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

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

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

Pierre Borgnat[(1)], Guillaume Dewaele[(1)], Kensuke Fukuda[(2)], Patrice Abry[(1)], Kenjiro Cho[(3)]
[(1)]Univ. Lyon, Lab. Physique ENS Lyon-CNRS, France [(2)]National Institute of Informatics, Japan [(3)]Internet Initiative Japan
[(1)] [(2)] [(4)]{firstname.lastname}@ens-lyon.fr, [(3)]kensuke@nii.ac.jp, [(5)]kjc@iijlab.net

ABSTRACT

It is often stated that *the Internet is a living beast*, constantly evolving with time, the implicit corollary being that its robust and sustainable analysis and modeling are impossible and that obtained results may prove to be outdated before being published. This paper aims at investigating this statement on a scientific basis. A longitudinal evolution study is performed over the traffic collected every day for seven years on a trans-Pacific backbone link (the MAWI dataset). Analyzing this unique dataset enables us to investigate long term characteristics of traffic evolution, both at the TCP/IP layers (packet and flow attributes) and application usage. This provides new insights into central issues, specifically: *Does link bandwidth increase and statistical multiplexing induce an evolution of traffic towards Poisson, Gaussianity and weaker Long Range Dependence?* Traffic in the MAWI dataset is subject to bandwidth changes, congestions, and to a number of short and long lived anomalies. This allows a comparison of their impacts on traffic statistical properties and yields an overview of traffic anomaly evolution. Also, we show and explain how and why random projection (sketch) based statistical procedures offer an efficient and robust tool to disentangle the impacts of actual long term evolutions from those of time localized events (anomalies and/or link congestions). A study of a 24-hour trace collected under the *A Day In the Life of the Internet* project on March 19th, 2008 complements these results with an understanding of the typical intraday traffic variability.

General Terms

Internet traffic; Analysis; Measurement; Modeling

Keywords

Longitudinal study; Sketch; Robust estimation; LRD

1. INTRODUCTION

One cannot attend a conference dedicated to the analysis of the Internet without hearing that *the Internet is a living beast, subject to numerous, violent, constant and rapid changes*, the underlying implication, often implicit, sometimes explicit, being that statistical analyses aiming at traffic modeling, resource optimization or anomaly detection proposed at a specific times, for specific links, are likely to be out-of-date and irrelevant by the time they are published. And, one can not deny that time scales in Internet science evolution are much faster than those of most, if not all, more traditional sciences. Also, there is a consensus around observations of strong and net changes in traffic volumes, link bandwidth or capacity, number of Internauts, types of applications and usages on the Internet over the last decade. Despite its apparent obviousness, this statement remains yet mostly informal and global hence vague. It often rises more questions than brings actual answers: What does actually change? What does not? Are changes mostly at the application layer or rather at the TCP/IP layer? Do they impact traffic statistical modeling procedure, anomaly detection schemes or network engineering procedures? Is the impact moderate or so drastic that last-year findings are useless today?

This paper aims at investigating some of the realities beyond this statement on a scientific basis and at proposing both methodological tools and objective elements of answer to shed light on these issues. To this end, we analyze traffic traces collected every day, for 15 minutes, from 2001, Jan. 1st, to 2008, March, 19th, over trans-Pacific backbone links (the MAWI repository, details in Sec. 3.1). Traffic circulating over such links are highly asymmetric as clients on the Japanese side mostly consist of academics while those on the US side are mostly on commercial ISPs. Analyses are hence conducted independently for the directions US to Japan (US2Jp) and Japan to US (Jp2US). It is commonly accepted that commercial and academic traffic differs in application and protocol mixture proportions, so that conclusions drawn from the MAWI dataset can be considered general. Moreover, to the best of our knowledge, no commercial traffic database collected over such a long period of time is publicly available.

Longitudinal analyses of long-term evolutions of traffic throughputs and protocol/application breakdown will yield our first conclusion: protocol and application characteristics remain surprisingly stable along the years (cf.

[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`

Sec. 3.2 and 3.3).

Then evolutions of the statistical characteristics relevant to the TCP/IP layer (packet or byte count aggregated time series) are analyzed. Following the framework proposed in [8, 26], marginal distributions (MD) are modeled as Gamma laws, and covariances, hence long range dependence (LRD), is analyzed by means of the standard wavelet based methodology proposed in [28] (technical material recalled in Sec. 4.1). A great difficulty in conducting a statistical longitudinal analysis of traffic consists in disentangling smooth long term evolution features from wild variations among day-to-day fluctuations that are likely to occur because there is no single day without anomalies or specific events. The risk is then strong that a longitudinal study actually boils down to the long report of a collection of singularities. This is discussed in Sec. 4.2 from examples chosen in purpose within the dataset. A major contribution of our proposal therefore lies in the construction of a robust estimation procedure based on sketches (or random projections) [22, 24]. It is shown in Sec. 4.3 how and why it enables to conduct long term analyses that are not polluted by specific traffic conditions or the occurrences of anomalies. Applied to the 7-year long datasets, this robust estimation procedure enables us to bring new insights, both in terms of methodology and of results, to the on-going debate related to *bandwidth increase and statistical multiplexing causing a return to Poisson and Gaussian together with the disappearance of long range dependence*: Traffic is characterized with a strong, stable and persistent LRD (cf. Sec. 5.1); MDs are constantly well modeled with Gamma laws along years, hence enabling us to revisit the Gaussianity issue (cf. Sec. 5.2). The MAWI datasets contain periods of congestions and of restricted traffic which permit to analyze their respective impacts of traffic statistics, on Gaussianity and long range dependence. Anomaly detection is automated using the sketch-multiresolution procedure previously proposed in [8]. Evolutions in direction, type, nature and number of anomalies are depicted, and related anomalies are discussed in Sec. 5.3.

Another pitfall comes from the fact that collected data last only 15 minutes, starting systematically at 2:00 pm: One may hence question representativity w.r.t. the natural intra-day variability as well as w.r.t. short duration. To address these points, a 24-hour trace collected on March, 18-19th 2008, within the framework of the world-wide *A day in the life of the Internet* project [20], is analyzed. This allows an illustration of the robustness of our sketch based statistical analysis procedure w.r.t. local anomalies. Also, it shows that our conclusions are not biased neither by this specific schedule nor by short measurements (cf. Sec. 6). Conclusions are drawn in Sec. 7.

2. RELATED WORK

Long Range Dependence: The discovery of LRD in Internet traffic constitutes one of the most epoch-making and fundamental issues in the recent Internet traffic research [23, 25]. Specifically, a striking characteristic related to LRD lies in the high variability of traffic fluctuations, yielding degradations of queueing performance [9]. Difficulties in empirically assessing LRD in real traffic time series have been thoroughly discussed [18, 28], showing the relevance of a wavelet-based analysis framework [28]. A number of authors discussed the fact that LRD in Internet traffic can be induced by higher-layer protocols [11, 31], as well as related to the heavy tail natures of the distributions of the file size to be transferred [7, 23]. The (heavy)-tail behaviors of IP flow size have been continuously investigated (see e.g., [30] for a recent report). Moreover, the impact of bottleneck and congestion on the existence and strength of LRD has been investigated in, e.g., [29].

Back to Poisson traffic? However, stability (or even existence) of LRD traffic is an ongoing debate and a hot research topic. An open issue is the conjectured disappearance of LRD as network loads increase, often referred to as the impact of *statistical multiplexing* [4]. Equivalently, it is often stated that traffic tends to return to Poisson for high network loads: for instance, Refs. [3, 4, 19] indicated that packet inter-arrival distributions tend to be well modeled with a Poisson process, when backbone loads increase. An important issue is in fact the time scales at which traffic is analyzed [13, 32]. For instance, it has been suggested that packet arrivals in recent backbones are well-modeled by the Poisson model for sub-second timescales, by nonstationarity at multi-second timescales, and that aggregated traffic still exhibits LRD at large timescales [19]. Zhang et al. [32] pointed out that backbone traffic time series are weakly correlated at fine time scale (1-100ms). These complex inter-relations between time scales have been interpreted in terms of multifractal (e.g., [10]). However, Ref. [14] made a strong case against multifractal models for Internet aggregated traffic and Ref. [13] indicated that small scales are related to the flow packet arrival process, which is consistent with a simple renewal process.

Gaussianity? In Ref. [32], it is shown that backbone traffic aggregated over 1s exhibits Gaussian-like marginals. Evolution toward Gaussianity is indeed a consequence of the return to Poisson and, as such, also been investigated. For instance, HTTP connection inter-arrival distributions may be modeled with Weibull laws, as the shape parameter is versatile enough to adjust both Gaussian and heavier tail laws [10]. Ref. [21] analyzed the Gaussianity of traffic MDs at a given time scale, and discussed the evolution toward Gaussianity with respect to aggregation levels and link loads. Fur-

thermore, Refs. [8, 26] showed that the marginal distribution of normal packet arrivals is well-modeled with Gamma laws at various scales. This is used here to investigate the evolution toward the Gaussianity issue.

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 major changes in traffic statistical characteristics, e.g., web [7], P2P [1, 17], video streaming [5],... There have been fewer studies aiming at quantifying the long term evolution of Internet traffic (statistics and applications). One of the oldest such reports is based on NSFNET traces (1988-1993) [16]. At that time, FTP and Mail traffic accounted for about 50% of the growing traffic volume, until web traffic becomes majority. In Ref. [12], a relation between packet rate and bit rate is investigated together with traffic statistical properties, based on more than 4000 traces collected from 1998 to 2003. Also, in Ref. [11], the correlation structures (from traffic collected at several measurement points) before and after the web are compared. The Hurst parameter evolution is not reported; However, it is pointed out that web traffic affects at least the fine time scales. For anomalies, an evolution of scanning activities through the LBL network for the past 12.5 years has been highlighted in [2].

Contributions: Ref. [30] pointed out the need for a general robust methodology to provide answers to these issues. In this spirit, we propose a median-sketch based method and provide analyses of the joint and long term evolutions for a few key statistical parameters, related to applications, protocols, anomalies, with a focus also on statistics (LRD, MD,...).

3. MAWI DATASET

3.1 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 to the public since 1999, and the total volume of the publicly available data exceeds 1TB as of April 2008 (cf. `http://mawi.wide.ad.jp/` and [6]).

Our main datasets are daily packet traces captured at Samplepoint-**B** (hereafter **B**) from 2001 to 2006/06, then at Samplepoint-**F** (hereafter **F**) from 2006/10 to 2008. These are transit links of the WIDE network, and the link of **B** was replaced in July 2006 by the link **F**. (However, traces just after the upgrade are missing unitl 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 in Japanese Standard Time (UTC+9), and the corresponding traces with IP addresses anonymized and payloads removed are made available to the public along with a summary information web page about the traffic. Occasionally, 24-hour or longer traces for these samplepoints are made available in the same manner.

The traffic of the WIDE transit link is mostly trans-Pacific commodity traffic between Japanese research institutions and non-Japanese commercial networks, as WIDE peers with all the major domestic ASes at the Internet Exchange Points it operates, and international traffic between academic networks goes through other international research networks. The traffic of the transit link is also asymmetric as WIDE has other trans-Pacific links, meaning that many flows can be observed only in one direction. This forces us to study traffic separately for each direction, being 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 contains various kinds of anomalies.

The datasets allow us to examine the evolution of the traffic over 7 years, under both congested and over-provisioned conditions. We also use 24-hour-long traces collected at **F** on 2008/03/19 to show in Sec. 6 that our findings are consistent with other time slots.

3.2 Throughput Evolution

Strong variability: Fig. 1 displays throughput evolutions, in bytes and packets, for both directions, and their intraday variabilities (measured by means of standard deviations (STD) computed around 1s time window averages). The first striking feature lies both in the wide range of observed throughput values and in their huge intraday variabilities (STD varies by a factor of 10). Also, there is a global increase of throughput from 100 kbps in 2001 to slightly 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).

Congestion periods: **B** experienced several long periods of congestions (marked with solid lines 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 periods of congestion are accompanied with a significant drop in STD (by a factor of 5 to 10). This means that the byte throughput remains close to a constant value with a very low level of fluctuations (this is important when discussing LRD in Sec. 4). Note that these drops in STD are long in duration, but their amplitude is not sufficient to detect congestion periods as short time fluctuations are of the same order.

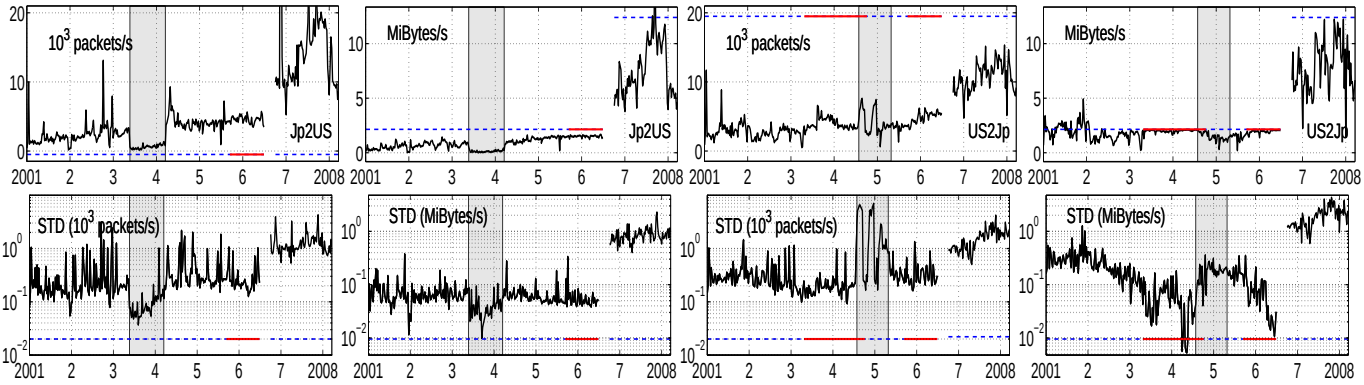

Figure 1: **Throughput vs. years.** Top: Pkt and Byte throughputs. Bottom: Intraday variability (measured in log of the standard deviation, computed from 1s time windows). Congestion periods are symbolized by solid lines beneath, or at the respective CAR bandwidth limitations for byte troughput. Left: Jp2US; Right: US2Jp.

Specific periods: Two periods with unusual traffic behavior (gray-shaded areas in Fig. 1) necessitate specific comments. From 2003/05 to 2004/03, Jp2US traffic underwent a severe volume decrease (Fig. 1, left). It is suspected that this volume reduction may have been caused by a change in the routing policy or by upward link congestions. Interestingly, we found that, despite this low volume, the traffic composition and its statistical characterization have not been significantly affected.

From 2004/07 to 2005/04, US2Jp (right on Fig. 1), strong fluctuations in packet number are observed (STD being extremely high). Our anomaly analyses reported in Sec. 5.3 enable us to conclude that it corresponds to a period of massive activities of the Sasser worm that strongly impacts traffic statistical characteristics.

3.3 Protocol and Application Breakdown

Anomalies will be discussed in details in Sec. 5.3. In the current section, we concentrate on traffic that can be regarded as legitimate.

Protocols: Over the 7-year period, in both directions, TCP and UDP continuously conveyed more than 90% of packets. ICMP (Ping) presents a noticeable share of the packets, and more frequent for Jp2US traffic ($\simeq$5%). At **F**, the situation is more even in this respect. Many INSLP packets (a security layer protocol) are also found in US2Jp traffic, from 2001 until early 2004. GRE (an encapsulation protocol) packets make a noteworthy part of the Jp2US traffic, notably around mid-2005 (more than 5% of traffic).

TCP/UDP contents: Details on the breakdown of TCP/UDP packets is provided in Fig. 2. A majority of them consists of Web traffic: At **B**, 40% 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 related to Peer to Peer (P2P) exchanges. Besides HTTP and P2P, common Internet services such as FTP, mail (SMTP, POP, IMAP,...), news protocols, etc. account together for around 5%; this ratio remains fairly stable over the years. Most of the remaining TCP/UDP traffic is targetting Microsoft services (such as MS RPC, MySQL, file sharing). This amounts to a couple percent of the US2Jp traffic. Those are probably anomalous and will be discussed further later on. Broadcasting protocols such as Realserver or Shoutcast also represent 1 or 2%. Finally, a quite large number of DNS packets are also present: At **B** Jp2US, from 2001 to 2006, DNS traffic is larger ($\simeq$15%) than for US2Jp ($\simeq$5%). For **F**, the situation is nearly inverted, likely due to the anycast deployment of M-root DNS server operated by WIDE.

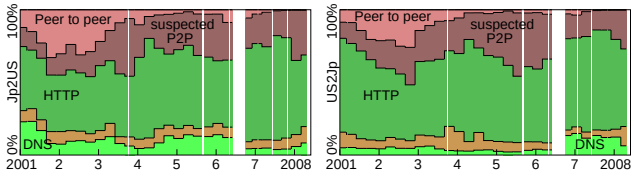

Figure 2: **Application breakdown vs. years.** Relative amounts of protocols (in legit traffic). Bottom to top: DNS (light green), common services (SSH, FTP, mail) (orange), HTTP (green), suspected P2P (red) and identified P2P (light red). Left: Jp2US; Right: US2Jp.

Peer to Peer: Traffic going to or from usual P2P ports constitutes around 30% of packets in each direction in 2002. At that time, the most popular protocol was Napster. Others such as Gnutella and its clones, Kazaa, WinMX, and Emule-Edonkey are also identified. This identified P2P traffic tends to disappear over the years and is quasi invisible since 2004. Some P2P traffic is still identified, mostly Bittorent (only around 2% of the traffic). However, the decline of the P2P traffic is only apparent and actually corresponds to a *P2P hiding* phenomenon [17]. A naive identification method based on source and destination port recognition no longer

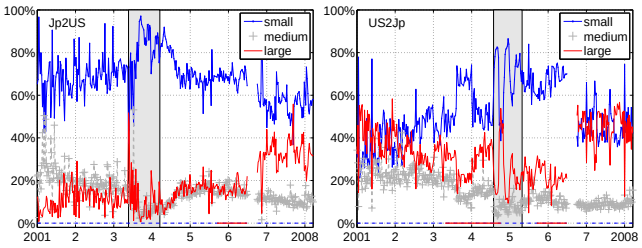

Figure 3: **Packet sizes vs. years.** Proportion of large ($\geq$ 1400B), small ($\leq$ 144B), or medium (in-between) packets. Left: Jp2US; Right: US2Jp.

works. Because of evolving P2P applications, firewall policies, etc., modern P2P software makes use of random ports as opposed to fixed ports. This is confirmed by the observation that a significant increase in TCP and UDP traffic volume between high ports ($\geq$ 4000), in the recent years, correlates with the decrease of identified P2P (cf. Fig. 2). Very few exchanges are conducted between high ports except for P2P, be they data sharing or (for a probably small part only) games. Hence, the conjecture that a large amount of P2P traffic is hidden there. Note that the dataset provides neither access to packet payloads, nor to complete flow information (only one direction of flow is usually available) so that more elaborated identification procedures (such as proposed in [17]) cannot be used here. P2P identification is anyway beyond the scope of this paper. Aggregating identified and unidentified P2P traffic shows (as expected) a significant increase in absolute volume at **F** (compared at **B**), though its relative share decreased slightly. As a side note, around half of (identified or suspected) P2P traffic (in number of packets) is carried by UDP (mostly small to medium size packets), which is usually signaling and request traffic; the other half corresponds to TCP and large packets, hence to file transfers.

Packet size distributions: As expected, most packets are either small ($\leq$ 144B signaling packets) or large ($\geq$ 1400B – usually data frames), cf. Fig. 3. Medium-size packets are less frequent, with a stable proportion since 2005. There is one notable evolution along the years: in the upgrade from **B** to **F**, the proportion of large packets has significantly increased. Also, we confirmed clear appearance of some typical intermediate size of packet over 7 years, probably used by specific applications (e.g., 660B used for P2P software).

Summary: These analyses show that, over the seven years, for both **B** and **F**, the content of (non anomalous) traffic does not change drastically. The protocol/application breakdown reported here well matches those provided in [12] whose traffic collected in 1998-2003. However, they are in clear contrast with those in [16], in the 90es when the majority of the traffic was FTP and email.

4. METHODOLOGY FOR ROBUST STATISTICAL CHARACTERIZATION

Let us now turn to the TCP/IP layer statistical characterization: analyses of aggregated packet or byte count time series, X_Δ and W_Δ. Following most of the studies reported in the literature, we concentrate on marginal distributions (one-point statistics) and on the covariance function (two-point statistics). A perennial issue lies in the determination of the aggregation time scale Δ leading to a relevant statistical description. This is addressed by performing a multiresolution statistical characterization, i.e., analyses jointly at various aggregation levels. Also, 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 Gaussianity and LRD (cf. [18, 23, 25, 26]). Therefore, analyses are confined to the estimation of parameters measuring these properties, overlooking other interesting attributes (e.g., those quantifying short time correlations).

4.1 Statistical description

Marginal distribution and Gaussianity: The marginal distributions (MD) of X_Δ and W_Δ are estimated by means of empirical histograms. They are analyzed 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 [8, 26], we use Gamma laws to model the necessarily positive X_Δ and W_Δ. A $\Gamma_{\alpha,\beta}$ distribution is defined as $\Gamma_{\alpha,\beta}(x) = (x/\beta)^{(\alpha-1)} \exp(-x/\beta)/(\beta\Gamma(\alpha))$. It has mean $\mu = \alpha\beta$ and variance $\sigma^2 = \alpha\beta^2$. While the scale parameter β mostly *feels* the volume of the data, the shape parameter α is used here as an indicator of closeness to Gaussianity. Indeed, skewness and kurtosis (relative third and fourth moments), which are 0 for Gaussian, behave respectively as $2/\sqrt{\alpha}$ and $3 + 6/\alpha$, for Γ. Hence, $1/\alpha$ controls the smooth transition of Γ from exponential to Gaussian. The shape and scale parameters are systematically estimated for Pkt and Byte count aggregated at the different Δ_j, and denoted by α_j and β_j.

Covariance functions and LRD: For stationary processes, two-point statistics are analyzed by means of the covariance function $\mathbb{E}\{X(t)X(t+\tau)\}$ ($\mathbb{E}$ denotes the expectation) or of its Fourier transform, the 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 [23, 28]. It is well-known that LRD is best analyzed in a wavelet framework through the relation: $S_j = (1/n_j) \sum_{k=1}^{n_j} |d_X(j,k)|^2 \sim C 2^{j(2H-1)}$, 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 constitute multiresolution quantities, i.e., 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

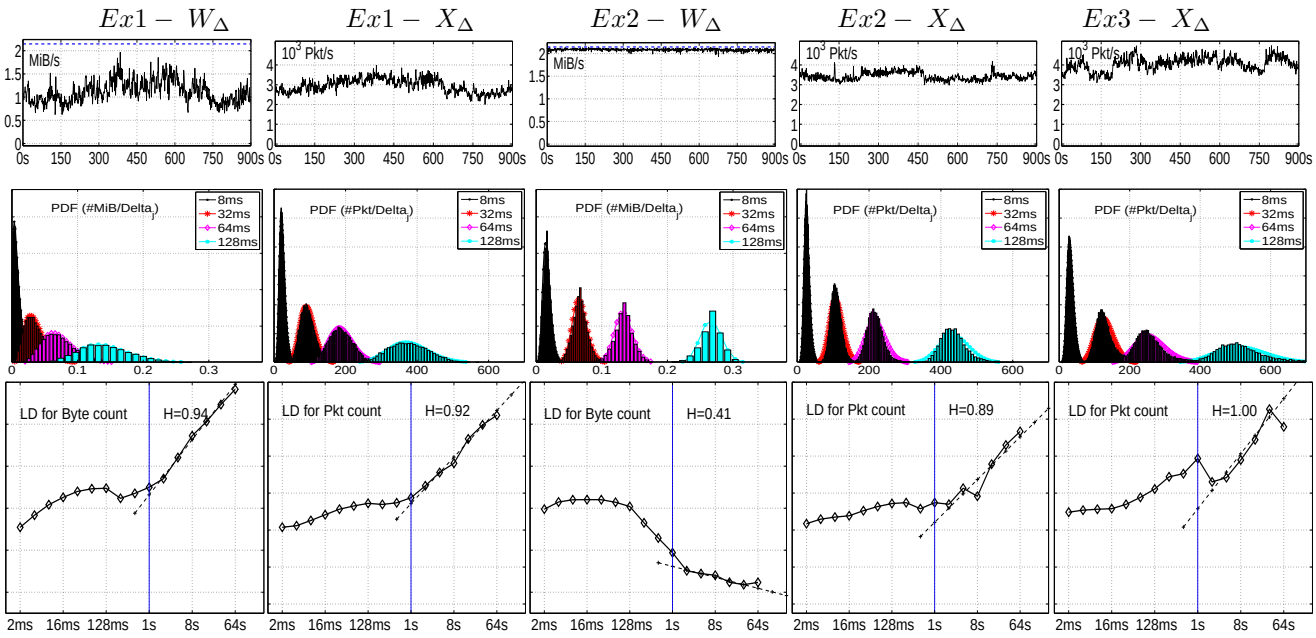

Figure 4: **Statistics in various traffic conditions.** Aggregated ($\Delta_0 = 1$ms) byte (W_Δ) or packet (X_Δ) count time series (top row); Marginal distributions (MD) for $\Delta_j = 4, 32, 64, 128$ms, both as empirical histograms (bars) and Γ fits (lines) (middle row); Logscale Diagrams (LD) (bottom row). $Ex1$ (cols. 1&2): **B**-US2Jp, 2005/07/11, anomaly-free. $Ex2$ (cols. 3&4): **B**-US2Jp, 2003/06/03, congestion. $Ex3$ (col. 5): **B**-Jp2US, 2004/09/21, anomalies (network scan, spoofed flooding, attack on Realserver).

as the basis for Hurst parameter estimation [28].

4.2 Impact of the variety of traffic conditions

Let us now illustrate these two statistical analyses on traffic collected under different conditions. Internet traffic is not intrinsically stationary (daily or weekly seasonality, anomalous events,...). However, for 15-minute long traces, stationarity is fairly well satisfied, hence the MD and LRD analyses described above are valid (cf. e.g., [3, 28] and discussions in Sec. 6).

Example 1 ($Ex1$ in Fig. 4) consists of US2Jp traffic collected on 2005/07/11, chosen because traffic is neither congested or restricted nor with anomalies. A careful human inspection (assisted with the anomaly detection procedure of [8] recalled in Sec. 6) enables us to conclude that it contains only a small number of (low volume) sure or suspected anomalies. Traffic MD at various Δ_j are satisfactorily modeled by Γ laws (2nd row). The LD exhibits a classical *knee-shaped* form that consists of two ranges of scales: 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), separated by a typical scale $2^{j_*}\Delta_0 \simeq 1$s. LRD is evidenced by the linear part of the LD at coarse scales (with estimated $H \simeq 0.95$). Such a shape for LDs is consistent with experimental observations continuously reported in the literature over the years [10, 13, 15, 26, 28] as well as with theoretical models such as Cluster Point Processes [13]. Fine scales are related to the packet arrival process while coarse scales are related to flow characteristics (notably heavy tail packet number distributions).

Example 2 ($Ex2$ in Fig. 4) corresponds to traffic collected under severe byte congestion (2003/06/03, US2Jp). Whereas the MD and LD for X_Δ are similar to those of $Ex1$, clear changes for W_Δ are observed. MD is still well modeled by Γ laws, though α_j and β_j parameters differ from those of $Ex1$ (values not reported, but differences easily inferred from the plots: larger means, but smaller variances). The LD shape is completely altered, notably with a disappearance of LRD at coarse scales. This can be easily interpreted: congestion implies that byte count remains quasi constant (Fig. 4 shown by a significant drop in standard deviation and hence in variability). By construction, no variability implies no LRD. Therefore, $Ex2$ illustrates that congestion impacts both the route toward Gaussianity and the actual value of the LRD parameter or even the existence of LRD. Congestion much less affects Pkt count as its variability is not strongly impacted by byte number saturation.

Example 3 ($Ex3$, 5th col. in Fig. 4) consists of traffic (Jp2US, 2004/09/21) containing several low-volume attacks: SYN flooding ($\simeq 6\%$ of traffic) looking for network open ports, SYN flooding from a single source, spoofed flooding (using spoofed IPsrc, 2% of traffic), attack targetting a Realserver through TCP port 554 (2% of traffic). It shows that LRD is usually not completely altered by the occurrence of (low-volume) 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 the findings in [8, 26]). This change is in agreement with that observed in the fine scale range of the LD. Indeed, a change in α_j for the range of scales $1\text{ms} \leq 2^j\Delta_0 \leq 1\text{s}$ 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 proposed in [8].

Discussion: These examples show that changes in traffic conditions (traffic restrictions, congestions, low-volume anomalies, major period of Sasser anomalies,...) drastically modify the parameters of the statistical modeling. Observations drawn from other days under congestion or with anomalies are always consistent. The study reported in Sec. 6 shows that there exists no day without numerous low-volume anomalies. This is a severe difficulty in performing long term evolution analyses of statistical characteristics of traffic intended here: there is a major risk that the study boils down to a long list of singularities, abnormalities or specific situations, with no possibility to identify *normal days* and hence to extract any global and long term features. This significantly impairs the possibility of performing automatic and unsupervised data analyses (a mandatory requirement to process a 7-year long dataset!) and hence of relevantly addressing issues related to long term evolution toward Gaussianity or LRD decrease or disappearance. Overcoming this difficulty is one of our key contributions and the solution proposed is referred to as a *robust* estimation procedure.

4.3 Sketches for robust estimation

In statistical signal processing, robustness in estimation is classically achieved by performing averages over independent copies of equivalent data. In our case, 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, we turn to the use of random projections (usually referred to as sketches), following the seminal contributions describing the benefits of their use for traffic analyses [22, 24].

Sketches: Let h_n denote a $k-$universal hash table of size M, computed using the fast-tabulation method in [27]. The hashing key A is chosen as one of the packet attributes (IPdst, IPsrc,...). The original collection of packets is then split into M sub-traces, each of them consisting of all packets with identical sketch output $m = h_n(A)$. 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 then aggregated, $X^{(m)}_{\Delta_0}, m = 1, ..., M$, and analyzed following the procedures used for the original trace. Robust estimation results from averaging, by means of *median*, estimates obtained independently from each sketch output.

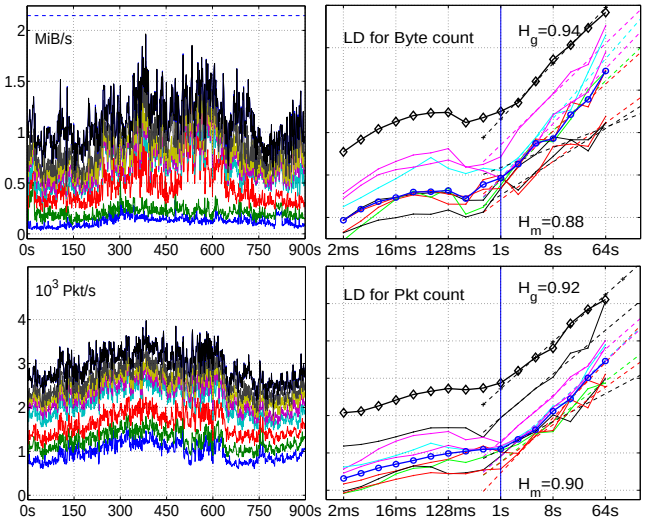

Figure 5: **Robust estimation. B**-US2Jp, 2005/07/11, no congestion: $Ex1$. Top: Bytes; Bottom: Packets. Left: Aggregated time series (1s) displayed cumulatively by sketch sub-trace; Right: LDs for global traffic (thick line, $\diamond$), for sketches (thin lines) and for median-sketch (thick line, o).

Example 1: Fig. 5 shows aggregated sketched sub-traces ($M = 8$) together with their LDs for $Ex1$. For this *quasi anomaly free* day, the M LDs display a weak variability around a well-defined average. Hence, the median LD is close to any of them and matches perfectly (up to a vertical shift, due to the division by M) the LD computed from the entire trace. Notably, comparing the estimates of the Hurst parameter obtained from the whole trace, H_g, from the median of the estimates over the M sketches, H_m, shows that they are perfectly consistent (and within the confidence intervals one of the other). As intuitively expected, all sketches are statistically equivalent. This validates the consistency of the median-sketch estimation procedure. It also shows that flow-sampling is compatible with LD estimates, better than ones being obtained by flow-preserving averages.

Example 2: Fig. 6 shows aggregated sketches and their LDs for $Ex2$ (congestion) (Byte counts only). The striking feature consists of each sub-traces having recovered a significant variability, when the original shows almost none. Accordingly, the sub-trace LDs (and hence their median) changed dramatically in shape (compared to that obtained from the entire trace) and exhibit back the *knee-shape* form with $j_* \simeq 9$ or 10 (0.5s to 1s) and

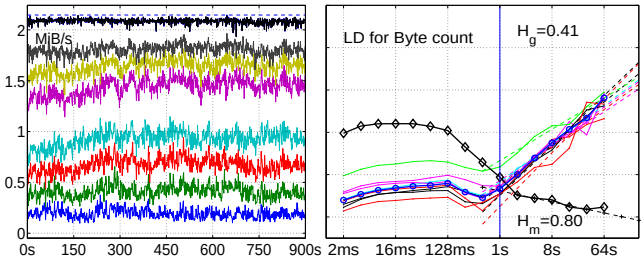

Figure 6: **Robust estimation. B**-US2Jp, 2003/06/03, with congestion: $Ex2$ (Byt). Left: Aggregated sketches (1s); Right: LDs for global traffic (thick line, $\diamond$), sketches (thin lines) and median-sketch (thick line, o).

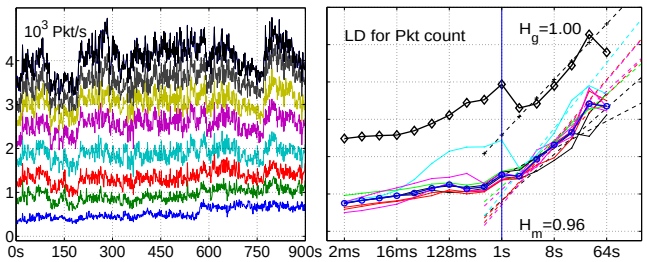

Figure 7: **Robust estimation. B**-Jp2US, 2004/09/21, with severe anomalies: $Ex3$ (Pkt). Left: Aggregated sketches (1s); Right: LDs for global traffic (line, $\diamond$), sketches (thin lines) and median-sketch (thick line, o).

estimated H in the usual range $[0.8, 1]$. This indicates that sub-trace are characterized by a clear and unquestionable LRD. This is a major finding of the present work: the global analysis of a trace under a congested day leads to the erroneous conclusion that congestion eliminates LRD. A sketch based analysis instead clearly reveals that the network mechanisms at work to create LRD remain equally and strongly active under congestion. Moreover, a relevant analysis and estimation of the LRD parameters can be automated by median over sketches. This provides a first justification in favor of qualifying this procedure as *robust*.

Example 3: Fig. 7 shows aggregated sketches and their LDs for $Ex3$ (anomalies). One observes that all sub-trace LDs almost superimpose but 2. Inspection confirms that the LDs resisting superimposition concentrate on the significant anomalies detected that day. Therefore, computing the median of the LDs results in an analysis of the traffic covariance structure as if not impacted by these significant anomalies. The median LD now significantly differs from the one computed from the entire time series, whose shape is mostly dominated by the contributions of the anomalies (global LD is affected by the shape of the dominant sub-trace LD, which may change from scale to scale). The differences are mostly seen in the fine scale range (0.1s to 2s), in agreement with our previous findings: low volume anomalies mostly affect traffic SRD [8]. This also indicates that the median-sketch based procedure provides a relevant estimate for H even when anomalies are present, and hence shows that traffic LRD per se is not affected nor varied by low-volume anomalies. 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 outliers (here anomalies). This raises a question of the choice of the number of sketch outputs M. It 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. This gives a second justification for *robust*: median-sketch procedure provides statistical estimates that are not (or weakly) impacted by low-volume anomalies (be they short or long lived). Obviously, if traffic mostly consists of major anomalies (like the Sasser period already mentioned) producing a dominant fraction of the traffic, estimates will be impacted and the proposed median-sketch procedure cannot help, as cannot any other procedure.

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 conclusions are drawn when studying MDs (and the α_j and β_j parameters). This procedure is also consistent with networking issues: sketches preserve flow structure and hence can be confronted to flow sampling tools such as NetFlow and sFlow.

Note that mitigating anomalies by sketches is achieved only when the relevant hashing key is chosen (e.g., IPdst hashing for scans). This is solved by using several hashing keys in parallel (this is the rationale behind the detection procedure described in [8] and used in Sec 6).

5. SEVEN YEARS OF RESULTS

When applied to the 7-year long MAWI dataset, the robust median-sketch analysis procedure yields the following results and conclusions.

5.1 Long Range Dependence

Constancy along time and global fluctuations: The significant variabilities of the LDs computed from the different days of the entire dataset yield large and wild fluctuations along time for the estimate H_g (cf. Fig. 8). This could incite to conclude that LRD is a versatile property significantly affected by changing traffic conditions and anomaly occurrences. This may (partly) explain the perplexingly large range of estimated H reported in the literature over the years from various traf-

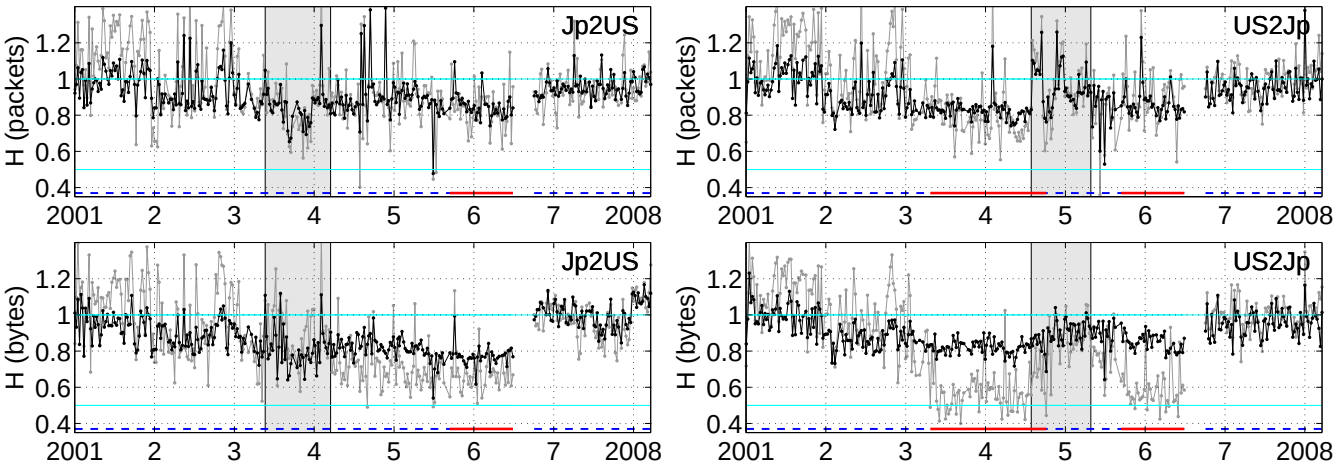

Figure 8: **LRD vs. years.** Global (gray lines) and median-sketch (black lines) estimates of H vs. years 2001-2008.

fic analyses. Conclusions drawn from the median-sketch procedure are markedly different and accurate. Median LDs (comparable to those shown in Fig. 5 to 7) remain with a strikingly 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 remains almost always in the range $0.8 \leq H_m \leq 1$ (cf. Fig. 8)), indicating thus a strong and persistent LRD.

Anomalies: A number of estimates H_m still significantly depart from the range $0.8 \leq H_m \leq 1$, for instance, during the Sasser period (Fig. 8, top right, period in gray). As explained, anomalies constitute most of the traffic during that period so that no robust estimation can be achieved. It has been manually checked that the small number of residual large values for H_m is explained by the use of a single hashing key (IPdst) to obtain the plots in Fig. 8. Robustness against some classes of anomalies is not achieved in this case and would require taking the median over estimates computed from different hashing keys (not shown here).

Congestions: Analyses of congested periods (notably US2Jp, bytes, Fig. 8, bottom right) indicate that the global estimate H_g are constantly close to 0.5 erroneously validating the claim that congestions induce the disappearance of LRD. Instead, even if closer to the lower bound $H \simeq 0.8$, median based estimates, clearly speak for the persistence of a very strong LRD. Therefore, the network mechanisms causing LRD are not (significantly) altered or modified by congestion occurrence, and the traffic is not returning to a simple Poisson process. Notably, the celebrated result by Taqqu et al. [23] relates LRD to the heavy-tail nature of the number of packet per flow distributions. Qualitative analyses enable us to indicate that congestions induce no major change in the shape of such distributions, hence no change in LRD. Quantitative analyses relating H to the heavy tail index are not possible because the 15-minute duration is *too short*. Ref. [15] indicates that knee-shaped LDs (and LRD) were found on traffic splitting or merging at non congested routers. Our result complement this by showing this is still valid on a link under congestion caused by traffic merging.

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. 8 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 ongoing 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. 9: $H(B)$ (byte) vs. $H(P)$ (packet). For the global H_g estimates (top row), despite a significantly positive correlation coefficients $\rho_g \simeq 0.65$ (both directions), 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 row) reveals a much clearer dependence, with $\rho_m \simeq 0.95$ indicating $H_m(B) \simeq H_m(P)$. This suggests that the same network mechanisms are at work to create the same LRD phenomenon in both byte and packet counts. This is consistent with Taqqu's fundamental theorem [23] as well as with some traffic models (e.g., [13]) usually pre-

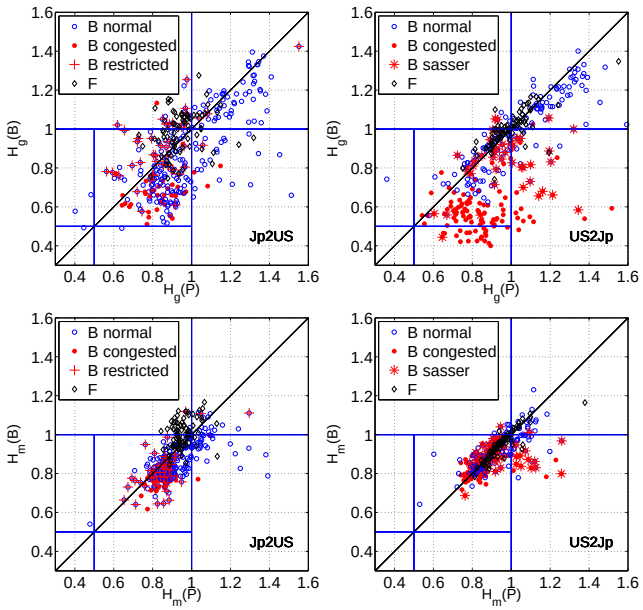

Figure 9: **Scatter plots of $H(B)$ (byte) vs. $H(P)$ (packet).** Global (top) and median-sketch (bottom) estimates. Symbols are: **o**: **B** without congestion; **•** : **B** with congestion; **+**: **B** anomaly (US2Jp) and restricted traffic (Jp2US); ⋄: **F**. Left: Jp2US; Right: US2Jp.

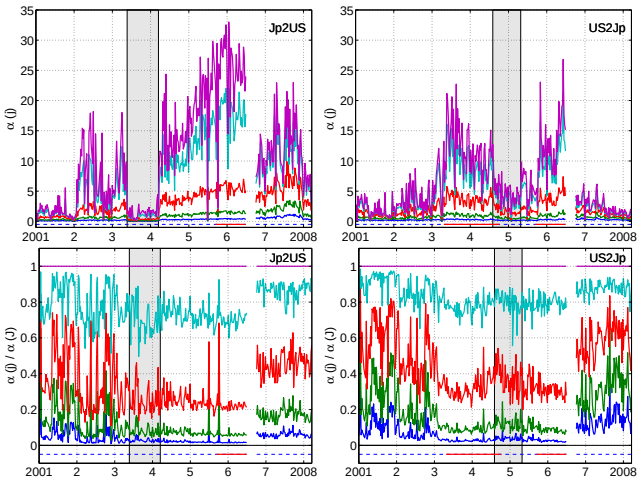

Figure 10: **Indices for closeness to Gaussianity.** Top: indices α_j, as a function of time, for different values of $j = 2, 4, 6, 8, 9$. Bottom: normalized indices $\alpha'_j = \alpha_j/\alpha_J$ ($J = 9$). Left: Jp2US; Right: US2Jp.

dicting the same Hurst exponent for packet and byte counts. Our findings confirm experimentally from real data these conceptual analyses, and complement them, showing that this identity remains valid during congestion periods and despite the occurrence of anomalies.

Summary: The median-sketch based analysis of the MAWI 7-year long dataset demonstrates that the LRD paradigm has been and remains 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 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).

5.2 Tendency to Poisson or Gaussianity ?

Numerous studies claim that the on-going increase in link bandwidth causes a return of traffic toward Poisson distribution for packet interarrival process and, hence mechanically, toward Gaussian aggregated times series (cf. e.g., [3, 32]). Such a statement per se hides a number of involved issues, some being addressed here.

First, Gaussianity of marginal distributions (MD) is in itself of only limited interest (see, for instance, [21]). One should rather consider Gaussianity for joint n-point distributions, as this is what actually matters for traffic modeling, performance assessment and network engineering. However, this is difficult to analyze in practice. Instead, we here handle this indirectly, by analyzing how MD evolve towards Gaussianity under an increase of the aggregation level. Second, as mentioned earlier, median-sketch LDs undergo little changes in shape despite significant variations in bandwidth or in the numbers of packets, bytes, etc. Notably, the onset (the knee) of the LRD scale range and the LRD parameters remain remarkably constant. This is in contradiction with any return to a Poisson behavior, which would indeed imply the disappearance of LRD, or, at least, a significant increase of the LRD onset scale $2^{j_*}\Delta_0$. The previous section showed that this onset scale remains within the range $[0.5, 1.5]$s over the entire period. Third, traffic collected on a large link is likely to appear more Gaussian than traffic from smaller link, due to a pure aggregation effect (central limit theorem), often referred to as *statistical multiplexing* or *multiplexing gain* [4]. The main point is that evolution toward Gaussianity with respect to an increase of the aggregation level matters more than closeness to Gaussianity at a given level. Therefore, there is a necessity to introduce some compensation involving the bandwith and the throughput to renormalize this closeness to Gaussianity accordingly.

Our contribution to this issue makes use of α_j, as indicators for closeness to Gaussianity. Because anomalies affect the evolution toward Gaussianity (cf. Sec. 5.3 and [8]), we use the robust median-sketch estimates of α_j. Obviously, periods with major anomalies (e.g., Sasser) modify globally the traffic statistics, hence they are not considered when drawing conclusions about the long term evolution of Gaussianity.

Fig. 10 (top row) shows α_j for different aggregation levels $2^j\Delta_0$, as a function of time. There are various periods of specific interest. Top left plot (Jp2US), reveals a regular increase of α_j with time (from 2004/03 to the

end of 2006). Comparisons with Fig. 1 indicate that this corresponds to a slow but regular increase of throughput (especially in bytes). This confirms that departure from (or closeness to) Gaussianity for aggregated traffic MD are impacted by traffic volumes, whatever the granularity. An analog conclusion can be drawn when observing the decrease in α_j on top right plot (US2Jp, from 2003/03 to the end of 2004) related to a byte congestion period and a concurrent packet decrease. This is in agreement with findings reported in [3]. However, comparing 2005 at **B** to 2007 at **F** in both directions shows that the question is intricate. Packet, byte or flow rates are higher at **F** yet producing smaller α_j (hence distributions that are farther to Gaussian) for the same j. A deeper analysis shows that the bandwidth occupancy ratio for **F** in 2007, is close to 50%, while it is close to 90% for **B** in 2005. Also, the bandwidth occupancy ratio for **B** in 2002 (Jp2US) is close to 50% yielding α_j of comparable order of magnitude to those at **F** in 2007 (Jp2US). This demonstrates that not only traffic volumes but also bandwidth occupancy ratio controls closeness to Gaussianity.

To further proceed, we renormalize α_j as $\alpha'_j = \alpha_j/\alpha_J$. This amounts to assuming, for an arbitrarily chosen aggregation level $2^J\Delta_0$, a level of closeness to Gaussianity accounting for global traffic volume effects. In Fig. 10 (bottom row), one sees that these $\alpha'(j)$ remain, strinkingly far more constant along time (notably over the two specific periods discussed above), and this for all j. This is in no way a trivial effect that could be induced by the chosen normalization procedure. It means that the evolution toward Gaussianity as a function of the different aggregation levels is kept constant along time, even at the upgrade from **B** to **F**. This suggests that changes in flow, byte or packet rates and/or bandwidth occupancy affect closeness to Gaussianity at a given aggregation level but do not impact the (speed of) evolution toward Gaussianity. In other words, an increase in throughput or bandwidth causes aggregated traffic to have MD being closer to Gaussian, but with unchanged speeds at which the joint distributions evolve to Gaussian. Such analyses and conclusions shed a new light of the issue of traffic Gaussianity: from a network engineering point of view, the evolution to Gaussianity is in itself far more important than absolute Gaussianity at a given level.

5.3 Anomalies

Finally, many features of the traffic are caused by the numerous anomalies found in the traces. The methodology proposed for robust analysis can be adapted to anomaly detection: departure from the median-sketch average behavior are considered anomalous. This automated anomaly detection procedure has been proposed and validated in [8]. In a nutshell, one splits traffic by sketching it in M outputs, using N different hash functions based on one of the packets attributes (usually source or destination IP). A sketch output is declared abnormal when the $\{\alpha_j, \beta_j\}$, obtained from MD computed jointly from several aggregation levels (multiresolution), differ significantly from the median taken over the sketch outputs. Anomalies (together with their IP attributes) are identified by tracking flows that are consistently hashed in abnormal sketch outputs — hence the use of several different hash functions. More details on this anomaly detection tool are shown in [8].

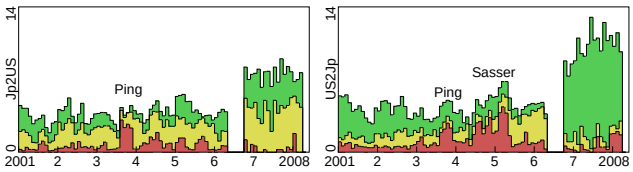

Figure 11: **Number of detected anomalies** using the Sketch+Multiresolution method. Hash on IPsrc. From top to bottom: "Suspected" (green): WWW, P2P, GRE, DNS. Mostly attacks (yellow): various mechanisms. "Sure attacks" (red): Ping/SYN floods, spoofed,... Left: Jp2US; Right: US2Jp.

Global features of anomalies: Applied to each day of the 7 year long dataset, at **B** and **F**, the detection procedure showed that around half a dozen (often many more) of suspicious significant events are usually identified in each trace as shown in Fig. 11. Inspecting them shows that they belong to numerous and different kinds of anomalies with variety of lasting time. A systematic detection and classification over the entire dataset is beyond the goals of the present contributions. Instead, we concentrate on two issues: recurrent anomalies and most prominent ones.

It is often quite difficult to automatically determine whether detected anomalies correspond to a real attack, a defective host, or is legitimate but unusual traffic that stands out in the trace. A decision requires a careful inspection of the alert, by examination of packet attributes, IPs, flows, host behaviors, etc. A preliminary yet automated procedure, based on simple attribute recognition, sorts alerts into three main categories (cf. Fig. 11): "sure" attacks (e.g., SYN floods, PING floods, packets with spoofing, etc.), unusual traffic patterns that are mostly classified as attacks if inspected closely (and manually), and "suspected" anomalies which are traffic with usual protocols and could hence be legitimate, but show unusual volume or statistical properties.

Most prominent anomalies: Ping flood, Sasser worm: Some anomalies are extremely large and contribute to a significant proportion of the traffic (more than 80% of the link capacity) as shown in Fig. 12. A first significant and long-lasting detected anomaly consists of a ping flooding (2003/08-12). Ping floods are quite common and can be regularly found during the

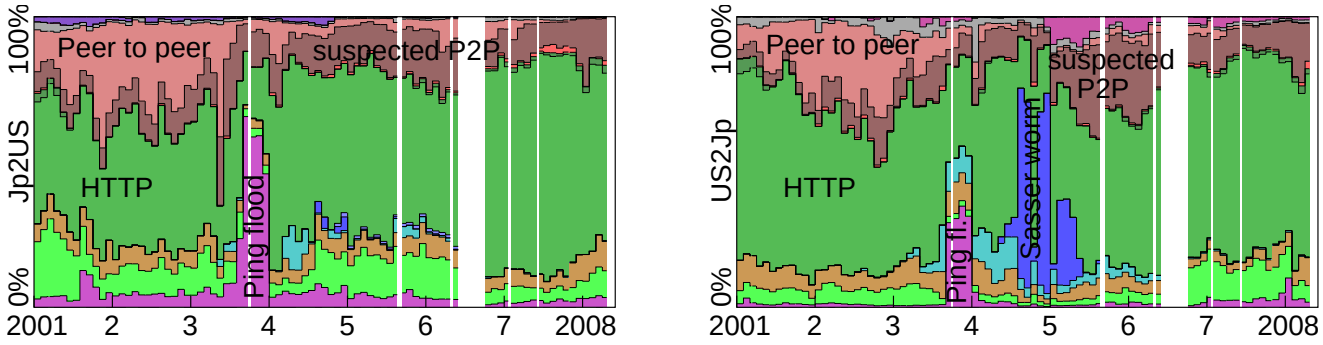

Figure 12: **Applications/anomalies breakdown vs. years.** Bottom to top : Ping, DNS, common services, MS vulnerarities, Sasser, HTTP, broadcast, suspected P2P, identified P2P, other TCP/UDP, INLSP (left) / GRE (right).

seven years. This case is special in that it lasts several months and with a very high volume: more than half of the packets on the link in Jp2US, and around 25% in the other direction, are ICMP packets. Another large detected anomaly has already been mentioned: the famous Sasser worm activity, between 2004/05 and 2005/05 mainly. From the traffic breakdown in Fig. 12, successive outbursts are observed (2004/08, 2004/12 and 2005/03): Sasser was on the verge of disappearance twice, yet came back (probably variants of the worm). This Sasser activity accounts for more than 50% of the US2Jp traffic, while barely noticeable in the opposite direction (which is likely due to a better defense against worms).

Anomalies in WWW exchanges: We also found that misbehaved HTTP carried on the unusual amount of large packets sent to Japanese servers in 2004/03-05. It could be just upload but this often involves more than 10^6 packets sent to a single server in each 15-min trace. This hence seems to be an attack, even if certainty is not at hand (without packet payload, no back-engineering of anomalies is possible). This is a typical example of "suspected" anomalies in Fig. 11. Those events are anomalous in every respect, yet one can not completely rule out the possibility of its being legitimate traffic.

Recurrent anomalies: Besides those major events, there existed many other anomalies. SYN scans and floods towards HTTP and other services are especially very common. Anomalies targeting any and all protocols and applications are regularly found, usually closely related to the popularity of the protocol itself. In particular, we found appearance of typical types of anomalies depending on the observed period; NNTP for the earlier days, SSH since 2004, MS security holes related from 2003/08, and so on.

Summary: The painting of a complete description of all the anomalies in traffic and of their evolution through 7 years remains a difficult exercise in a restricted space. The salient point is that *normal* traffic seems to be never observed, numerous significant anomalies are consistently found in each 15-minute long daily traces; major anomalies that consume more than half of the throughput during several consecutive months are also found; there exist anomalies that continue over the span of months, even years (some from or to the same hosts). Hence, our findings underline the need for estimation procedures that are robust against anomalies, when performing long term evolution analyses.

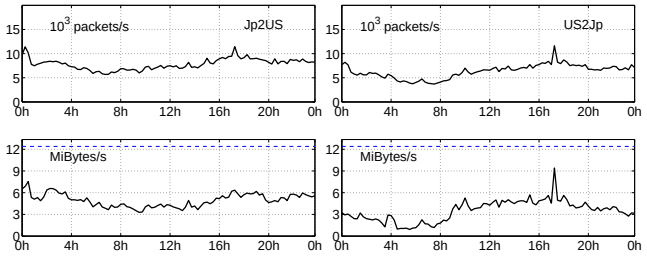

Figure 13: **One day: Throughput vs. hours.**

6. RESULTS ON A 24-HOUR LONG TRACE

We now analyze a 24-hour long trace collected on 2008/03/19, within the framework of the *A Day in the Life of Internet* project [20]. This enables us to 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 (not shown here) are satisfactorily modeled with Gamma laws. LDs systematically present the usual knee-type shape, with $\Delta_0 2^{j_*} \simeq 0.5$s. Fig. 13 reveals a smooth modulation of the traffic volume with respect to the hours of the day, which simply amounts to a vertical shift in LDs. This discrepancy is fixed by normalizing the traces w.r.t. their volume. Yet, despite normalization, Fig. 14(a) reveals a significant variability of the LDs around the knee-type shape, affecting a large range of scales, from

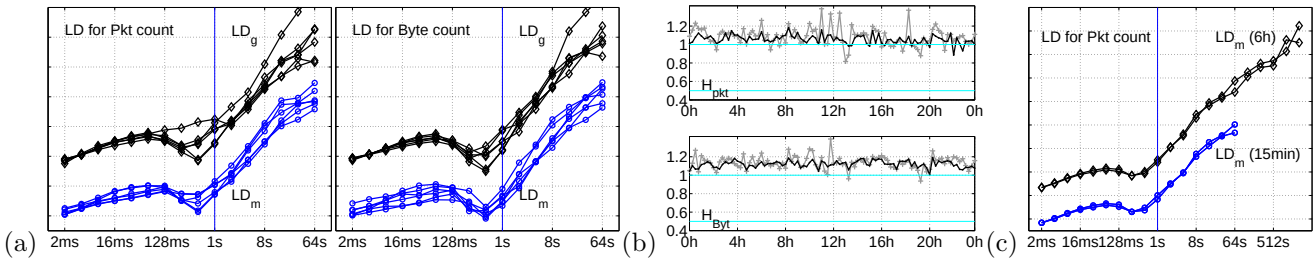

Figure 14: **One day: Statistics** (Jp2US). (a) LDs for 15min long traces (Left: Pkt; Right: Byte): Global (black $\diamond$) and median-sketch (blue $\bullet$). (b) Estimates for H vs. time: Global (gray), median-sketch (black). (c) LDs for 6h-long traces (Jp2US): Average LD over 15-min long traces (blue $\bullet$) and LD computed fully on 6h trace (black $\diamond$).

a few ms to the minute. Variability is further confirmed in Fig. 14(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 were 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. Also, again, byte and packet count median-sketch based estimates for H are found to be closely tied together (scatter plots not shown). Hence, as have been expected, the median-sketch methodology proves robust against intraday variability as well and permits to disentangle smooth evolution along the day from localized events (such as anomalies). Moreover, the median-sketch procedure reveals a strong and persistent LRD (with constant Hurst parameter) irrespectively of the time of the day. This can be interpreted as follows. Japan and the USA belonging to very different time zones, traffic is always important and there always active sessions (and Internauts), and hence the network mechanisms producing LRD remain constantly active.

Stationarity time scale: There is an ongoing debate questioning the existence of LRD w.r.t. non stationary effects, to which the analysis of a 24-hour long trace can contribute. Inspired by the methodology developed in [28], the 24-hour trace is split into adjacent and non-overlaping sub-traces over which LDs are computed independently. Fig. 14(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) is clearly and consistently expanded at coarser scales (1 min to 1h), confirming its existence and hence the meaningfulness of the estimates reported in Sec. 5. A careful analysis indicates a slight decrease of the estimated H when measured at the coarsest scales (1 min to 1h), compared to those obtained from the range 1s to 1min. This is consistent with previous analyses indicating that, LRD being a coarse scale asymptotic property, it may be difficult to measure H precisely when traces are not long enough [13]. This may explain that, for **F**, H is slightly over estimated (≥ 1): the bandwidth increase may imply that LRD should be actually measured at coarser scales, but only partially available for 15-min long traces. Still, our main conclusions (LRD remains constant and strong) cannot be questioned.

Summary: The analysis of this 24-hour trace enables us to further illustrate the relevance and potentials of the proposed median-sketch estimation procedure: it is robust against intraday variabilities. Also, this analysis reveals a strong and constant LRD irrespectively of the time of the day. Hence, the long term evolution study reported in Sec. 5.1 is not significantly dependent on the specific data collection time. The trace actually collected lasted 72 hours. The other 48 hours yield equivalent conclusions. Moreover, the MAWI datasets contain several other long traces. For comparison, that of 2005/09/22 has also been studied (not shown) and yields similar conclusions, hence further validating the stability of the statistical characterization along the years.

7. CONCLUSION

A unique day-by-day longitudinal analysis of a 7 year (and one day) long dataset has been conducted. It shows that the estimations of the parameters entering the traffic statistical characterization exhibit a huge daily variability, likely due to traffic condition variations (congestions, restrictions,...) as well as a wide variety of low-volume anomalies constantly but randomly occurring. Such wild fluctuations significantly impair the possibility of drawing long term evolution conclusions. Therefore, our first major contribution is methodological: to disentangle long time evolution from day-by-day incidental variabilities, we have proposed the recourse to an estimation procedure based on sketches and median average, and shown that it brings robustness against congestions and low-volume anomaly impacts as well as against intra-day variability.

The analysis of each day traffic yields our second major contribution: the parameters describing the statistical characterization of Internet traffic remain surprising stable along the entire period. LRD (for both packet

and byte) remains remarkably strong, persistent and stable. The LRD onset scale of time remains stable (0.5s to 1s). Our robust analysis showed (for one day traces) that LRD persists over hours. Also, the LRD parameter H for bytes and packets are closely related (almost identical). This indicates that 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 are constantly well modeled with Gamma distributions. This enabled us to argue that it is not how a MD at a given aggregation level is close to Gaussian that matters but rather how fast it evolves toward Gaussianity under aggregation. We found that this speed of evolution also remains stable along the years. These persistence and stability of these two major statistical properties lead us to conclude that they are intrinsic and unavoidable features of aggregated Internet traffic and also that there is no evidence for a return to Poisson inter-arrival process, even when the capacity or the loads of the links are significantly increased. This might be seen as a pessimistic conclusion w.r.t. traffic and network engineering: the failure of the Poisson model still holds. However, the remarkably stable traffic characterization can also be exploited. 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 remains also relatively stable, despite the intuitive and heuristic claims often made, forecasting dramatic changes in Internet traffic. The application usage has slowly shifted from pure web traffic to P2P applications along the whole period, together with changes in 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. A further study will be to extend these analyses to multiple measurement points, to obtain more global view of traffic statistics.

8. REFERENCES

[1] W. Acosta and S. Chandra. Trace driven analysis of the long-term evolution of gnutella peer-to-peer traffic. In *PAM'07*, pages 42–51, 2007.

[2] M. Allman, V. Paxson, and J. Terrell. A brief history of scanning. In *IMC'07*, pages 77–82, 2007.

[3] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. On the nonstationarity of internet traffic. In *ACM SIGMETRICS/Performance'01*, pages 102–112, 2001.

[4] 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*, 2002.

[5] 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*, pages 1–14, 2007.

[6] K. Cho, K. Mitsuya, and A. Kato. Traffic data repository at the WIDE project. In *USENIX 2000 Annual Technical Conference: FREENIX Track*, pages 263–270, June 2000.

[7] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Trans. Networking*, 5(6):835–846, 1997.

[8] 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*, pages 145–152, 2007.

[9] A. Erramilli, O. Narayan, and W. Willinger. Experimental queuing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Networking*, 4(2):209–223, 1996.

[10] A. Feldmann, A. C. Gilbert, and W. Willinger. Data networks as cascades: Explaining the multifractal nature of internet wan traffic. In *SIGCOMM'98*, pages 42–55, 1998.

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

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

[13] N. Hohn, D. Veitch, and P. Abry. Cluster processes, a natural language for network traffic. *IEEE Trans. Signal Proc., Spec. Issue on Sig.Proc. in Networking*, 8(51):2229–2244, Oct. 2003.

[14] N. Hohn, D. Veitch, and P. Abry. Multifractality in TCP/IP traffic: the case against. *Comp. Networks*, 48:293–313, 2005.

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

[16] k claffy, G. C. Polyzos, and H.-W. Braun. Tracking long-term growth of the nsfnet. *Commun. ACM*, 37(8):34–45, 1994.

[17] T. Karagiannis, A. Broido, N. Brownlee, k claffy, and M. Faloutsos. Is p2p dying or just hiding? In *IEEE GLOBECOM'04*, 2004.

[18] T. Karagiannis, M. Molle, and M. Faloutsos. Long-range dependence - ten years of internet traffic modeling. *IEEE Internet Computing*, 8(5):57–64, 2004.

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

[20] kc claffy. A day in the life of the internet: Proposed community-wide experiment. *ACM Comp. Com. Rev.*, 36(5):39–40, 2006.

[21] J. Kilpi and I. Norros. Testing the gaussian approximation of aggregate traffic. In *IMW'02*, pages 49–61, 2002.

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

[23] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. *IEEE/ACM Trans. Networking*, 2(1):1–15, 1994.

[24] S. Muthukrishnan. Data streams: Algorithms and applications. In *ACM SIAM SODA*, page 413, Jan. 2003.

[25] V. Paxson and S. Floyd. Wide area traffic: The failure of poisson modeling. *IEEE/ACM Trans. Networking*, 4(3):209–223, 1995.

[26] 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.*, 4(1):56–70, Jan. 2007.

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

[28] D. Veitch and P. Abry. A statistical test for the time constancy of scaling exponents. *IEEE Trans. Signal Proc.*, 49(10):2325–2334, Oct. 2001.

[29] A. Veres, Z. Kenesi, S. Molnar, and G. Vattay. On the propagation of long-range dependency in the internet. In *SIGCOMM'00*, pages 243–254, 2000.

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

[31] 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. Networking*, 5(1):71–86, 1997.

[32] Z.-L. Zhang, V. Ribeiro, S. Moon, and C. Diot. Small-time scaling behaviors of internet backbone traffic: an empirical study. In *IEEE INFOCOM'03*, pages 1826–1836, 2003.