (ii) the amount of variance explained by the access link capacity factor becomes clearly negligible. Concerning the former conclusion, note that the amount of variance that remains unexplained for the mean busy-hour traffic μ has been reduced to 17% and 19% for the upload and download directions, respectively. Remark that, in the previous model which only took into account the access link capacity, the unexplained variance was much higher, up to 90% and 87% (see Table 4) for the upload and download directions, respectively.

Response variable (direction)		Source of variation	Sum of squares	df	Adj. mean square	F	p-value
Upload	μ	Popula.	42015	1	42015	79.6	0.000
		Capacity	814	1	814	1.54	0.229
		Error	10027 (17%)	19	528		
		Total	57706	21			
	σ	Popula.	360	1	360	28.1	0.000
		Capacity	1.62	1	1.62	0.126	0.727
		Error	243 (38%)	19	12.833		
		Total	635	21			
Download	μ	Popula.	48407	1	48407	67.5	0.000
		Capacity	2091	1	2091	2.92	0.104
		Error	13623 (19%)	19	717		
		Total	71296	21			
	σ	Popula.	314	1	314	34.1	0.000
		Capacity	15.5	1	15.5	1.68	0.210
		Error	175 (32%)	19	9.21		
		Total	553	21			

Table 5: ANCOVA table with access link capacity as factor, population size as covariate and μ and σ parameters as response variables (in both directions)

Indeed, the amount of variance explained by the access link capacity was due to the correlation between the population size and the access link capacity of universities, rather than on the latter factor only. This null effect of access link capacity is consistent with the premise of negligible capping effect explained in Section 2.3.

Given that the access link capacity is not significant, the following section is focused on a simplified model that only takes into account the population size via linear regression.

4.3. Focusing on the population size: Linear Regression

As stated before, ANCOVA performs a linear regression in order to remove the variance explained by the covariates. Next, we assess whether such a linear regression can be useful to estimate the busy hour traffic distribution $N(\mu, \sigma)$ based on the university's population size only.

Note that the previous model (Eq. (11)) assumes a different β^{group} for each group of universities G_{high} and G_{low}. The next model simplifies this

issue assuming a common slope β for all universities. Such assumption is known as the homogeneity of regression coefficients [37, chapter 31]. We did not find any evidence that such assumption is violated, consequently we can use the same β parameter for all groups (G_{high} and G_{low}), given by the following simplified model:

$$y_{U_j} = k_y + \beta P_{U_j} + \epsilon_{U_j} \tag{12}$$

which only takes into account the university's population size P_{U_j} as the only source of influence in the busy-hour traffic distribution $N(\mu, \sigma)$. We remark that the value of β represents the slope in the linear regression model, and can be viewed as the amount of traffic that each network user contributes to the average busy-hour traffic value μ_{U_j}. This is a parameter of key importance in the capacity planning of university access links based on their population size.

After applying the linear regression, Table 6 shows the regression coefficients for each response variable, together with their 95% confidence intervals. The fourth column in the table provides the coefficient of determination (R^2) which gives the amount of variance explained by the linear regression model. The results show that the population size explains 81% and 78% of the variance of the busy-hour process mean μ in the upload and download directions of traffic, respectively. For σ, the experiment results give 61% and 66% of explained variance, again in the upload and download directions, respectively.

Response variable (direction)		Coefficients (Mb/s)	95% confidence intervals	R^2
Upload	μ	k = -8.448 β = 0.0027	-26.632 / 9.735 0.0021 / 0.0033	81.21%
	σ	k = 1.707 β = 0.00024	-1.028 / 4.442 0.00015 / 0.00033	61.34%
Download	μ	k = 1.695 β = 0.0029	-20.197 / 23.586 0.0022 /0.0036	77.96%
	σ	k = 1.863 β = 0.00024	-0.547 / 4.273 0.00016 / 0.00032	65.55%

Table 6: Regression coefficients for μ and σ in both directions

Furthermore, Fig. 6 shows the regression lines estimated by ANCOVA for each parameter along with the data, on attempts to provide a visual

contrast of the results.

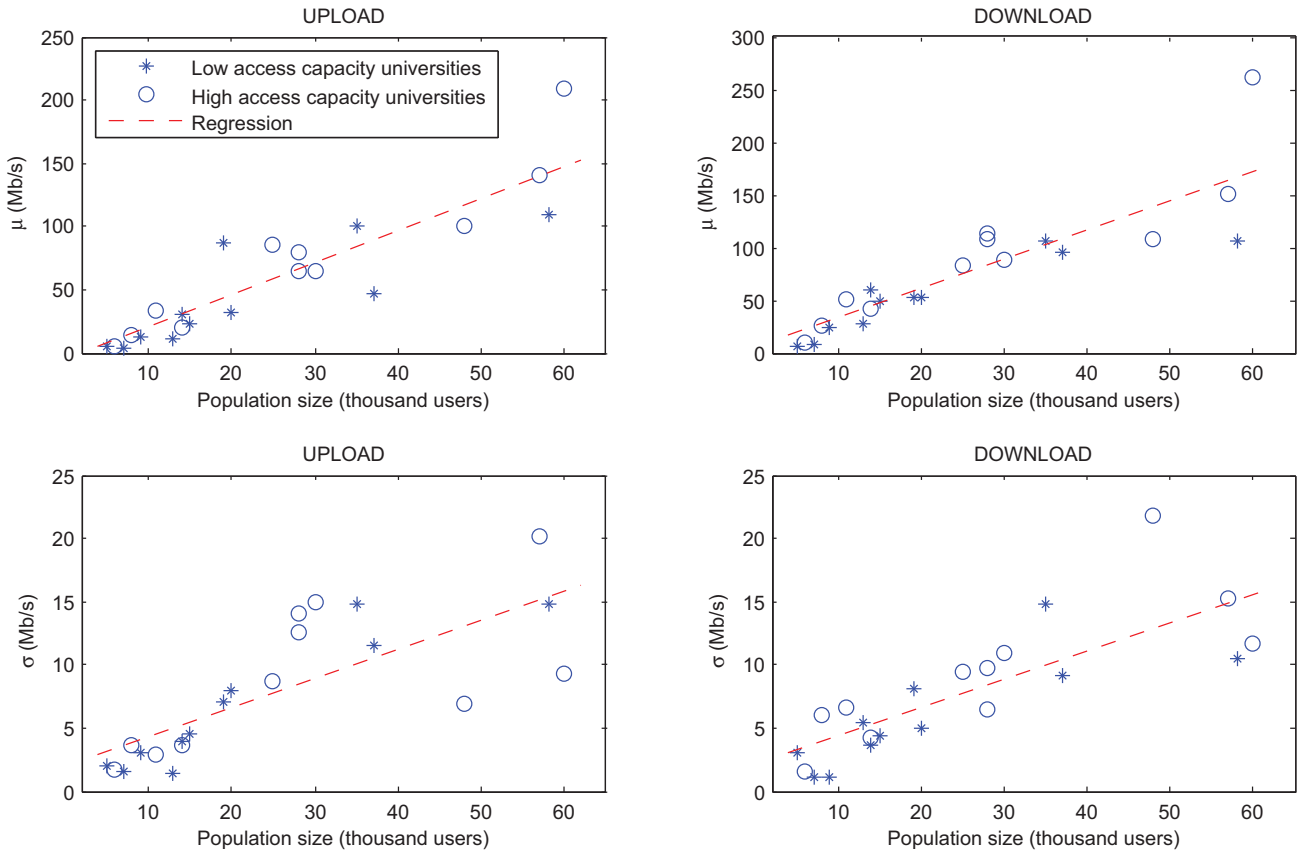

Figure 6: ANCOVA linear regression for the μ and σ parameters in the upload (left) and download (right) directions of traffic

Concerning capacity planning, a given university U_j with population size P_{U_j} requires a capacity in the download direction of $k = 1.69$ Mb/s constant plus 2.9 Kb/s (and depending on the selected significance level an extra addend that takes into account the deviation) per user, as given by Table 6, since the error ϵ is zero-mean Gaussian distributed (modeled with $N(0, \sigma_\epsilon)$). This provides a simple rule for dimensioning the access link capacity of a new university based on its expective population (number of users), as:

$$
\begin{array}{lll}
C_{U_j} & \text{such that} & \text{Prob}\,(d \cdot X^{U_j} > C_{U_j}) \leq \varepsilon, \\
 & \text{with} & X^{U_j} \sim N(\mu_{U_j}, \sigma_{U_j}) \\
 & \text{where} & \mu_{U_j} = k_\mu + \beta_\mu P_{U_j}, \\
 & \text{and} & \sigma_{U_j} = k_\sigma + \beta_\sigma P_{U_j} \\
 & & \text{for each of the directions}
\end{array} \tag{13}
$$

This constitutes a further refinement of Eqs. (1) and (9).

This methodology makes it possible to estimate the demand for bandwidth for new universities, over which no previous measurement experiments have been carried out, in contrast with Eq. (9) which requires a set of busy-hour daily traffic measurements. Furthermore, Eq. (13) can be used to estimate the bandwidth demands for a university network whose population changes with time, that is, whose student body either increases or decreases every academic year.

4.4. Validation model

This section checks whether or not the bandwidth dimensioning model of Eq. 13 is valid for a set of eight new universities, which were not included in the model characterization. This set of new universities were not originally considered in the ANCOVA analysis because of the following reasons: (i) The access link capacity changed during the measurement period (January, 2009 to April, 2009), or (ii) the measurement process failed during several days of the measurement period for those universities. As the previous section showed the access capacity is not a significant factor for the RedIRIS' users demand, and all the universities have at least 2 months worth of data, we regain such discarded set of universities and use them as validation set. The features of this new set are summarized in Table 7. In this light, Fig. 7 shows the measured values of μ and σ in upload and download directions for the eight new universities. Additionally, the model data and regression for μ and σ in both directions, together with the model 95% prediction intervals are depicted. The results show that, in most cases, such values fall within the model's 95% prediction intervals, which supports the applicability of the model. There are two universities (V_2 and V_3) whose measured values fall outside the prediction intervals. These two universities are technical universities which are exclusively devoted to engineering degrees. This may be the reason why they generate much more traffic than other less-technical universities. In the original set of 22 universities used

Networks	Population size (thousand)	Capacity (Mb/s)	$\hat{\mu}_{U_j}$ (Mb/s)	$\hat{\sigma}_{U_j}$ (Mb/s)
V_1	50	2000	148 / 165	13 / 16
V_2	37	1000	163 / 162	14 / 16
V_3	33	2000	140 / 190	19 / 19
V_4	29	1000	65 / 69	8.5 / 7.1
V_5	28	155	25 / 65	9.2 / 9.7
V_6	22	1000	82 / 101	13 / 12
V_7	12	155	11 / 35	3.7 / 4.0
V_8	9	100	4.1 / 10	1.1 / 2.0

Table 7: Description of the validation set of universities and measured values for μ and σ parameters in upload / download direction

to build the model, only one was a technical university, which, in turn, also showed the same behavior (higher traffic demands per user than expected). This result suggests that the model may underestimate traffic demands per user for technical universities, which calls for a future model refinement that accounts for such phenomenon (i.e, including a factor, type of university).

4.5. On the relationship between heavy-hitters and population size

Previous studies have pointed out that most of the Internet traffic is generated by a small fraction of network users [23, 38, 39], often referred to as *heavy-hitters*. As shown in Figs. 8 and 9, there is a clear correlation between the population size of a given university and the number of heavy-hitters observed during its busiest hour. Fig. 8 considers as heavy-hitters all those IP addresses which account for 90% of the total traffic, and Fig. 9 defines heavy hitters as those users who exchange more than 1 Gb (about 100 times the average) of traffic, both measured in the busy hour. The points depicted are computed as the average number of heavy-hitters per day found during the busy hour over the four-month experiment. In the plots, we have removed those university networks in which the use of NAT is a common practice, resulting in a set of eleven networks under study.

These results give support to the idea that heavy-hitters are homogeneously distributed with respect to the population across the universities. In other words, the larger a university is in terms of population size, the more heavy-hitters are expected to be found in its busy-hour traffic measurements. Nevertheless, it is more interesting to define link dimensioning rules based on well-documented intrinsic characteristics such as the population size of universities rather than on measurement-based metrics such as

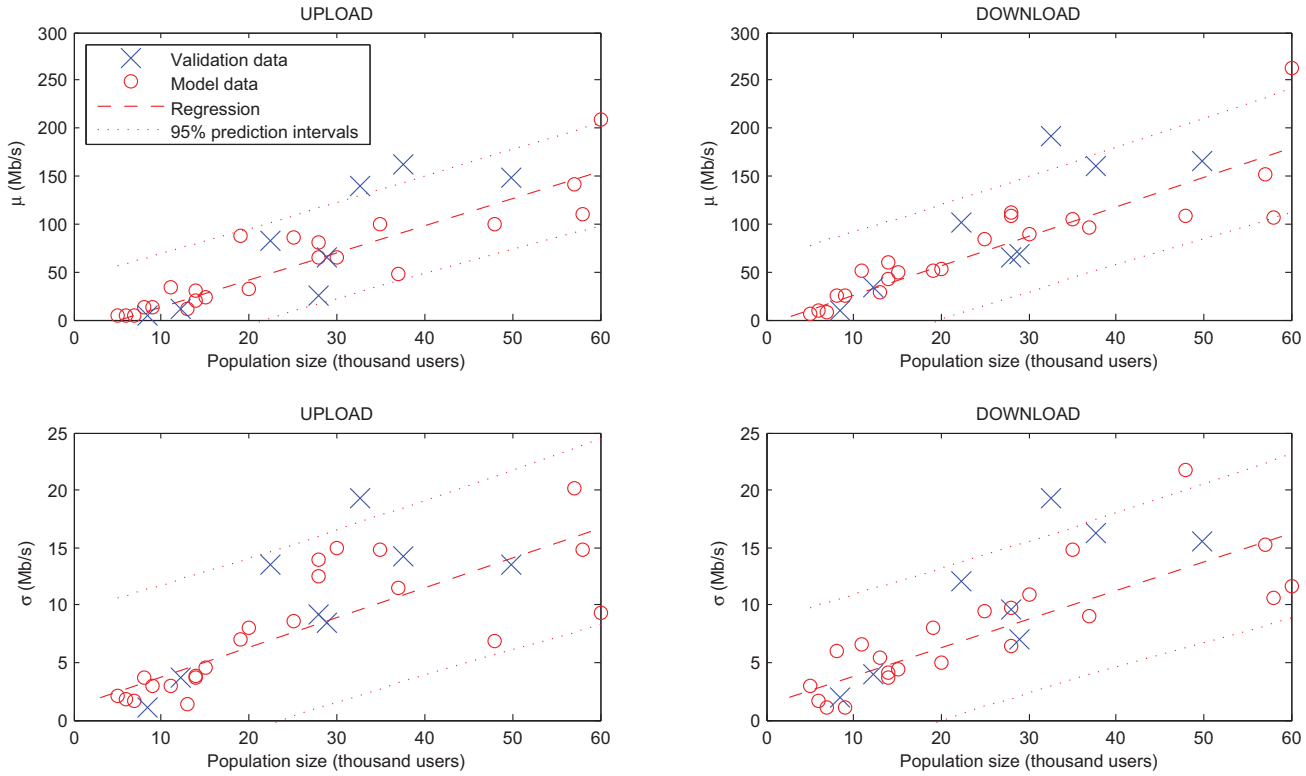

Figure 7: Validation and model set of universities along with model regression and 95%-prediction intervals

the number of heavy-hitters since the latter requires extensive measurement experiments and computational analysis.

5. Summary and conclusions

This work provides an in-depth analysis of the traffic volumes observed during the busy hour of the access links of universities, regional networks, and Internet exchange points in the Spanish Research and Education Network from a long/mid term point of view. First, it is shown that such busy-hour traffic is Gaussian distributed, and shows no correlation between measurements over different days, hence accurately characterized by a white Gaussian process. Therefore, the traffic of a network during several months can be summarized by means of only two parameters (i.e., the mean and

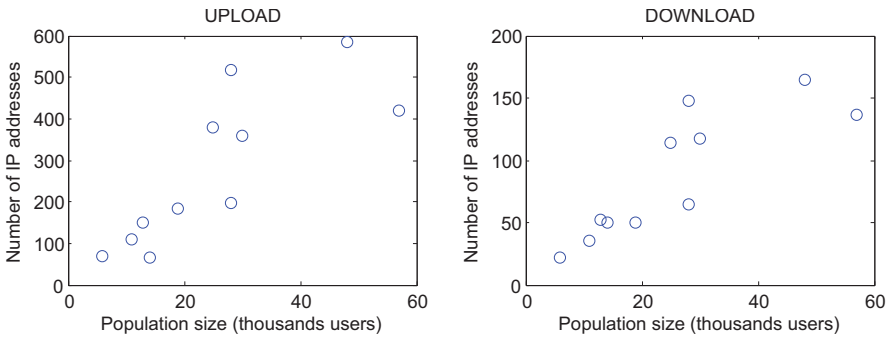

Figure 8: Average number of different IP addresses per day that account for 90% of the total upload/download traffic during the busy hour

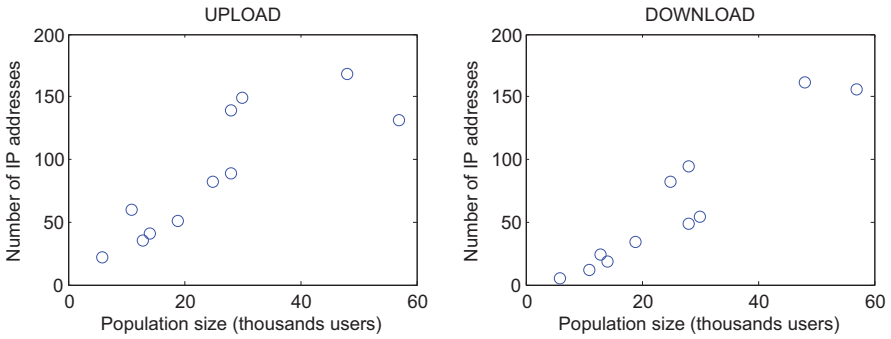

Figure 9: Average number of different IP addresses per day that send/receive more than 1 Gb during the busy hour

standard deviation of such a white Gaussian model) which is useful for capacity planning purposes.

Nevertheless, the network operator must use this methodology only after proving that the stationarity of the measurement set remains, that is, with no infrastructure upgrades, new killer applications appraisal, P2P filtering policy changes, etc. Otherwise, the operator must restart the traffic measurement experiments. Specifically, in this study, we have shown that the stationarity assumption holds in the range of several months for an extensive set of academic networks. As future work, we plan to repeat the proposed methodology with measurements collected on different scenarios, for instance Small-Office/Home-Office (SOHO) networks, and compare the results.

Additionally, this work goes one step further and aims to characterize

the mean and variance of such white Gaussian process based on the population of the network for which the measured access link gives service. The ANOVA and ANCOVA methodology are applied over the Gaussian models that characterize the busy-hour traffic volumes measured for different universities on attempts to check whether or not the universities' intrinsic features (population size and access link capacity) explains part of the busy-hour traffic volume generated. The experiments show that the access link capacity feature show little influence on the busy-hour traffic for networks whose maximum utilization are far from reaching the maximum available capacity. On the other hand, the population size accounts for the majority of explained variance in the ANCOVA test. Furthermore, the test provides a linear regression model and estimates its parameters, making it possible to perform capacity planning for university networks based on their population size. However, after applying ANCOVA the unexplained variance still accounted for some percentage. The use of more features may improve the results that we have shown obtaining more accurate estimations. To this end, we are currently working on finding and analyzing other features such as the type of university under study (technical versus non-technical) as shown in Section 4.4, or the ratio between staff members and students.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Education and Science under project *DIOR* (TEC2006-03246), *ANFORA* (TEC2009-13385), and the F.P.I. Research Fellowship program of Spain. The authors would also like to acknowledge the support of the Spanish National Research and Education Network RedIRIS, which have provided valuable support and experience in the development of this work.

Finally, the authors would like to thank Dr. José Manuel Rojo for the fruitful discussions during the statistical analysis of the measurements.

References

[1] S. Floyd, E. Kohler, Internet research needs better models, ACM SIGCOMM Computer Communication Review 33 (1) (2003) 29–34.

[2] R. Adler, R. Feldman, M. S. Taqqu (Eds.), A practical guide to heavy tails: statistical techniques for analyzing heavy tailed distributions, Birkhauser Verlag, 1998.

[3] J. A. Hernández, I. W. Phillips, J. Aracil, Discrete-time heavy-tailed chains, and their properties in modeling network traffic, ACM Transactions on Modeling and Computer Simulation 17 (4), article 17.

[4] K. Papagiannaki, N. Taft, Z.-L. Zhang, C. Diot, Long-term forecasting of Internet backbone traffic, IEEE Transactions on Neural Networks 16 (5) (2005) 1110–1124.

[5] H. van den Berg, M. Mandjes, R. van de Meent, A. Pras, F. Roijers, P. Venemans, Qos-aware bandwidth provisioning for IP network links, Computer Networks 50 (5) (2006) 631–647.

[6] C. Fraleigh, F. Tobagi, C. Diot, Provisioning IP backbone networks to support latency sensitive traffic, in: Proceedings of IEEE INFOCOM, San Francisco (USA), 2003, pp. 375–385.

[7] R. van de Meent, M. R. H. Mandjes, A. Pras, Smart dimensioning of IP network links, in: Proceedings of IFIP/IEEE International Workshop on Distributed Systems: Operations and Management, San José (USA), 2007, pp. 86–97.

[8] R. McGorman, J. Almhanaa, V. Choulakian, Z. Liua, Empirical bandwidth provisioning models for high speed Internet traffic, in: Proceedings of Annual Communication Networks and Services Research Conference, Moncton (Canada), 2006, pp. 188–195.

[9] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, F. J. Montserrat, E. Robles, T. P. de Miguel, On the duration and spatial characteristics of Internet traffic measurement experiments, IEEE Communications Magazine 46 (11) (2008) 148–155.

[10] 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.

[11] V. Paxson, S. Floyd, Wide area traffic: the failure of Poisson modeling, IEEE/ACM Transactions on Networking 3 (3) (1995) 226–244.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Transactions on Networking 2 (1) (1994) 1–15.

[13] A. Berger, Y. Kogan, Dimensioning bandwidth for elastic traffic in high-speed data networks, IEEE/ACM Transactions on Networking 8 (5) (2000) 643–654.

[14] S. Abeck, A. Farrel (Eds.), Network management: know it all, Morgan Kaufmann, 2009.

[15] S. Floyd, V. Paxson, Difficulties in Simulating the Internet, IEEE/ACM Transaction on Networking 9 (4) (2001) 392–403.

[16] H. T. Marques-Neto, J. M. Almeida, L. C. D. Rocha, W. Meira, P. H. C. Guerra, V. A. F. Almeida, A characterization of broadband user behavior and their e-business activities, SIGMETRICS Performance Evaluation Review 32 (3) (2004) 3–13.

[17] A. W. Moore, K. Papagiannaki, Toward the accurate identification of network applications, in: Proceedings of the Passive and Active Measurement Conference, Boston (USA), 2005, pp. 41–54.

[18] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, C. Diot, Packet-level traffic measurements from the Sprint IP backbone, IEEE Network 17 (6) (2003) 6–16.

[19] J. Kilpi, I. Norros, Testing the gaussian approximation of aggregate traffic, in: Proceedings of ACM SIGCOMM Workshop on Internet Measurement, Marseille (France), 2002, pp. 49–61.

[20] R. van de Meent, M. R. H. Mandjes, A. Pras, Gaussian traffic everywhere?, in: Proceedings of IEEE International Conference on Communications, Vol. 2, Istanbul (Turkey), 2006, pp. 573–578.

[21] T. Oetiker, MRTG - the multi router traffic grapher, in: Proceedings of USENIX Systems Administration Conference, Boston (USA), 1998, pp. 141–148.

[22] S. Leinen, RFC 3955: Evaluation of candidate protocols for IP flow information export (IPFIX), verified on January 2011: `http://www.ietf.org/rfc/rfc3955.txt` (Oct. 2004).

[23] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, F. True, Deriving traffic demands for operational IP networks: methodology and experience, IEEE/ACM Transactions on Networking 9 (3) (2001) 265–279.

[24] R. Sommer, A. Feldmann, Netflow: information loss or win?, in: Proceedings of the ACM SIGCOMM Workshop on Internet measurement, Marseille (France), 2002, pp. 173–174.

[25] Instituto Nacional de Estadística, verified on January 2011: http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft13%2Fp405&file=inebase&L=0.

[26] N. Vicari, S. Köhler, Measuring Internet user traffic behavior dependent on access speed, in: ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management, Monterey (USA), 2000.

[27] Y. d'Halluin, P. Forsyth, K. Vetzal, Managing capacity for telecommunications networks under uncertainty, IEEE/ACM Transaction on Networking 10 (4) (2002) 579 – 587.

[28] A. Odlyzko, Data networks are lightly utilized, and will stay that way, Review of Network Economics 2 (3) (2003) 210–237.

[29] A. Nucci, K. Papagiannaki, Design, measurement and management of large-scale IP networks, Cambridge University Press, 2008.

[30] R. B. D'Agostino, M. A. Stephens (Eds.), Goodness-of-fit techniques, Marcel Dekker, Inc., 1986.

[31] H. W. Lilliefors, On the Kolmogorov-Smirnov test for normality with mean and variance unknown, Journal of the American Statistical Association 62 (318) (1967) 399–402.

[32] D. R. Cox, P. A. W. Lewis, The statistical analysis of series of events, Chapman and Hall, 1966.

[33] S. F. Olejnik, J. Algina, Parametric ANCOVA and the rank transform ANCOVA when the data are conditionally non-normal and heteroscedastic, Journal of Educational Statistics and Behavioral Statistics 9 (2) (1984) 129–149.

[34] G. V. Glass, P. Peckham, J. R. Sanders, Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance, Review of Educational Research 42 (3) (1972) 237–288.

[35] O. J. Dunn, V. A. Clark, Applied Satistics: Analysis of Variance and Regression, John Wiley and Sons, Inc., 1974.

[36] R. Jain, The Art of Computer System Performance Analysis, John Wiley and Sons, Inc., 1991.

[37] M. P. Allen, Understanding Regression Analysis, Springer-Verlag, 2004.

[38] S. Sen, J. Wang, Analyzing peer-to-peer traffic across large networks, in: Proceedings of ACM SIGCOMM Workshop on Internet measurement, Marseille (France), 2002, pp. 137–150.

[39] N. Brownlee, Understanding Internet traffic streams: Dragonflies and tortoises, IEEE Communications Magazine 40 (2002) 110–117.