

Seven Years and One Day: Sketching the Evolution of Internet Traffic

Pierre Borgnat, Guillaume Dewaele, Kensuke Fukuda, Patrice Abry, Kenjiro Cho

▶ **To cite this version:**

Pierre Borgnat, Guillaume Dewaele, Kensuke Fukuda, Patrice Abry, Kenjiro Cho. Seven Years and One Day: Sketching the Evolution of Internet Traffic. Proceedings of the 28th IEEE INFOCOM 2009, Rio, Brazil. pp.711-719. ensl-00290756v3

HAL Id: ensl-00290756

https://ens-lyon.hal.science/ensl-00290756v3

Submitted on 4 Feb 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Seven Years and One Day: Sketching the Evolution of Internet Traffic

Pierre Borgnat*, Guillaume Dewaele*, Kensuke Fukuda†, Patrice Abry* and Kenjiro Cho‡
* CNRS, Univ. Lyon, Lab. Physique ENS Lyon, France. Email: {firstname.lastname}@ens-lyon.fr
† National Institute of Informatics, Tokyo, Japan / PRESTO, JST. Email: kensuke@nii.ac.jp
‡Internet Initiative Japan, Tokyo, Japan. Email: kjc@iijlab.net

***Abstract*—This contribution aims at performing a longitudinal study of the evolution of the traffic collected every day for seven years on a trans-Pacific backbone link (the MAWI dataset). Long term characteristics are investigated both at TCP/IP layers (packet and flow attributes) and application usages. The analysis of this unique dataset provides new insights into changes in traffic statistics, notably on the persistence of Long Range Dependence, induced by the on-going increase in link bandwidth. Traffic in the MAWI dataset is subject to bandwidth changes, to congestions, and to a variety of anomalies. This allows the comparison of their impacts on the traffic statistics but at the same time significantly impairs long term evolution characterizations. To account for this difficulty, we show and explain how and why random projection (sketch) based analysis procedures provide practitioners with an efficient and robust tool to disentangle actual long term evolutions from time localized events such as anomalies and link congestions. Our central results consist in showing a strong and persistent long range dependence controlling jointly byte and packet counts. An additional study of a 24-hour trace complements the long-term results with the analysis of intraday variabilities.**

Keywords: Traffic Analysis; Longitudinal study; Sketch; Robust estimation; LRD

I. Introduction

The Internet is a fast evolving world, or beast, an implicit corrolary often stated being that its robust and sustainable analysis and modeling are impossible and that obtained results may prove to be outdated before being published. This contribution investigates some of the realities beyond this statement and proposes both methodological tools and objective elements of answers to shed light on these issues. We conduct longitudinal analyses of traffic statistics long-term evolutions, for traces collected every day, for 15 minutes, from 2001 to 2008, over trans-Pacific backbone links (the MAWI repository, providing publicly available anonymized traces [1]). Our first major observation is that packet and protocol characteristics remain stable along the years; on the application side, the changes of the applications used on the Internet does not seem to have a major impact on those characteristics. The statistics of aggregated packet or byte count timeseries at the TCP/IP layer are then analyzed, with focus on the evolutions with time of their marginal distributions (MDs) and of long range dependence (LRD). One key difficulty in performing statistical longitudinal analysis is to disentangle smooth long term evolution features from day-to-day fluctuations, as there is no single day without anomalies or specific events. Therefore, our first contribution consists of proposing a robust estimation method based on sketches (random projections) [2], [3], that enables long term analyses without being affected by specific traffic conditions or anomalies. Applied to the 7-year long datasets, this robust estimation procedure brings new insights into the on-going debate related to *bandwidth increase and statistical multiplexing causing the disappearance of long range dependence*. The second contribution lies in finding that, once the impacts of local events such as anomalies and congestions are filtered out, the traffic statistics remain stable along years with persistent LRD and MDs being well modeled with Gamma laws. A concern with this longitudinal study is that data last only 15 min., starting systematically at 2:00 pm. One may question the representativity w.r.t. both the natural intra-day variability and short duration observations. To address this, 24-hour traces were analyzed. We report results for data collected on March 19th 2008, as part of the *Day in the life of the Internet* project [4]. They confirm that the proposed analysis yields robust conclusions, biased neither by the specific schedule nor by short time measurements.

[0]Work supported by Strategic International Cooperative Program between CNRS (France) and JST (Japan). All data used are publicly available at `http://mawi.wide.ad.jp`

II. Related Work

Longitudinal studies: Traffic analyses often consist of snapshot studies of application behaviors, for instance, focused on the impact of the latest killer application, likely to cause changes in traffic statistical characteristics, e.g., web [5], P2P [6], video streaming [7],... There have been fewer studies quantifying the long term evolution of Internet traffic (statistics and applications). One of them is based on NSFNET traces (1988-1993) [8]. At that time, FTP and Mail accounted for about half of the growing traffic volume, until web traffic becomes majority. More recently [9], relations between packet rate, bit rate, and traffic statistics are investigated based on more than 4000 traces collected from 1998 to 2003. In [10], traffic correlation structures before and after the web emergence are compared, showing that web traffic affects at least the finest time scales; however, evolution of longer range correlations (such as LRD) is not reported. A recent study focused on scanning activities on the LBL network for the past 12 years [11].

Long Range Dependence: The discovery of LRD in Internet traffic was epoch-making and raised fundamental issues

[12], [13]. Specifically, a characteristic related to LRD is the high variability of traffic fluctuations, yielding degradations of queueing performance [14]. Difficulties in empirically assessing LRD in real traffic time series have been thoroughly discussed [15]–[17], showing the relevance of a wavelet-based analysis framework [16]. However, stability (or even existence) of LRD traffic is an ongoing debate and claims were made predicting its disappearance on backbone or when loads increase (cf. e.g., [18], [19]). A number of authors discussed the fact that LRD in Internet traffic can be induced by higher-layer protocols [10], [15], as well as related to the heavy tail natures of the distributions of the file size to be transferred [5], [12]. The (heavy)-tail behaviors of IP flow size have been continuously investigated (see e.g., [20]). However, the practical validity of the control of LRD of heavy-tail in actual traffic has only been assessed in recent studies [21] as (heavy)-tail behaviors of IP flow size is an elusive characteristic to estimate. Therefore, we here concentrate only on proposing a robust and reliable method, combining LD and sketches, to evaluate the relevance of LRD.

III. MAWI Dataset

A. Monitoring point

The MAWI traffic repository archives traffic data collected from the WIDE backbone networks. The WIDE network (AS2500) is a Japanese academic network connecting universities and research institutes. The MAWI repository has been providing anonymized packet traces since 1999 (total volume of available data exceeds 1TB as of April 2008, cf. `http://mawi.wide.ad.jp/` and [1]). A specific note here is that the data used here are all publicly available on the website.

Our main datasets are daily packet traces captured at Samplepoint-**B** (hereafter **B**) from 2001/01 to 2006/06, then at Samplepoint-**F** (hereafter **F**) from 2006/10 to 2008/03. These are transit links of the WIDE network, and the link of **B** was replaced in July 2006 by the link **F**. Traces just after the upgrade are missing until 2006/10. At **B**, congestions were frequently observed, the link was a 100Mbps, with 18Mbps Committed Access Rate. The link for **F** is over-provisioned, it started as a full 100Mbps link and upgraded to a 1Gbps link with the capped bandwidth of 150Mbps in June 2007. Daily packet traces are captured from 2:00 pm to 2:15 pm everyday (Japanese Standard Time, UTC+9). The traces, with anonymized IP addresses and without payloads, are made available to the public along with a summary information web page about the traffic. Occasionally, 24-hour or longer traces are made captured. The 24-hour-long traces collected at **F** on 2008/03/19 is used in Sec. IV-E to show the consistency of the obtained results. The WIDE transit link mostly carries trans-Pacific commodity traffic between Japanese research institutions and non-Japanese commercial networks, as WIDE peers with most major domestic ASes at the Internet Exchange Points it operates, and international traffic between academic networks goes through other international research networks. Traffic is also asymmetric as WIDE has other trans-Pacific links, meaning that many flows can be observed in one direction only. This compels us to study traffic separately for each direction, labeled US2Jp, for traffic going to Japan, and Jp2US, for outgoing traffic, as most traffic is between Japan and the USA. The traffic is highly aggregated: A 15-minute-long trace usually contains 300k-500k unique IP addresses, and various kinds of anomalies.

Because the traces are taken on links used in real traffic conditions, the ground truth is not always known about the whys and whens for some specific events. For instance, there is no control and only scarce explanation usually about why the used bandwidth increases or not. One goal of this study is to show that a proper methodology as proposed here gives most of the information about what happened in a trace, both in terms of statistics and flow or packet characteristics.

B. Throughput Evolution

Strong variability: Fig. 1 displays throughput evolutions, and their intraday variabilities (measured as standard deviations (STD) computed around 1s time window averages). A wide range of throughput values and huge intraday variabilities (STD varies by a factor of 10) are observed, together with a global increase of throughput from 100 kbps in 2001 to more than 12 Mbps in 2008. At **B**, the load steadily increases over years up to the link capacity. The upgrade from **B** to **F** induced a significant change in average throughput (currently varying between 5 and 10 Mbps). The datasets enable then the study of the evolution of the traffic over 7 years, under both congested and over-provisioned conditions.

Congestion periods: B experienced several long lasting congestions (shown in Fig. 1): US2Jp, from 2003/04 to 2004/10 and from 2005/09 to 2006/06; Jp2US, from 2005/09 to 2006/06. The byte throughput is close to constant with a low level of fluctuations (80% drops in STD). Although drops in STD are long in duration, they do not allow detection of congestion, as short time fluctuations can locally have amplitudes of same order.

Specific periods: Two periods with unusual traffic behavior (gray-shaded areas in Fig. 1) call for specific comments. From 2003/05 to 2004/03, Jp2US traffic underwent a severe volume decrease (Fig. 1, left). This had likely been caused by a change in the routing policy or by upward link congestions. Interestingly, despite this low volume, the traffic composition and its statistical characterization have not been significantly affected. From 2004/07 to 2005/04, US2Jp (Fig. 1, right), strong fluctuations in packet number are observed (STD being extremely high) due to massive activities of the Sasser worm (see also Sec. III-C).

C. Protocol and Application Breakdown

Methodology: Protocols, applications and anomaly breakdowns (shown in Fig. 2 for both directions) are first obtained from classical procedures, using protocol breakdown then a Port number based identification. Unknown port numbers are found associated to dynamic ports larger than 1024. Those packets were shown to be mostly linked to *P2P hiding*, [6].

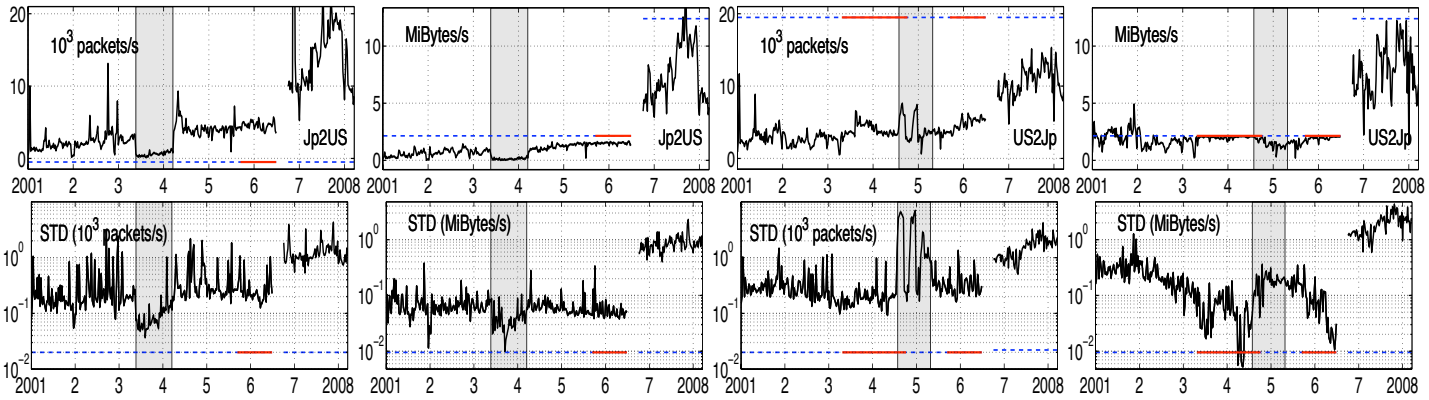


Fig. 1. **Throughput vs. years.** Top: Pkt and Byte throughputs. Bottom: Intraday variability (log of the standard deviation, computed from 1s time windows). Congestions are marked with solid lines (below for Pkt, at CAR bandwidth limitations for byte). Left: Jp2US; Right: US2Jp.

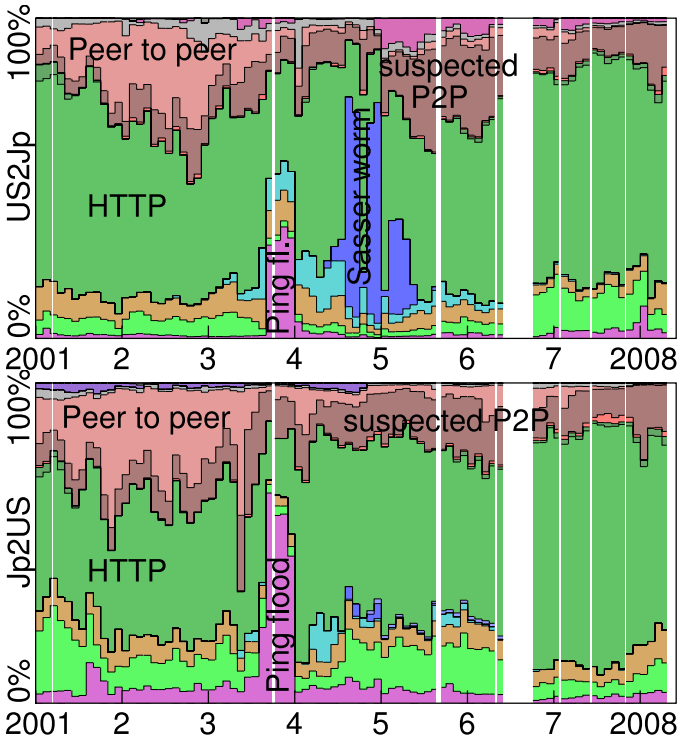


Fig. 2. **Protocols / Applications / Anomalies breakdown vs. years.** Bottom to top : Ping, DNS, common services, MS vulnerabilities, Sasser, HTTP, broadcast, suspected P2P, identified P2P, other TCP/UDP, GRE (US2Jp) or INLSP (Jp2US).

A further refinement of the classification in terms of types of applications is made possible by combining the Port number based identification with heuristic rules (based on proportions of types of packets, on sizes of packets, and so on) in the spirit of [22]. This set of heuristic rules is not detailed here. It was already used in post-processing of the packet level anomaly detection procedure proposed and validated in [23]. Finally, using this specific anomaly detection method [23], it is found that most of the remaining unidentified traffic consists of anomalies. In the situation of this study, note that this methodology is state-of-the-art because there is neither access to the payload of the packets (so deep packet inspection methods are not possible), nor any possibility to use traffic classification method based on the reconstruction of the exact sequence in the flows.

Protocols: Over the 7-year period, TCP and UDP continuously conveyed more than 90% of packets. ICMP (Ping) presents a noteworthy share of the packets, more frequent for Jp2US traffic ($\simeq$5%). Ping floods are quite common and have been regularly and automatically detected along the 7 years. Particularly, a ping flood detected during 2003/08-12 lasts several months, with a very high volume: more than 50% packets for Jp2US, and around 25% for US2Jp, are ICMP packets. A number of anomalies were long-lasting ones or contributed to a significant proportion of traffic (locally more than 80% of the link capacity). Also some unexpected protocols found is GRE (around 5% of traffic, Jp2US) or INSLP a security layer protocol (US2Jp).

TCP/UDP: A large proportion of packets consists of Web traffic: At **B**, 40% of legitimate traffic for Jp2US, and $50-55\%$ for US2Jp. After the link upgrade to **F**, it increases to roughly 60% for both directions. The second largest group is P2P traffic. Common Internet services such as FTP, mail (SMTP, POP, IMAP,...), news protocols,... account together for only about 5%; this remains stable over years. Most remaining traffic is targeting Microsoft services (such as MS RPC, MySQL, file sharing), up to 2% for US2Jp, which are often associated to malicious activities. Streaming protocols (Realserver, Shoutcast) represent 1 or 2%. At **B** Jp2US, DNS traffic is larger ($\simeq$15%) than for US2Jp ($\simeq$5%). For **F**, this is inverted, likely due to the anycast deployment of M-root DNS server operated by WIDE.

Peer to Peer: In 2002, traffic using known P2P ports constitutes around 30% of the packets, mostly Napster, and others such as Gnutella and clones, Kazaa, WinMX, and Emule-Edonkey. This identified P2P traffic tends to disappear over the years to quasi invisible after 2004. Some (but only $\simeq$ 2%) P2P traffic (Bittorent) is still identified. However, P2P traffic decline is only an appearance and actually corresponds to *P2P*

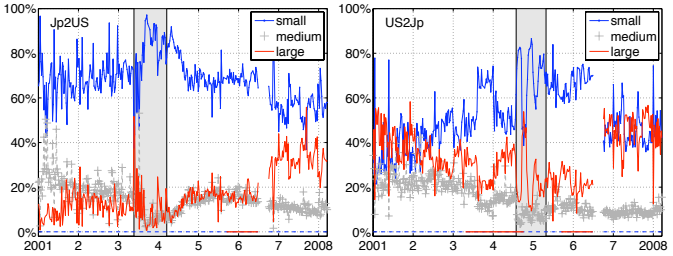


Fig. 3. **Packet sizes vs. years.** Proportion of large ($\geq$ 1400B), small ($\leq$ 144B), or medium (in-between) packets. Left: Jp2US; Right: US2Jp.

hiding [6]: port number based identification methods no longer work, as modern P2P software uses random higher port numbers. And, indeed, a significant increase over recent years of TCP and UDP traffic between high ports (≥ 4000) correlates well with identified P2P decline. Aggregating identified and unidentified P2P traffic yields a significant increase in absolute volume at **F** (compared to **B**), yet a slight decrease of relative share.

Anomalies: Applied to each day of the 7 year long dataset, the detection procedure in [23] showed that usually around half a dozen (often many more) of suspicious anomalous events are identified in 15 min. traces. Some of them last weeks, or months; others are short (from seconds to minutes) – justifying the use of a multi-scale detection method like that of [23]. One large anomaly corresponds to the Sasser worm activity (2004/05 and 2005/05 mainly). Successive outbursts (cf. Fig. 2) are observed (2004/08, 2004/12 and 2005/03): Sasser was on the verge of disappearing twice, yet came back as variants of the worm. Sasser activity accounts for more than 50% of the US2Jp traffic, while is barely noticeable for Jp2US (likely due to a better than average defense vs. worms on academic computers). Besides those major events, many other anomalies exist: recurrent ones are SYN scans and floods towards HTTP. Anomalies targeting any and all protocols and applications are also commonly found, usually in close relation with the protocol popularity itself: specific types of anomalies are indeed found depending on the analyzed period: e.g., NNTP for the earlier days, SSH since 2004, MS security holes from 2003/08.

Packet size distributions: Finally, packet size distributions is reported in Fig. 3. As expected, most packets are either small ($\leq$ 144B signaling packets) or large ($\geq$ 1400B – usually data frames). The only notable evolution is the upgrade from **B** to **F**, when the proportion of large packets increased: due to congestion, there was a latent demand for more data exchange. Yet, other statistics does not change at that time, as seen in the next sections. This is a non-event for traffic analysis. Also, we confirmed clear appearance of some typical intermediate size of packet over 7 years, used by specific applications (e.g., $660B$ used for P2P software). A consequence of the slow (if any) evolution of packet size distribution is the high correlation between $H(B)$ and $H(P)$ that is reported later on in Sec IV-D.

Summary: Over the 7 years, for both **B** and **F**, the content of (non anomalous) traffic does not change significantly. The protocol/application breakdown reported here well matches those provided in [9] for traffic collected in 1998-2003. However, they are in clear contrast with those in the 90s when most of the traffic consisted of FTP and email [8]. A salient result lies in the fact that *normal* traffic is never observed! Numerous and significant anomalies are consistently found each day, ranging from major anomalies consuming more than half of the throughput during several consecutive months to short-lived anomalies existing only one day. These findings underline the need for estimation procedures that untangle anomaly impact from long term evolution when performing longitunal studies of traffic.

IV. Robust statistical characterization

For the statistical characterization of aggregated packet (X_Δ) or byte count time series (X_Δ and W_Δ), our goal is not to propose self-consistent statistical models for Internet traffic, but rather to focus on the long term evolutions of some of its salient features, namely marginal distributions and LRD, i.e., one and two-point statistics (cf. [12], [13], [17]). Hence, analyses are confined here to these properties, overlooking other interesting ones (e.g., short time correlations).

A. Statistical description

Marginal distributions and Gaussianity: The marginal distributions (MD) of X_{Δ_j} and W_{Δ_j} are analyzed via empirical histograms, for $\Delta_j = \Delta_0 2^j$, with $j = 1, \ldots, J$, $\Delta_0 = 1$ms and $J = 10$, that is from 1ms to 1s. Following [17], [23], Gamma laws are used to model the necessarily positive X_Δ and W_Δ. A $\Gamma_{\alpha,\beta}$ distribution is defined as $\Gamma_{\alpha,\beta}(x) = (x/\beta)^{(\alpha-1)} \exp(-x/\beta)/(\beta\Gamma(\alpha))$. While the scale parameter β mostly represents the volume, the shape parameter α is used here as an indicator of closeness to Gaussianity. Indeed, skewness and kurtosis, which are 0 for Gaussian, behave respectively as $2/\sqrt{\alpha}$ and $6/\alpha$, for Γ. Hence, the smooth transition of Γ from exponential to Gaussian is controled by $1/\alpha$. The shape α_j and scale β_j parameters are systematically estimated for X_{Δ_j} and W_{Δ_j}.

Spectrum and LRD: For stationary processes, two-point statistics are analyzed via their spectrum $f_X(\nu)$. LRD is defined as: $f_X(\nu) \sim C|\nu|^{-(2H-1)}$, when $|\nu| \to 0$. H is referred to as the Hurst parameter [12]. It is well-known that LRD is best analyzed in a wavelet framework through the relation: $S_j = (1/n_j)\sum_{k=1}^{n_j} |d_X(j,k)|^2 \sim C2^{j(2H-1)}$, when $2^j \to +\infty$ and where the $d_X(j,k)$ are the (Discrete) Wavelet Coefficients of X_{Δ_0}, at scale $2^j\Delta_0$ and time position $k2^j\Delta_0$. By nature, wavelet coefficients indeed consist of aggregated versions of X at level $2^j\Delta_0$. The plots $\log_2 S_j$ versus $\log_2 2^j = j$ are commonly referred to as logscale diagrams (LD), and serve as the basis for Hurst parameter estimation [16].

B. Impact of the various traffic conditions

Internet traffic is not intrinsically stationary (daily or weekly seasonality, anomalies,...). However, for 15-minute long traces,

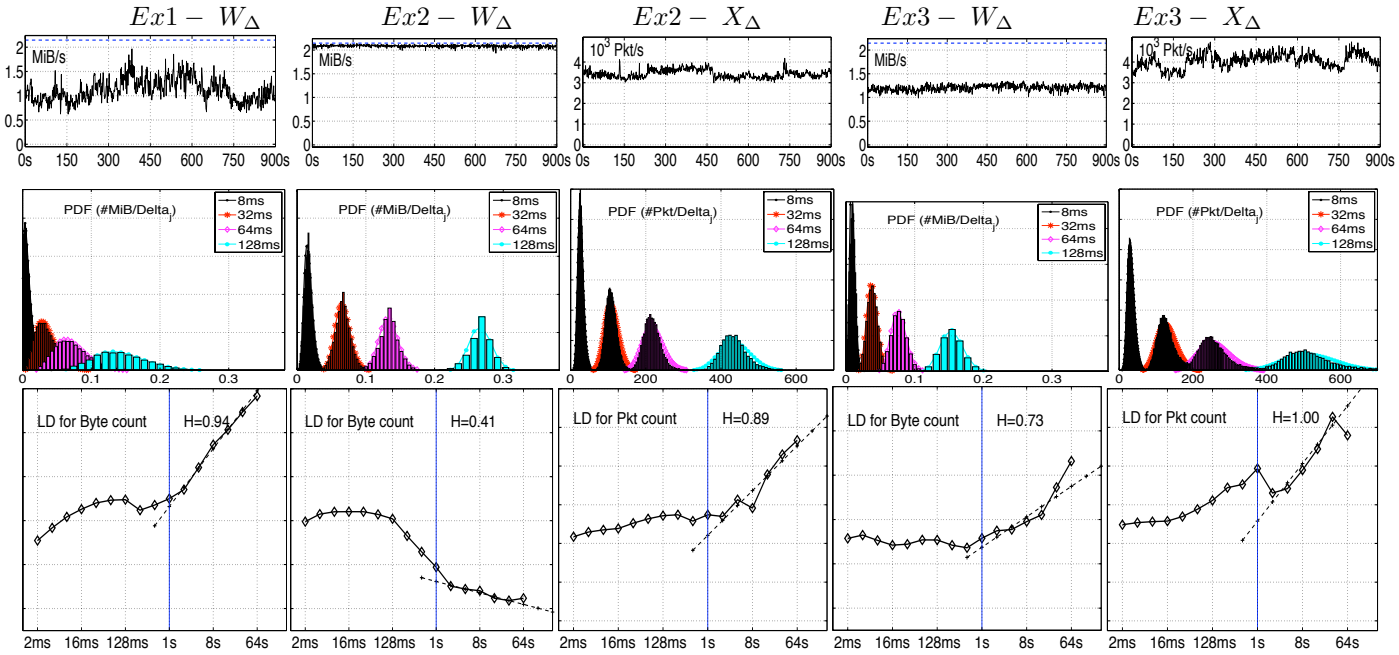


Fig. 4. **Statistics in various traffic conditions.** Aggregated ($\Delta_0 = 1$ms) byte (W_Δ) or packet (X_Δ) count time series (top row); Marginal distributions (MD) for $\Delta_j = 4, 32, 64, 128$ms, both as empirical histograms (bars) and Γ fits (lines) (middle row); Logscale Diagrams (LD) (bottom row). $Ex1$ (col. 1): **B**-US2Jp, 2005/07/11, anomaly-free. $Ex2$ (cols. 2&3): **B**-US2Jp, 2003/06/03, congestion (in term of Byte count (W_Δ), not on X_Δ). $Ex3$ (cols. 4 &5): **B**-Jp2US, 2004/09/21, anomalies (network scan, spoofed flooding, attack on Realserver).

stationarity is fairly well satisfied, as discussed in Sec. IV-E, and in consistence with [16].

Example 1 (Fig. 4, US2Jp, 2005/07/11, $Ex1$) has been chosen because traffic is neither congested, restricted nor with anomalies (assessed by a careful human inspection assisted with the anomaly detection procedure of [23]). MDs, at various Δ_j, are well modeled with Γ laws. The LD exhibits a *knee-shaped* form with both short range dependencies (SRD) at fine scales (from 1ms to less than 1s), and long range dependencies (LRD), at coarse scales (from 1s to 500s, with estimated $H \simeq 0.95$), separated by a typical scale $2^{j_*}\Delta_0 \simeq 1$s. This is consistent with observations (and models) reported in the literature over the years [10], [17], [24], [25]. Common knowledge is that fine scales are related to the packet arrival process while coarse scales are related to flow characteristics, notably the heavy tail packet number distributions. Note that here, only Byte count is displayed but plots shown for X_Δ and W_Δ are comparable in typical situations.

Example 2 (Fig. 4 US2Jp, 2003/06/03, $Ex2$ 2 & 3th col. in Fig. 4) is collected under byte congestion. Clear changes for W_Δ mostly are observed. MD can still be modeled by Γ laws, though the α_j and β_j (not reported here) significantly differ from those of $Ex1$. The LD is strongly altered: LRD no longer exists at coarse scales. This is due to the congestion, inducing that the byte count remains quasi constant, and the absence of variability implies that of LRD. However, there is no reason for a change in heavy-tailness of packet number distributions, hence calling into question this disappearance of LRD. Moreover, X_Δ is still displays LRD.

Example 3 (Jp2US, 2004/09/21, $Ex3$, 4 & 5th col. in Fig. 4) Traffic here contains several attacks (identified using [23] and validated by manual packet inspections): Network port scan with SYN, single source SYN flooding, distributed spoofed flooding, attack against a Realserver through TCP port 554. LRD is not altered by these attacks: The LRD onset remains around $2^{j_*}\Delta_0 \simeq 1$s and the Hurst parameter is not markedly varied. However, anomalies impact the range of fine to intermediate scales of the LD, and therefore the SRD of X_{Δ_0}. Simultaneously, MDs remain well modeled with Γ laws, despite the occurrence of attacks. However, α_j, hence the route toward Gaussianity, is significantly modified when anomalies occur (in consistence with [17], [23]): Changes in α_j, for the range of scales 1ms $\leq 2^j\Delta_0 \leq$ 1s, can only result from a change in the structure of the short time correlation in the data. This is the grounding ingredient of the anomaly detection procedure that was proposed in [23].

Discussion: These examples show that changes in traffic conditions (congestions, anomalies,...) drastically affect the parameters of the statistical modeling. Observations drawn from other days under congestion or with anomalies are consistent with these reports and, as mentioned above, there is almost no normal day (i.e., without numerous low-volume anomalies). This is a severe difficulty in performing longitudinal statistical study of traffic as intended here (especially automatic and unsupervised): The risk being that the study boils down to a long list of specific situations, without any possibility to identify normal or expected behaviours, and hence no global and long term features. To overcome this, a *robust* estimation procedure is now proposed.

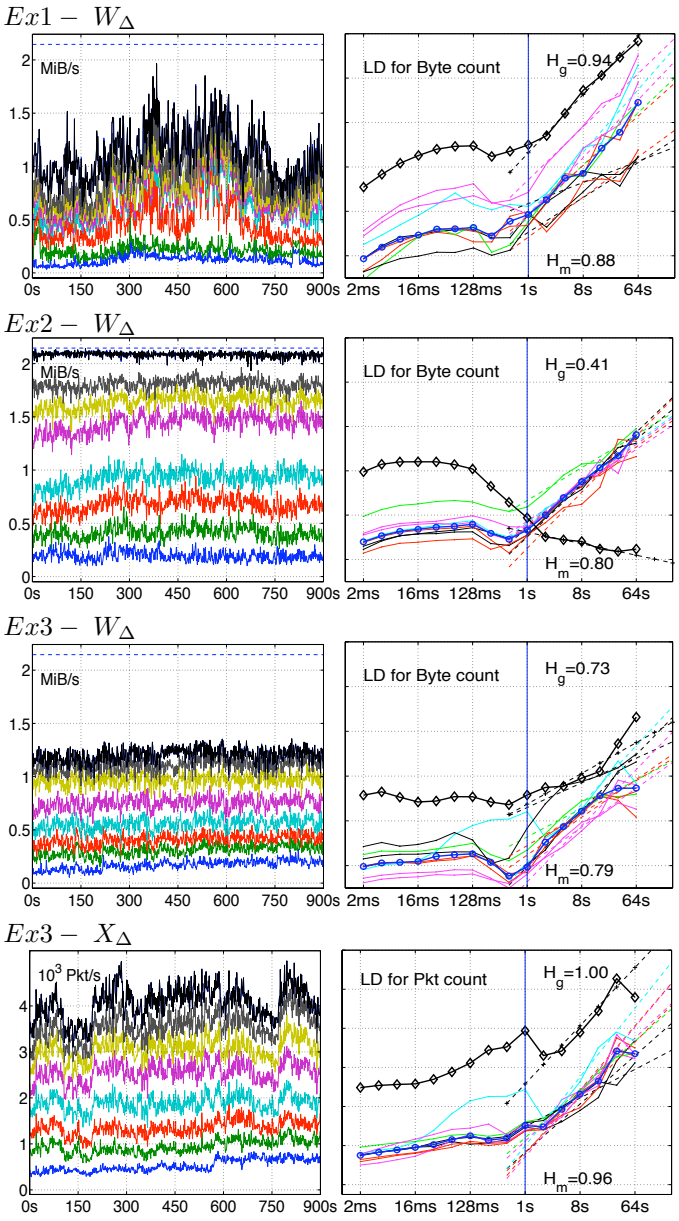


Fig. 5. **Robust estimation.** Left: Aggregated byte time series (1s) displayed cumulatively by sketch sub-trace; Right: LDs for global traffic (thick line, $\diamond$), for sketches (thin lines) and for median-sketch (thick line, o). Top row, **B**-US2Jp, 2005/07/11, no congestion: $Ex1$ (Byt). Second row, **B**-US2Jp, 2003/06/03, with congestion: $Ex2$ (Byt). Bottom rows, **B**-Jp2US, 2004/09/21, with severe anomalies: $Ex3$ in Bytes and Packets.

C. Sketches for robust estimation

In statistical signal processing, robustness in estimation is classically achieved by performing averages over independent copies of equivalent data. Here, this would mean either split data into shorter traces or average equivalent days, but 15min long data are too short for trace spliting and identifying equivalent days is a complex and dubious solution. Instead, following [2], [3], [23], we turn to the use of random projections (usually referred to as sketches).

Sketches: Let h_n denote a $k-$universal hash table of size M (computed using a fast-tabulation [26]). The original collection of packets is split into M sub-traces, each of them consisting of all packets with identical sketch output $m = h_n(A)$, where the hashing key A is chosen as one of the packet attributes (IPdst, IPsrc,...). This amounts to performing random projections, preserving flow structures (packets belonging to a given flow are assigned to the same sub-trace). Each sub-trace is aggregated, $X_{\Delta_0}^{(m)}, m = 1, ..., M$, and analyzed following the procedures used for the original trace. Robust estimation results from averaging, by means of *median*, over the sketch outputs.

Example 1: Fig. 5, top row, shows aggregated sketched ($M = 8$) for W_Δ and their LDs. The M LDs display a weak variability around a well-defined average: All sketches are statistically equivalent. Hence, the median LD matches perfectly (up to a vertical shift, due to the division by M) the LD computed from the entire trace. The Hurst parameter estimated from the median of the estimates over the M sketches, H_m is consistent with H_g estimated from the whole trace. This also shows that flow-sampling is compatible with LD estimates, in a better way than flow-preserving averages.

Example 2: Fig. 5, second row, shows aggregated sketches and their LDs for W_Δ. Each sub-traces has recovered a significant variability, when the original showed almost none. Accordingly, the sub-trace LDs exhibit back the *knee-shape* form with $j_* \simeq 9$ or 10 (0.5s to 1s) and estimated H in the usual range $[0.8, 1]$. This indicates that they are characterized by an unquestionable and significant LRD. Whereas the global analysis of a trace under a congested day leads to the erroneous conclusion that congestion eliminates LRD, a sketch based analysis reveals that the network mechanisms at work to create LRD remain equally and strongly active under congestion. Moreover, a relevant estimation of the LRD parameters can be automated by median over sketches.

Example 3: For $Ex3$, Fig. 5, last two bottom rows in Bytes and Packets, all sub LDs are quasi identical, but two. Inspection confirms that these two LDs correspond to sketch output that convey the anomalies of that day. Computing the LD median results in an analysis of the traffic covariance structure that is not impacted by these significant anomalies. The median LD differs from the one computed from the entire traffic, mostly in the fine scale range (0.1s to 2s), in agreement with previous findings [23]: Low volume anomalies mostly affect short time-scales. The median-sketch based procedure provides a relevant estimate for H even when anomalies are present: Traffic LRD per se is not affected nor varied by low-volume anomalies.

Summary: These case studies show that the proposed median-sketch estimation procedure is statistically consistent and provides robustness against severe traffic condition changes (congestions, restrictions, low-volume anomalies,...). Analyses have been carried over LDs, yet equivalent (not reported here) conclusions are drawn when studying MDs (and the α_j and β_j). These observations justify the crucial choice of the *median*, instead of the simpler *mean*, to average estimates: Median is a non linear procedure providing *robustness* against

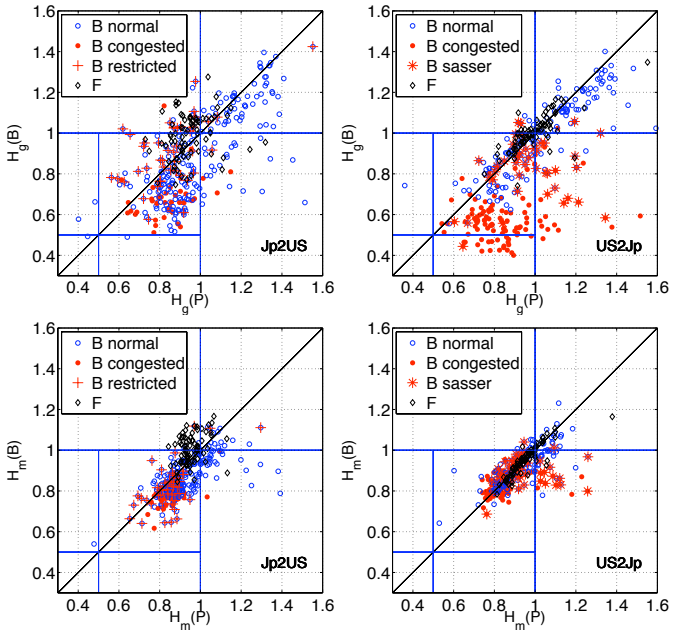


Fig. 7. **Scatter plots of $H(B)$ (byte) vs. $H(P)$ (packet)**. Global (top) and median-sketch (bottom) estimates ; Jp2US (left) and US2Jp (right). Symbols are: **o**: **B** without congestion; • : **B** with congestion; *: **B** anomaly (US2Jp) or low-volume restricted traffic (Jp2US); ◇: **F**.

outliers (e.g., anomalies). The choice of M obviously resorts to a trade-off: Larger M decreases the impact of outliers (hence of anomalies); However, larger M also implies less traffic in each output and hence a larger inter-sketch variability and larger confidence intervals for estimates. Empirical investigations yield $M = 8$ as satisfactory. The procedure is also consistent with networking issues: Sketches preserve flow structure and can be confronted to flow sampling tools such as NetFlow and sFlow.

D. Seven Years of Results

Results obtained with the robust median-sketch analysis from the 7-year long dataset are now discussed.

Constancy along time and global fluctuations: There has been a perplexingly large range of estimated H reported in the literature over the years from various traffic analyses, leading to the conclusion that LRD is a versatile property. Whereas the estimate H_g computed from different days over the entire dataset show large and wild fluctuations (cf. Fig. 6), conclusions drawn from the median-sketch procedure are markedly different. Median LDs remain with a constant knee-type shape over the entire period. The separation scale $2^{j_*}\Delta_0$ is constantly in $[0.5, 1.5]$s and the median based estimate H_m of H are almost always in the range $0.8 \leq H_m \leq 1$ (cf. Fig. 6), thus confirming a strong and persistent LRD.

Anomalies and Congestions: A number of estimated H_m depart from the range $0.8 \leq H_m \leq 1$ (e.g., during the Sasser activity). Obviously, if traffic mostly consists of anomalies, or if they are a dominant part of the traffic, estimates will be impacted and the proposed procedure cannot help when using a single hashing key (IPdst here). Robustness against those anomalies can be achieved by taking the median over estimates computed from different hashing keys (not shown here). Turning now to congested periods (notably US2Jp, bytes) indicate that the global estimate H_g that is constantly close to 0.5 would erroneously validate the claim that congestions induce the disappearance of LRD. Instead, median based estimates speak for the persistence of a strong LRD ($H_m \simeq 0.8$). The network mechanisms causing LRD [12], [21] are not altered by congestion occurrence, neither is traffic returning to a simple Poisson process. Indeed, qualitative analyses indicate that there is no major change in the heavy-tail distributions of the number of packet per flow, hence no change in LRD (quantitative analyses of the heavy tail index are not possible because of the 15-min. observation duration).

Bandwidth and bandwidth occupancy rate: Fig. 1 shows that the bandwidth occupancy rate has been regularly increasing on **B** (Jp2US) over the years up to saturation. Meanwhile, H_m remained fairly constant. Also, the switch from **B** to **F** is accompanied with a significant increase in bandwidth. Fig. 6 indicates that the H_m for **F** are systematically closer to the upper bound of (yet within) the range $0.8 \leq H_m \leq 1$. This suggests that bandwidth and/or bandwidth occupancy rate changes do not cause nor suppress LRD and only marginally impact the LRD parameter: Low bandwidth occupancy rate favoring (slightly) higher H.

Bytes vs. Packets: Another debate regarding LRD consists of deciding whether it should be measured on packet or byte counts, or both. This is examined by means of scatter plots, Fig. 7: $H(B)$ (byte) vs. $H(P)$ (packet). For the global H_g estimates (top), a large variability and dispersion are observed, explained both by numerous outlier (anomaly) days and long congestion periods yielding unreliable estimates for $H(B)$. This would lead to conclude that LRD observed on both packet and byte counts are only partially related, suggesting that they may be induced by different mechanisms. Considering instead the median-sketch estimates H_m (bottom) reveals a much clearer dependence, with $\rho_m \simeq 0.95$ indicating $H_m(B) \simeq H_m(P)$. This confirms, experimentally from real data, conceptual analyses of [12] or models (e.g., [24]) that predict the same Hurst exponent for packet and byte counts.

Summary: The median-sketch based analysis of the 7-year long dataset demonstrates that the LRD paradigm is a relevant and central feature of Internet traffic statistics, even during congestion or traffic restriction periods or anomaly occurrences. It also shows that the Hurst parameter remained constant, and high, $0.8 \leq H \leq 1$, along the years. It tends to be slightly modulated by the bandwidth occupancy rate (loaded link yields estimates closer to 0.8). Hence, this shows that LRD is not suppressed, nor even diminished, with increased bandwidth or statistical multiplexing. Moreover, knee-shaped LDs (and LRD) were reported in [25] for traffic splitting or merging at non congested routers. Our result complements this by showing this is still valid on a link under congestion caused by traffic merging.

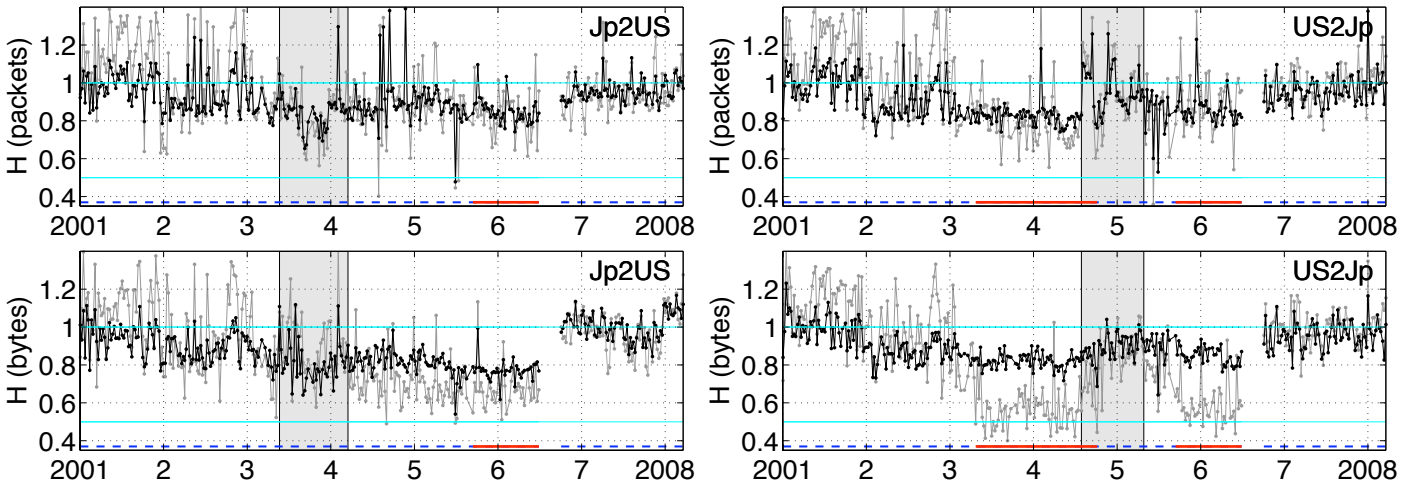


Fig. 6. **LRD vs. years.** Global (gray lines) and median-sketch (on IPdst) (black lines) estimates of H vs. years 2001-2008.

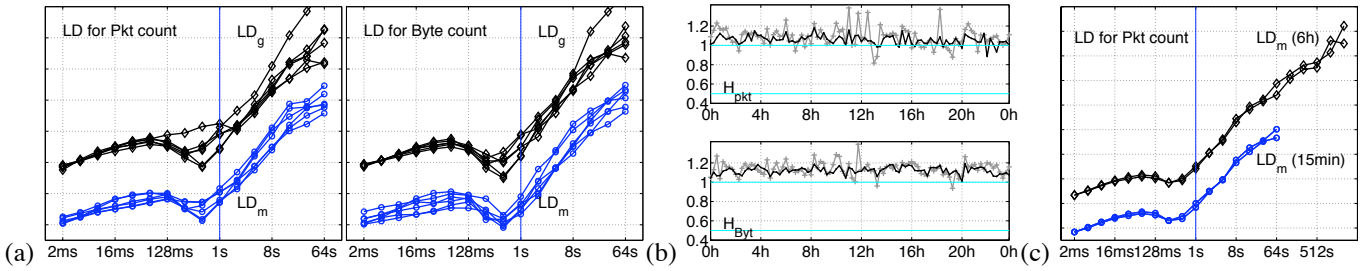


Fig. 8. **One day: Statistics** (Jp2US). (a) LDs for 15min long traces (Left: Pkt; Right: Byte): Global (black ⋄) and median-sketch (blue •). (b) Estimates for H vs. time: Global (gray), median-sketch (black). (c) LDs for 6h-long traces (Jp2US): Average LD over 15-min long traces (blue •) and LD computed fully on 6h trace (black ⋄).

E. Results on a 24-hour long trace

A 24-hour long trace was collected on 2008/03/19, within the *Day in the life of Internet* project [4]. By analysing it, we address i) representativity of 15-minute long trace vs. intraday variability or volume trends, and ii) stationarity over periods longer than 15 minutes.

Intraday variability: Splitting the 24-hour long trace into 15-minute long sub-traces enables consistent comparisons against previous results. MDs are satisfactorily modeled with Gamma laws. LDs systematically present the usual knee-type shape, with $\Delta_0 2^{j_*} \simeq 0.5$s. After normalization to accout for the smooth day-night modulation of the traffic volume, LDs still reveal significant variability around the knee-type shape, cf. Fig. 8(a). This is further confirmed in Fig. 8(b) where the fluctuations of the estimated Hs are large. This is consistent with the fact that, continuously along the day, a number of low volume various anomalies are detected. Applied to the 96 sub-traces, the median-sketch procedure produces LDs almost superimposing one onto the others and hence estimates for H with far less variability. Again, our findings are that there is a strong and persistent LRD (with constant Hurst parameter) irrespectively of the time of the day, and that byte and packet count median-sketch based estimates for H are closely tied together (scatter plots not shown).

Stationarity time scale: Another debate is to question the existence of LRD w.r.t. non stationary effects. Following [16], the 24-hour trace is split into adjacent and non-overlaping sub-traces over which LDs are computed independently. Fig. 8(c) illustrates that the stationarity hypothesis cannot be rejected for time scales up to at least 2h $= 2^{21}\Delta_0$, and hence shows that LRD can not be confused with any spurious non stationarities: LRD measured on 15-min traces (in the range 1s to 1min) clearly and consistently expands at coarser scales (1 min to 1h), confirming its existence and hence the meaningfulness of the estimates reported in Sec. IV-D, even from short duration traces.

Conclusions: Our findings are in favor of strong and constant LRD irrespectively of the time of the day. The trace actually collected lasted 72 hours. The other 48 hours yield equivalent conclusions. The MAWI datasets contain other long traces (one each year) whose analyses yield similar conclusions: LRD is stable and constant within days and throughout years.

V. Conclusion

A unique day-by-day longitudinal analysis of a 7 year (and one day) long dataset shows that the estimations of traffic statistics exhibit a huge variability, largely due to traffic condition variations (congestions, restrictions,...) and

anomalies constantly but randomly occurring. This impairs the possibility of drawing long term evolution conclusions. Therefore, our first contribution is methodological: The recourse to an estimation procedure based on sketches and median average brings robustness for estimations, and enables to disentangle long time evolution from day-by-day incidental variabilities. Our second contribution is to show that the statistics of Internet traffic remain stable along the entire period. LRD (for both packet and byte) remains strong and persistent: The LRD onset scale of time is always between 0.5s and 1s. Moreover, LRD persists over hours and the LRD parameter H for bytes and packets are usually identical. The same network mechanisms are creating a unique LRD phenomenon over both count time series. The robust analysis also showed that despite a significant reduction of volume variability during congestion periods, traffic still presents a strong and clear LRD. MDs remains constantly well modeled with Γ distributions, and there is no evidence for a return to Gaussianity. Non Gaussianity and LRD (and hence *Poisson modeling failure*) are still a forefront property for traffic and network engineering, even when the capacity or the loads of the links are significantly increased. Our conclusions also open rooms for further investigations: Could the bandwidth occupancy ratio be a key control parameter rather than the absolute statistical multiplexing gain? May an increase of any of them be accounted for by a simple shift in time scales?

At the application level, traffic proportion are also stable, despite the intuitive and heuristic claims often made, forecasting dramatic changes in Internet traffic. The dominant uses remain web traffic and P2P applications, with slowly shifting to P2P modalities (higher ports,...) in the recent years. Surprisingly, traffic has been found to contain each and every day (for 7 years) a large number and a variety of anomalies. This significantly questions the notion of *normal* or *regular* traffic, and put the emphasis for the need and benefits of the proposed robust median-sketch estimation procedure.

REFERENCES

[1] K. Cho, K. Mitsuya, and A. Kato, "Traffic data repository at the WIDE project," in *USENIX 2000 Annual Technical Conference: FREENIX Track*, Jun. 2000, pp. 263–270.

[2] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-based change detection: Methods, evaluation, and applications," in *IMC'03*, 2003, pp. 234–247.

[3] S. Muthukrishnan, "Data streams: Algorithms and applications," in *ACM SIAM SODA*, Jan. 2003, p. 413.

[4] kc claffy, "A day in the life of the internet: Proposed community-wide experiment," *ACM Comp. Com. Rev.*, vol. 36, no. 5, pp. 39–40, 2006.

[5] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Trans. Network.*, vol. 5, no. 6, pp. 835–846, 1997.

[6] T. Karagiannis, A. Broido, N. Brownlee, kc claffy, and M. Faloutsos, "Is P2P dying or just hiding?" in *IEEE GLOBECOM'04*, 2004.

[7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated video system," in *IMC'07*, 2007, pp. 1–14.

[8] kc claffy, G. C. Polyzos, and H.-W. Braun, "Tracking long-term growth of the nsfnet," *Communications of the ACM*, vol. 37, no. 8, pp. 34–45, 1994.

[9] M. Fomenkov, K. Keys, D. Moore, and k claffy, "Longitudinal study of internet traffic in 1998-2003," in *WISICT'04*, 2004.

[10] A. Feldmann, A. C. Gilbert, W. Willinger, and T. Kurtz, "The changing nature of network traffic: Scaling phenomena," *ACM Comp. Com. Rev.*, vol. 28, pp. 5–29, 1998.

[11] M. Allman, V. Paxson, and J. Terrell, "A brief history of scanning," in *IMC'07*, 2007, pp. 77–82.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic," *IEEE/ACM Trans. Network.*, vol. 2, no. 1, pp. 1–15, 1994.

[13] V. Paxson and S. Floyd, "Wide area traffic: The failure of poisson modeling," *IEEE/ACM Trans. Network.*, vol. 4, no. 3, pp. 209–223, 1995.

[14] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queuing analysis with long-range dependent packet traffic," *IEEE/ACM Trans. Network.*, vol. 4, no. 2, pp. 209–223, 1996.

[15] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of ethernet lan traffic at the source level," *IEEE/ACM Trans. Network.*, vol. 5, no. 1, pp. 71–86, 1997.

[16] D. Veitch and P. Abry, "A statistical test for the time constancy of scaling exponents," *IEEE Trans. Signal Proc.*, vol. 49, no. 10, pp. 2325–2334, Oct. 2001.

[17] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry, "Non gaussian and long memory statistical characterisations for internet traffic with anomalies," *IEEE Trans. Dep. and Secure Comp.*, vol. 4, no. 1, pp. 56–70, Jan. 2007.

[18] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "Internet traffic tends toward poisson independent as the load increases," in *Nonlinear estimation and classification*, C. Holmes and et al., Eds., 2002.

[19] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido, "A nonstationary poisson view of internet traffic," in *IEEE INFOCOM'04*, 2004.

[20] W. Willinger, D. Alderson, and L. Li, "A pragmatic approach to dealing with high-variability in network measurements," in *IMC'04*, 2004, pp. 88–100.

[21] P. Loiseau, P. Gonçalves, G. Dewaele, P. Borgnat, P. Abry, and P. Vicat-Blanc Primet, "Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility," *preprint*, 2008.

[22] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: multilevel traffic classification in the dark," in *SIGCOMM'05*, 2005.

[23] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho, "Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedure," in *SIGCOMM LSAD'07*, 2007, pp. 145–152.

[24] N. Hohn, D. Veitch, and P. Abry, "Cluster processes, a natural language for network traffic," *IEEE Trans. Signal Proc.*, vol. 8, no. 51, pp. 2229–2244, Oct. 2003.

[25] N. Hohn, D. Veitch, and T. Ye, "Splitting and merging of packet traffic: measurement and modeling," *Performance Evaluation*, vol. 62, pp. 164–177, 2005.

[26] M. Thorup and Y. Zhang, "Tabulation based 4-universal hashing with applications to second moment estimation," in *ACM SIAM SODA*, Jan. 2004, pp. 615–624.