Internet Traffic: Analysis, Modeling with real-world aspects

Pierre BORGNAT

CNRS – ENS Lyon, Laboratoire de Physique (UMR 5672)

TERA-NET – 07/2010

- Internet traffic metrology: some basics
- Analysis: Scale Invariance, LRD, Robust Estimation
- Modeling: LRD / Heavy-Tails
- Anomaly Detection; Host classification

- Acknowledgements
  - P Abry, G Dewaele, P Flandrin, A Scherrer, P Gonçalves, P Loiseau, P Primet (Lyon, ENSL, CNRS & INRIA)
  - Ph Owezarksi, N Larrieu (LAAS-CNRS) Metrosec (ACI Sécurité & Informatique), ANR OSCAR JL Guillaume, M Latapy, C Magnien (LIP6)
  - K Fukuda, R Fontugne, Y Himura (NII), K Cho (IIJ) (Tokyo)
  - D Veitch, N Hohn (Melbourne Univ.)
  - O Michel (GIPSA-lab, INPGrenoble)

- Internet traffic metrology: some basics
- Analysis: Scale Invariance, LRD, Robust Estimation
- Modeling: LRD / Heavy-Tails
- Anomaly Detection; Host classification

- **Acknowledgements**
  - P Abry, G Dewaele, P Flandrin, A Scherrer,
    P Gonçalves, P Loiseau, P Primet
    (Lyon, ENSL, CNRS & INRIA)
  - Ph Owezarksi, N Larrieu (LAAS-CNRS)
    Metrosec (ACI Sécurité & Informatique), ANR OSCAR
    JL Guillaume, M Latapy, C Magnien (LIP6)
  - K Fukuda, R Fontugne, Y Himura (NII), K Cho (IIJ) (Tokyo)
  - D Veitch, N Hohn (Melbourne Univ.)
  - O Michel (GIPSA-lab, INPGrenoble)

Traffic & Network Measurement

Overview of networks properties

- Heterogeneity
  (of information, devices, topologies, geography,...)
- Evolve with time (new services, increased usage,...)
- Complexity
  - individual elements $\nRightarrow$ behaviour of the whole
  - interplay: architecture / protocols / usages
- Crucial choice: level of description
  - Information flows? $\rightarrow$ Signals
  - Network's level? $\rightarrow$ Graphs, or Multivariate Signals

$\rightarrow$ Need for a statistical approach

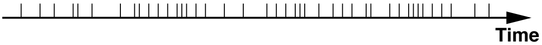
Traffic & Network Measurement: What for?

- Analysis of networks:
  (protocols, routeurs, provisioning,...)
- Modeling of traffic and of its properties
- Classification or recognition of traffic
  (with new needs: Peer to Peer, real-time, wireless,...)
- Définition of service agreements
  (Pricing, QoS, Committed QoS...)
- Security of Networks; Intrusion Detection Systems; Anomaly Detection
  (DDoS, scans, computer virus, worms, outages...)

[ACI METROPOLIS 2001, AS Métrologie des réseaux de l'Internet 2003, ACI METROSEC 2007,...]

Passive Measurements of traffic

- On networks: *Internet Protocol* → Packets+information
- Monitoring facilities: add a time-stamp to data (dynamics)
  - **link level**, monitor packets: intercept (port-mirroring, splitter,...); capture (tcpdump, DAG, GNET,...); filter (...)



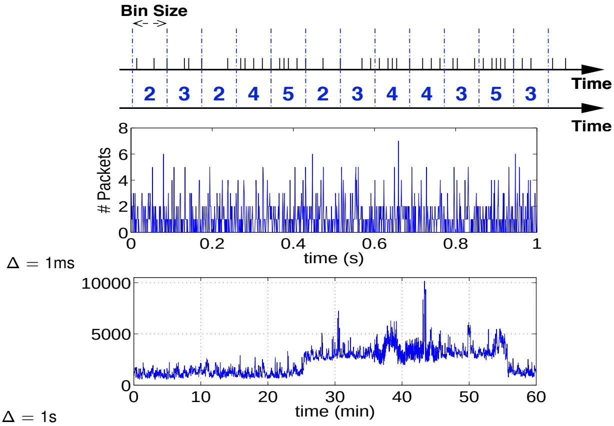
IP protocol	Source Address	Destination Address	Source Port	Destination Port

→ Point processes (marked)

  - **node level** (routeur) → multivariate data
    Device: routeur ! Netflow (CISCO), flow-tools (Juniper)
  - **network level** → multivariate data, graph
    Synchronising several link or node monitoring?

Passive Measurements of traffic

- $\rightarrow$ Huge stream of data.
- Aggregated cout process = # of packets during Δ



- **Problematic:** understand the features of traffic

Short Biblio. on Longitudinal Traffic Analysis

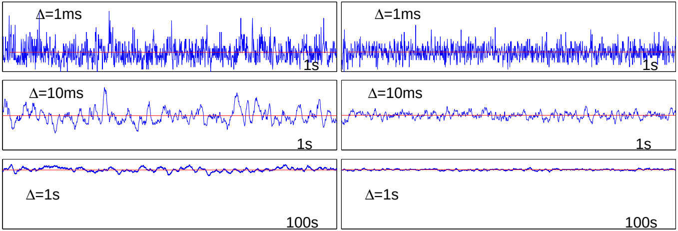
- Many works during the past 15 years.
- Some Focus on newest application at the time:
  - FTP, Mail in early 90's [kc claffy *et al.*, Comm. ACM 94]
  - Web, mid-90's [Crovella & Bestravos, ToN 95]
  - P2P, early 2000's [Karagiannis *et al.*, Globecom'04]
  - Video Streams, late 2000's [Cha *et al.*, IMC'07]
  - ...
  - Anomalies: History of Scanning [Allman *et al.*, IMC'07]
  - Wireless, Mobile,...
- Some focus on non-classical statistical properties:
  - 'Failure of Poisson modeling' / Self-similarity / Scaling / LRD [Leland *et al.*, 94] [Paxson & Floyd, 95], [Willinger *et al.*, 97], [Veitch & Abry, 01], [Cao *et al.*, 02], [Karagiannis *et al.*, 04], [Hohn *et al.*, 05], [Robeiro *et al.*, 05]

Internet traffic: not a simple renewal process

The Failure of Poisson Modeling. Paxson & Floyd 1994

- If Internet ≃ phone
  - Packets would follow a Poisson process
  - Short-range correlations only
  - Aggregated traffic: Gaussian law (per Central Limit Thm)

- The thruth: much more variabilities and burstiness



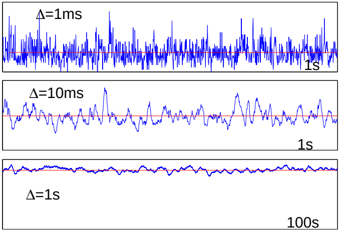
IP Traffic | Poisson Traffic

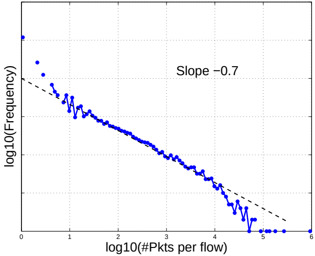
Internet traffic: not a simple renewal process

The Failure of Poisson Modeling. Paxson & Floyd 1994

- If Internet $\simeq$ phone
  - Packets would follow a Poisson process
  - Short-range correlations only
  - Aggregated traffic: Gaussian law (per Central Limit Thm)

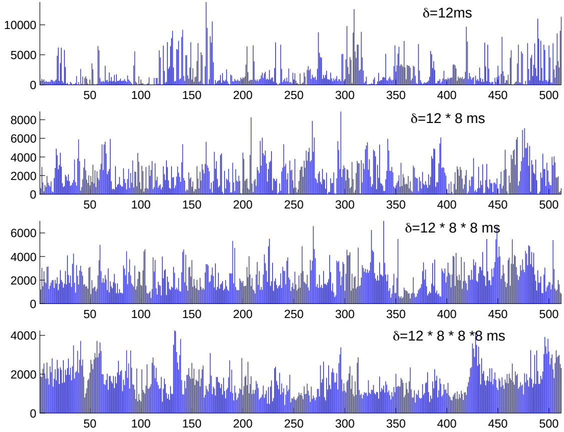
- The thruth: much more variabilities and burstiness





- # packets per $\Delta \neq$ Poisson distrib.
- waiting times $\neq$ Exponential distribution
- correlations $\neq$ short-range only

Traffic series: aggregation at several time-scales

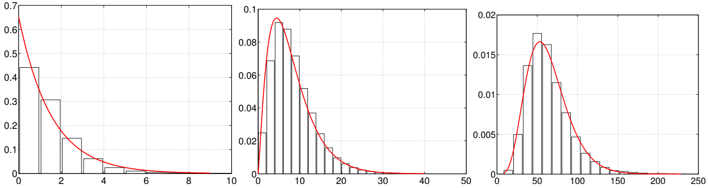


- Same kinds of fluctuations seens at all the different levels

Marginal probability distributions

Traffic trace LBL-TCP-3 (1994)

- Empirical histograms of the # of packets per Δ
- Estimation: count the number of occurrences



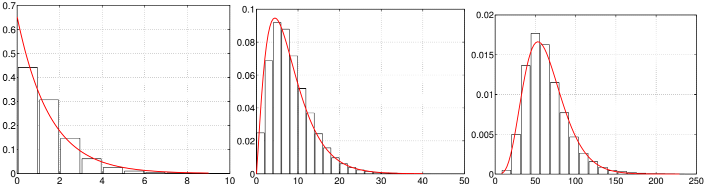
$\Delta = 4$ms $\quad \Delta = 32$ms $\quad \Delta = 256$ms

- Exp. $p(x) = e^{-x/\beta}/\beta$ $\qquad$ Gaussian: $p(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$
- Fit/Model: Gamma $\Gamma_{\alpha,\beta}(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\frac{x}{\beta}\right).$

Marginal probability distributions

Traffic trace LBL-TCP-3 (1994)

- Empirical histograms of the # of packets per Δ
- Estimation: count the number of occurrences



$\Delta = 4$ms $\Delta = 32$ms $\Delta = 256$ms

- Exp. $p(x) = e^{-x/\beta}/\beta$ Gaussian: $p(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$
- Fit/Model: Gamma $\Gamma_{\alpha,\beta}(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\frac{x}{\beta}\right).$

Long-Range Dependence (or Long Memory)

The Self-Similar Nature of Ethernet Traffic. Leland, Taqqu, Willinger & Wilson 1993

Property of Long-Range Dependence (LRD)

Covariance tends to a non-summable power-law (at large lags)

$\Rightarrow$ Spectrum $F_X(\nu) \sim c|\nu|^{-\gamma}, \ |\nu| \to 0, \quad$ avec $0 < \gamma < 1$.

- Spectrum – (Wiener-Khintchine) $\rightarrow$ Correlation

$$F_X(\nu) = \left| \frac{1}{T} \int_0^T e^{-i2\pi\nu t} X(t) \mathrm{d}t \right|^2 = \int C_X(\tau) e^{-i2\pi\nu\tau} \mathrm{d}\tau$$

Self-similarity: statistical invariance under dilatation

A random process $\{X(t), t \geq 0\}$ is **self-similar** with index H ("H-ss") if **for all** dilation factor $\lambda > 0$,

$$X(\lambda t) \stackrel{d}{=} \lambda^H X(t), \quad t > 0.$$

- H-ss for $H > 0.5 \Rightarrow$ LRD.

Time-Scale Representation

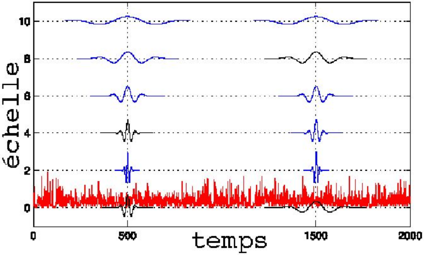
Definition : Wavelet transform

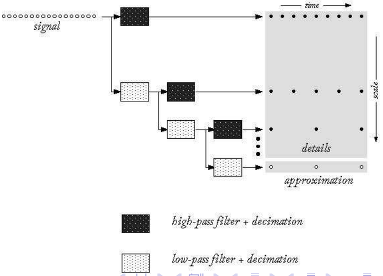
Shifted (time) and dilated (scale) versions of ψ_0 :

$$\psi_{j,k}(t) = 2^{-j/2}\psi_0(2^{-j}t - k).$$

Wavelet coefficients:

$$d_{X_\Delta}(j,k) = \langle \psi_{j,k}, X_\Delta \rangle.$$





Efficient Algo. [Mallat 1989]

Self-Similarity and Wavelets

- Signature of self-similarity
  $\mathbb{E}(d(j,k))^2 = 2^{j(2H+1)}\mathbb{E}(d(0,k))^2$.
- Decorrelation of wavelet coefficients (due to N, number of null moments for the wavelet). If $N > H + 1/2$:
  $\mathbb{E}(d(j,k)d(j',k')) \simeq |2^j k - 2^{j'} k'|^{2H-2N}$ si $|2^j k - 2^{j'} k'| \to \infty$.

Wavelet Spectrum: $S_2(j) = \dfrac{1}{n_j}\sum_{k=1}^{n_j} |d_{X_\Delta}(j,k)|^2$

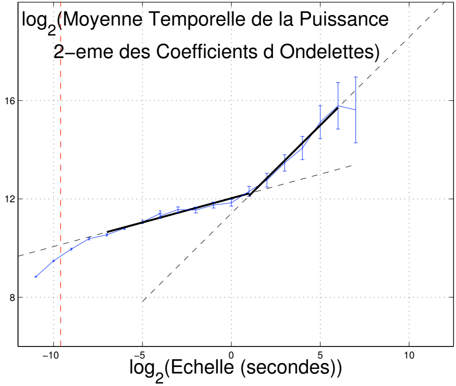
$$\mathbb{E}\left\{S_2(j)\right\} = \int F(\nu)2^j|\Psi_0(2^j\nu)|^2 d\nu \to \hat{F}\left(\nu = \frac{\nu_0}{2^j}\right) \simeq S_2(j).$$

- H-ss $\Longrightarrow \mathbb{E}\left\{S_2(j)\right\} \sim c\, 2^{j(2H+1)}$.
- LRD $\Longrightarrow \mathbb{E}\left\{S_2(j)\right\} \sim c\, 2^{j\gamma}$ if $2^j \to +\infty$.

[Abry & Veitch '98; Abry, Flandrin, Veitch & Taqqu '00]

Log-scale Diagrams (LD)

- Test of this linear behaviour: $\log_2 S_2(j)$ vs. $\log_2 2^j = j$



Traffic from Auckland-IV (2001)

- Current knowledge: At least two ranges of scales:
  - Scale invariance $H \sim 0,8$ for the large scales
  - Small scales: no clear multi-scaling

What about a Robust Longitudinal Analysis?

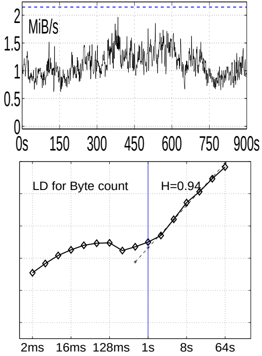
Is this a robust feature of traffic over the years?

- **Topics in Statistical analysis of traffic**
- Diversity of expected traffic: http, P2P, mail, DNS,...
- Variety of conditions: used bandwidth, congestion,...
- Frequent anomalies: scans, viruses&worms, DDoS,...
- ...

- Intuition: **One trace is not enough!**
  (for longitudinal, empirical data analysis)
- MAWI dataset: more than 7 years of daily traces
- WIDE network (AS2500); trans-pacific backbone
- 2TB of (anonymized) packet traces (still growing...)
- Sample point **B**: 18Mbps CAR (on a 100Mbps link)
- Then **F**: full 100Mpbs, then 150Mpbs CAR (on 1Gbps)
- `http://mawi.wide.ad.jp/`

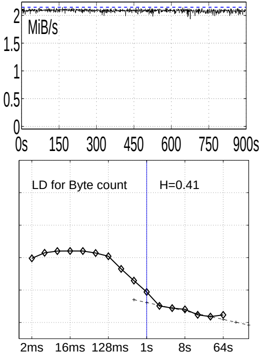
This is real network!...

day OK



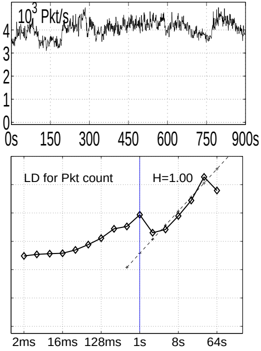
B-US2Jp, 2005/07/11

w/ congestion



B-US2Jp, 2003/06/03

w/ anomalies



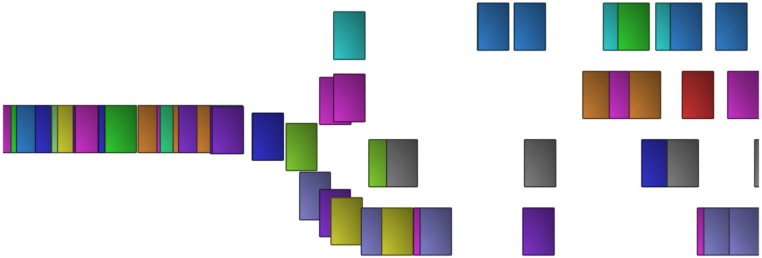
B-Jp2US, 2004/09/21
(network scans, spoofed flooding, attacks on RealServer,...)

Question of methodology

How can we be certain of the validity of what is seen?

- Text-book solution: **averaging**... over what? *along time?*
- **However**: Anomalies, failures, non-stationarities,...

- Proposition: **use Sketches**
  = *M* sub-traces taken by random projections (of flows)
- Averages over outputs → reduce variance of estimation.
- Average using **median** = robust estimator



Sketched Traffic

Sketches = ensemble of outputs of random hash table

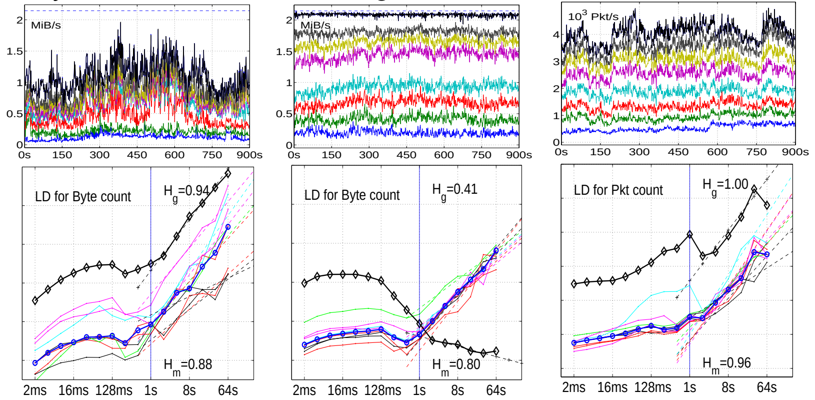
[Muthukrishnan'03, Krishnamurty'03,...] [Abry+ SAINT'07, Dewaele+ Sigcomm LSAD'07]

- Random Hash Functions : h_n
  - $y = h(x)$,
  - $M-$outputs: $y \in [1, \ldots, M]$,
  - $k-$universal Hash functions.

- Hash the Traffic :
  - Packet: $i-$th packet has: $t_i, PTscr_i, PTdst_i, IPsrc_i, IPdst_i$
  - Choose one specific key, e.g., Destination Address
  - Hash according to this key: $m_i = h(IPdst_i) \in [1, \ldots, M]$,
  - All packets with same m_i = one sub-trace, sampled by random projection.
  - **Aggregate** traffic $\{t_i, m_i\}_{i \in I}$ into M series $X^m_\Delta(t)$, bins of Δs.

Robust Estimation of LRD with Sketches

day OK | w/ congestion | w/ anomalies



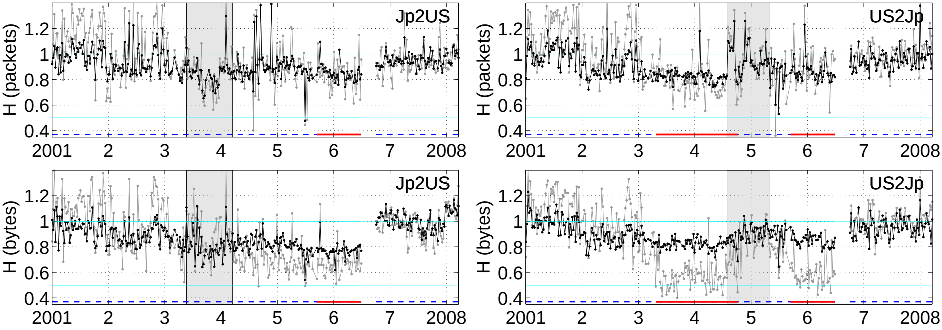
B-US2Jp, 2005/07/11 | **B**-US2Jp, 2003/06/03 | **B**-Jp2US, 2004/09/21

- Sketches = **random flow sampling**
  $\rightarrow$ filters out anomalies, congestion, accidents,...
- **Median** on Sketches = $H \simeq 0.9$ + LDs have similar looks

Longitunal study: Estimation of LRD, H parameter

MAWI dataset (backbone) [Borgnat et al. INFOCOM 2009]

H vs Year 2001-2008. From Japan (left) and To Japan (right)

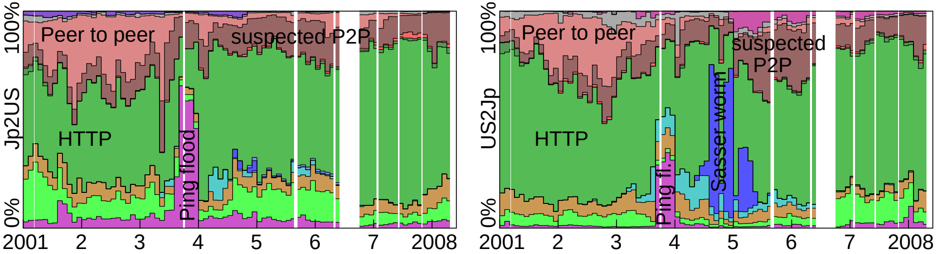


- Congestion = global traffic goes to $H \simeq 0.5$
- However the flows still see relevant LRD: **median** on sketch's outputs $\sim$ usual traffic, $H \simeq 0.8$ to 0.9

Longitunal study: LRD is a robust feature of traffic!

[Borgnat et al. INFOCOM 2009]

- Analysis over 7 years of data
- Diverse conditions of traffic (congestion or not,...)
- Diverse composition of traffic (with large proportion of "hidden" P2P, and of anomalies!)



Bottom to top : Ping, DNS, common services, MS vulnerabilities, Sasser, HTTP, broadcast, suspected P2P, identified P2P, other TCP/UDP, INLSP (left) / GRE (right) – (Left: Jp2US; Right: US2Jp).

Traffic Modelling

- Choice of details: aggregated series, packet processes, complete trace?
- Self-similarity paradigm $\neq$ one model (e.g., fBm)
- Main statistical properties to satisfy:
  - Long Range Dependence
  - Non Poisson Statistics
  - Heavy-Tailed Probability Distributions for # of packets/flow; Flow durations; File sizes on WWW,...

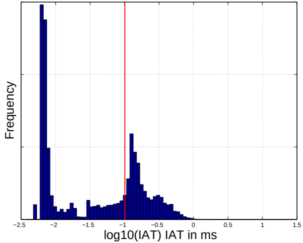
Def.:there is $\alpha > 0$ s.t. $P(X > x) \sim cx^{-\alpha}$ when $x \to \infty$.

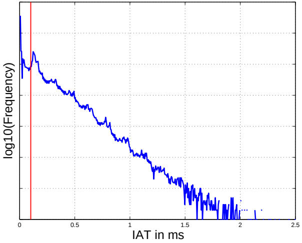
Heavy-Tailed Probability Distributions in the WWW. Crovella, Taqqu & Bestavros 1998

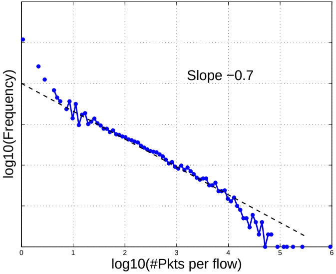
On the relationship between file sizes, transport protocols, and self-similar network traffic. Park, Kim & Crovella 1996

Heavy-Tails in Traffic

Inter-Arrival Times





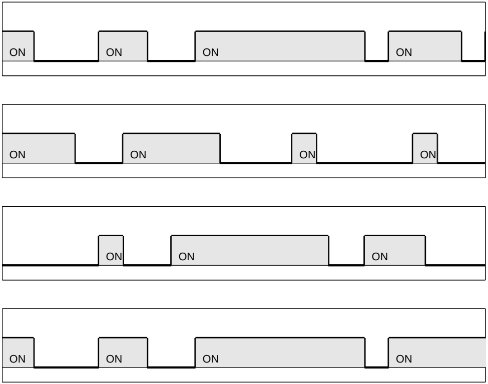


packets/flows $\longrightarrow$ power law

From Heavy-Tails to LRD

Proof of a Fundamental Result in Self-Similar Traffic Modeling. Taqqu, Willinger & Sherman 1997

- Superposition of activity sessions that are **independent**

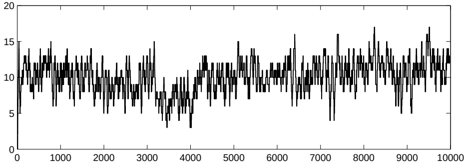


- PDF of the durations τ :
  - of activity (ON) : heavy-tailed law with exponent α
  - of inactivity (OFF) : heavy-tailed law with exponent β, or law without heavy-tail

From Heavy-Tails to LRD

Proof of a Fundamental Result in Self-Similar Traffic Modeling. Taqqu, Willinger & Sherman 1997

- $S_N(t) = \sum_{i=1}^{N} X_i(t)$



- Limiting Cumulative Process: there is $c >$ s.t.

$$Y_N(t) = \int_0^{Tt} S_N(s)ds \overset{d}{\sim} \frac{\mathbb{E}(\tau_{on})}{\mathbb{E}(\tau_{on}) + \mathbb{E}(\tau_{off})} NtT + c\sqrt{N}T^H B_H(t)$$

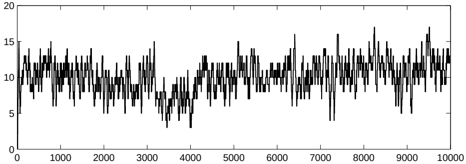
if $N \to \infty$, $T \to \infty$ and $H = \frac{3-\alpha^*}{2}$ (for $\alpha^* = \min(\alpha, \beta, 2)$)

- Consequence: LRD if $\alpha \in [1, 2]$ (infinite variance)

From Heavy-Tails to LRD

Proof of a Fundamental Result in Self-Similar Traffic Modeling. Taqqu, Willinger & Sherman 1997

- $S_N(t) = \sum_{i=1}^{N} X_i(t)$



- Limiting Cumulative Process: there is $c >$ s.t.

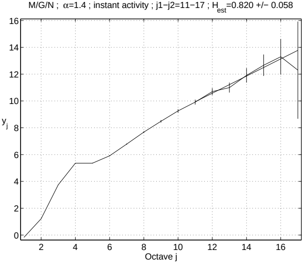
$$Y_N(t) = \int_0^{Tt} S_N(s)ds \overset{d}{\sim} \frac{\mathbb{E}(\tau_{on})}{\mathbb{E}(\tau_{on}) + \mathbb{E}(\tau_{off})} NtT + c\sqrt{N} T^H B_H(t)$$

if $N \to \infty$, $T \to \infty$ and $H = \dfrac{3-\alpha^*}{2}$ (for $\alpha^* = \min(\alpha, \beta, 2)$)

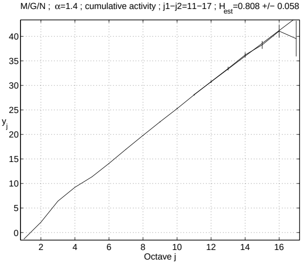
- Consequence: LRD if $\alpha \in [1, 2]$ (infinite variance)

From Heavy-Tails to LRD

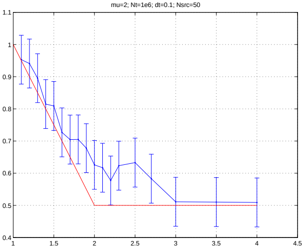
Theoretical (and numerical) evidences



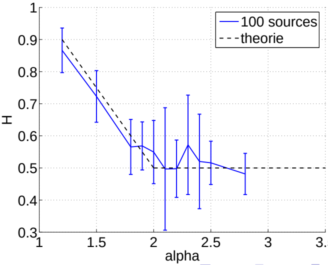
LD S_N



LD Y_N



Simul.

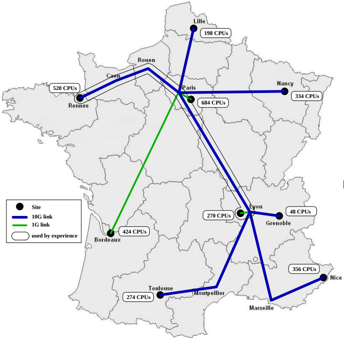


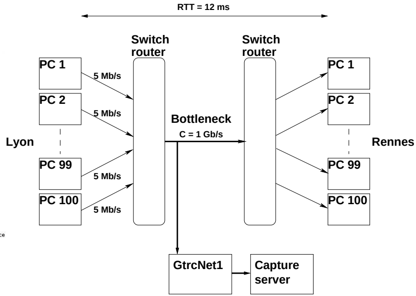
NS

From Heavy-Tails to LRD

Experimental measurements

- Controlled experiences on Grid5000
- Flow's PDF constrained, passive monitoring of resulting traffic.

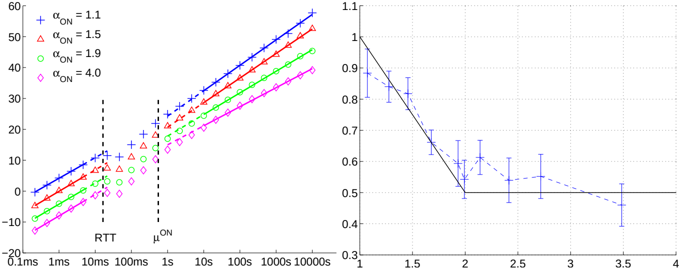




[Loiseau et al., "Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility", IEEE ToN (2010)]

From Heavy-Tails to LRD

Experimental measurements



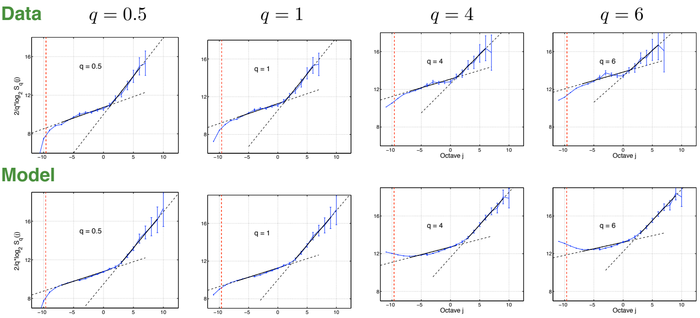
[Loiseau et al., "Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility", IEEE ToN (2010)]

Some more refined models

- **Cluster-Point Processes**: packets arrive in clusters

  [*Cluster Processes, a Natural Language for Network Traffic.* Hohn, Veitch & Abry 2003]

  - Comparison to experimental data [Auckland-IV]



  - Good model for LRD; marginal PDF; intermediate scales. Point process at small scales

Some more refined models

- **Gamma-farima** model = effective model of traffic (simpler!)

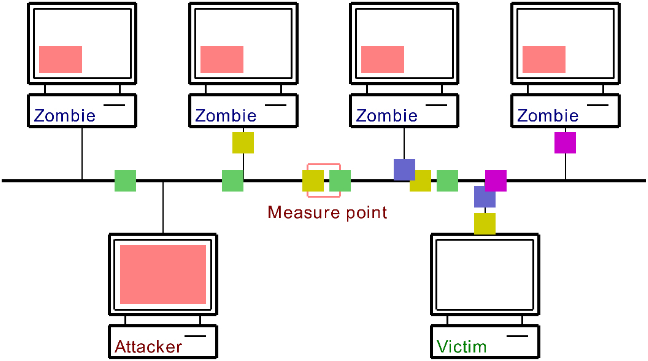
  [*Non-Gaussian and Long Memory Statistical Characterizations for Internet Traffic with Anomalies.* Scherrer, Larrieu, Owezarski, Borgnat & Abry 2007]

  1. **Marginal PDF** as Gamma laws
  2. **farima** = fractionally Intregrated ARMA, models the LRD + short-range correlations

- Some use:
  - traffic model for normal/abnormal situations ($\rightarrow$ detection?)
  - traffic synthesis
  - simulation of chips traffic [Scherrer et al. 2006]
  - simulation of queueing effects [Janowski et. al 2007, 2009]

Anomalies in Internet Traffic – Detection?

- Schematic scenario of DDoS



- Attack with packets without specific signatures
- Objective: detection in low SNR

Anomalies in Internet Traffic – Detection?

Overview of strategies for anomaly detection

- Methods based on **signatures**
  - recognition of packets
  - avantage: robust
  - drawbacks: limited to known anomalies, with specific signatures, scalability with increasing number of anomalies?
- Methods based on **anomalies** or statistical profile
  - use statistical properties of traffic: normal vs. abnormal
  - avantage: versatile, indifferent to number of signatures
  - drawbacks: variability of traffic
  - statistics $\rightarrow$ false alarm vs. detection prob. trade-off

Some ref.: [Brutlag '00], [Barford '02] Lakhina '04] [Kim '06]

Algorithm for detection and identification of anomalies

[*Sketch based Anomaly Detection, Identification,....* Abry, Borgnat, Dewaele. SAINT'07]
[*Extracting Hidden Anomalies using Sketch and Non Gaussian Multiresolution Statistical Detection Procedures.*
Dewaele, Fukuda, Borgnat, Abry & Cho. LSAD Sigcomm'07]

Key Steps:

- A- Sketches (random projection/sampling)

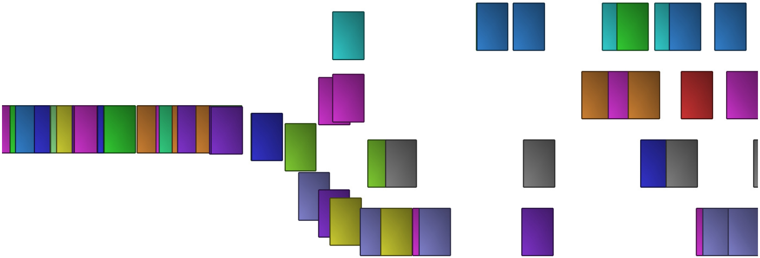
  $\rightarrow$ reference without any prediction or model in time

- B- Multi-scale aggregation (several scales at the same time)

- C- Modelling with non-Gaussian statistics (based on Gamma-farima)

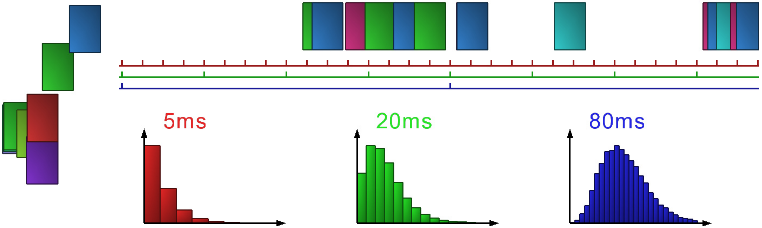
- Detection Test: comparison of traffic across the Sketches

A- Sketches: random projection/sampling



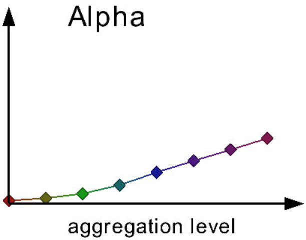
- Output of size N
- key for hashing = IP source , IP destination...

B- Multi-scale Aggregation



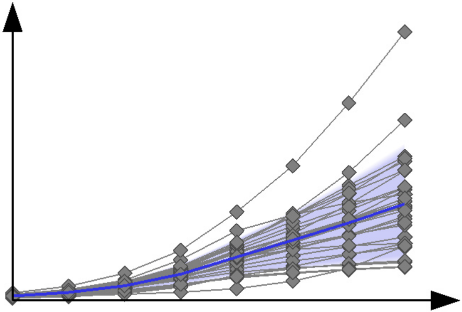
- Aggregated traffic with scales: 5ms, 10ms, ..., 1s

C- Modelling with non-Gaussian statistics



- Gamma laws: parameters $\alpha(\Delta)$ and $\beta(\Delta)$

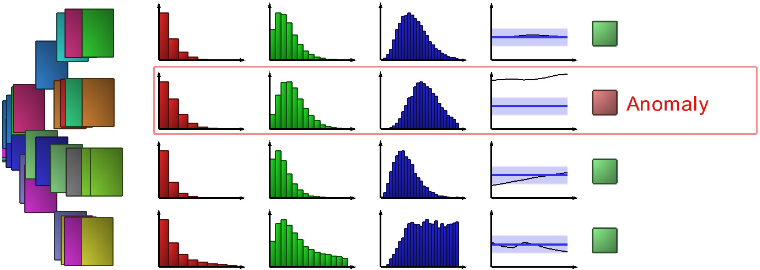
Detection: comparison of traffic across the Sketches



- Compute average and standard deviation across boxes.
- Anomaly = an output is far from the average.

In Mahalanobis distance: $D_\alpha = \left(\frac{1}{J} \sum_{j=1}^{J} \frac{|\alpha^n_{\Delta_j} - \alpha^{Ref}_{\Delta_j}|^2}{\sigma^2_{\alpha,\Delta_j}} \right)^{1/2}$ >threshold.

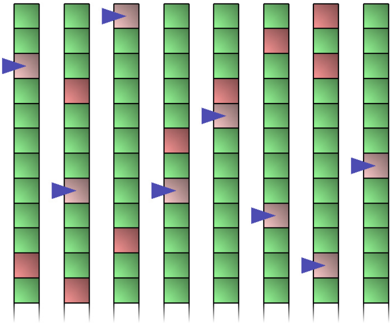
Algorithm: sketches + multiresolution + Gamma statistics



Avantages:

- Enhanced contrast of anomalie wrt the rest of traffic of the output
- Reference extracted from traffic (no problem if evolution)
- Identification of IP responsible or victim of anomalous traffic.

Identification of IP involved



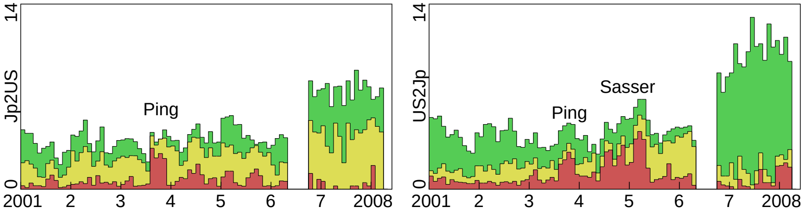
10.23.7.59	Safe
52.27.143.78	Safe
67.12.121.59	Safe
69.22.21.132	Safe
81.82.133.132	Safe
81.82.133.241	Safe
85.102.0.1	Safe
92.131.141.61	Safe
112.27.29.51	?
113.65.56.31	Safe
127.91.66.67	Safe
145.25.10.52	Safe
...	

$N > 5$ sketches: no expected collisions.

- IP that are not always in anomalous outputs = normal
- IP that are **always** in anomalous outputs = anomalies

Results: Longitudinal analysis of anomalies

MAWI dataset: 15’ per day, trans-pacific backbone



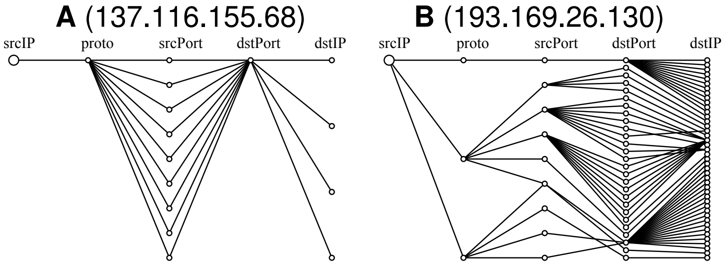
- “Suspected” (green): WWW, P2P, GRE, DNS.
- Mostly attacks (yellow): various mechanisms.
- “Sure attacks” (red): Ping/SYN floods, spoofed,...

Some requirements for “traffic classification”

- High-speed links of Backbones:
  - No bi-directionality
  - No packet payload (useful for a posteriori & online work)
  - Robustness to sampling
- Unsupervised classification:
  - Allow finding new classes of traffic
  - No need for labelled training set
- Host-level analysis
  - vs. usually: flow or packet-level approaches
  - Strengths: cases of mix traffic; network administrator point of view ($\rightarrow$ IP)

Inspiration: Host connection described with Graphlets

BLINC: Multilevel Traffic Classification in the Dark, Karagiannis et al., SIGCOMM 2005.



However, some drawbacks:

- Representation in infinite-dimension space
- Hosts with mixed types of traffic → complex graphlets

Set of quantitative features of connection patterns

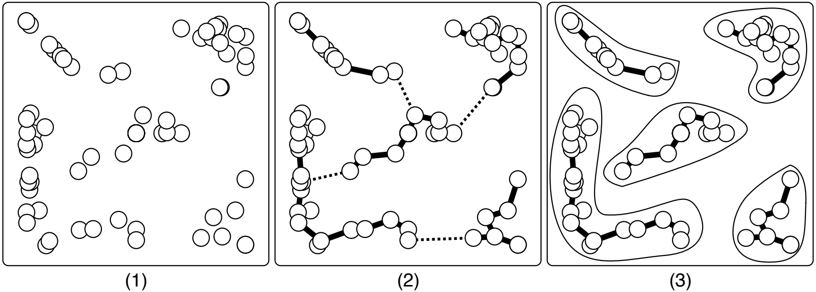
- ***I. Network connectivity***
  - i) the number of peers (or destination IPs)
  - ii) the number source ports, divided by the # of peers (dst IPs)
  - iii) the number of destination ports, divided by the # of peers (dst IPs)
- ***II. Connection dispersion in the network.***
  - iv) the ratio of the entropies of the second and fourth bytes of IPdst
  - v) the ratio of the entropies of the third and fourth bytes

  Entropy $S = -\sum_i p_i \log p_i$
- ***III. Host traffic content.***
  - vi) the mean number of packets per flow
  - vii) the percentage of small size packets ($\leq$ 144 bytes)
  - viii) the percentage of large size packets ($\geq$ 1392 bytes)
  - ix) the entropy of the distribution of medium size packets

These features obey a Parsimony / Relevance trade-off.

Clustering: edge-cut of Minimum Spanning Tree

- (1) A set of hosts into a (reduced 2D) feature space



- (2) the MST with the longest edges in dashed lines
- (3) edge cutting procedure, yields the clusters

Cross-validation with port-based analysis

Id	HTTPr	HHTPa	P2P	Ping	SYN	SMTPr	SMTPa	DNSr	DNSa	SSHr	SSHa	Mix	#Hosts
T_1	**6771**	121	**3357**	427	1	3	59	55	53	46	24	41	11637
T_2	3	**5581**	364	0	0	112	0	0	0	0	8	5	6344
T_3	16	**539**	**802**	9	0	7	0	0	0	3	4	14	1626
T_4	2	197	**892**	250	0	6	0	0	43	2	16	16	1591
T_5	7	22	**382**	13	0	6	0	0	0	2	8	15	572
T_6	51	21	41	**622**	0	0	16	133	58	2	1	7	986
T_7	0	0	**583**	1	0	0	0	0	0	0	0	0	586
C_1	**6138**	0	130	3	18	115	0	119	0	43	2	1003	7875
C_2	**2271**	2	215	16	0	1	1	37	0	12	0	57	2765
C_3	69	0	0	78	**220**	11	0	83	0	0	0	25	524
C_4	**2057**	4	144	1	3	18	0	5	0	1	2	49	2389
C_5	**751**	0	248	0	3	49	0	1	0	17	0	151	1566
C_6	**147**	0	60	0	10	0	0	1	0	1	0	**309**	608
C_7	**224**	0	30	0	8	2	0	0	0	3	0	**193**	530
S_1	0	**4648**	171	0	0	1	0	0	16	0	2	340	5383
S_2	0	**1637**	65	0	0	2	0	0	0	0	3	22	1772
S_3	12	**369**	257	11	0	0	**442**	212	29	1	60	337	1760
S_4	14	**221**	193	6	1	0	**309**	14	124	0	26	47	991
S_5	7	**561**	47	0	0	10	0	0	0	1	2	19	690
S_6	0	**3849**	45	0	0	1	0	0	3	0	2	123	4225
S_7	17	**3578**	191	0	0	63	0	0	0	0	4	32	4056
S_8	0	302	33	0	0	0	116	0	37	0	**1136**	17	1694
S_9	0	**455**	7	0	0	0	0	0	0	0	0	3	476
S_{10}	0	**421**	11	0	0	0	0	0	0	0	0	3	442
P_1	719	186	523	12	44	111	272	239	38	0	29	**1922**	4461
P_2	9	5	**235**	0	15	5	0	1	0	0	5	**251**	560

Comments: Cross-validation with a "Ports"

- The table is relatively sparse: good coherence
- Identified clusters: they fall mostly in the proper "port-based" class
  - T_1 = requests in HTTP and P2P; T_2 = answers over HTTP; T_3 and T_4 = P2P plus some web browsing,
  - C and S well separated in requests / answers
  - P = P2P + mix, not easily in a "port-based" class
- Clusters with a large # of anomalies (T_4, T_6, C_3, C_7): Not found by port-based classes (Exc.: with SYN-flag rule).
- Conclusion: clusters are better representative of hosts than "port-based" classes

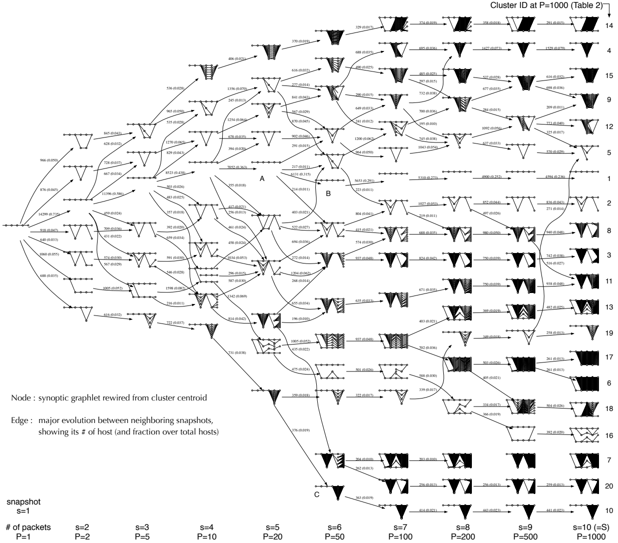
[*Unsupervised host behavior classification from connection patterns.* Dewaele et al., IJNM 2010]

Perspectives in Host & Traffic Classification

- Computation load: takes less than real-time
- Future integration with port-based classifier + anomaly detection + BLINC for automation of cluster labelling
- Methods to compare results of detectors of classifiers
- $\rightarrow$ MAWILab: first attempt of automatic host profiling and anomaly labeling on 9 years of traffic

Perspectives in Host & Traffic Classification

- Automatic Characteristics of Synoptic Graphlets



Conclusion

- Traffic Measurement:
  a tool to understand traffic and network behaviours
- Input from Statistical Signal Processing:
  advanced analysis methods + models (of complexity tailored to applications)
- Some Examples:
  Traffic models; Anomaly detection; Host Classification
- Perspectives :
  - multi-variate setting = several links (or nodes)
  - dynamical models = of the network itself

`perso.ens-lyon.fr/pierre.borgnat`

Conclusion

- Traffic Measurement:
  a tool to understand traffic and network behaviours
- Input from Statistical Signal Processing:
  advanced analysis methods + models (of complexity tailored to applications)
- Some Examples:
  Traffic models; Anomaly detection; Host Classification
- Perspectives :
  - multi-variate setting = several links (or nodes)
  - dynamical models = of the network itself

`perso.ens-lyon.fr/pierre.borgnat`

Supplementary slides

Long-Range Dependence (or Long Memory)

Property pertaining to estimation

- Let X_t be a stationary process with long memory. Then, with $H = 1 - \gamma/2 \in (0.5, 1)$,

$$\lim_{n\to\infty} \mathrm{Var}\left(\sum_{t=1}^{n} X_t\right) / [c\sigma^2 n^{2H}] = \frac{1}{H(2H-1)}.$$

- Aggregation of processes with long-range dependence results in power-law behaviour of the variance of the aggregated processes:

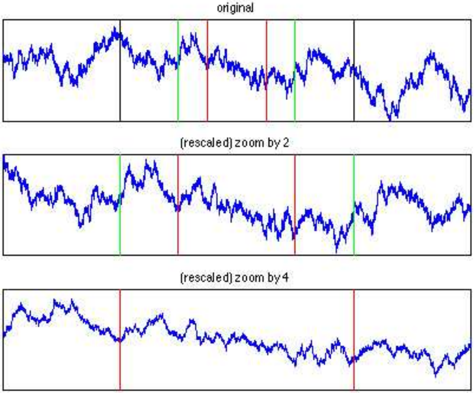
$$\mathbb{E}\left|\frac{1}{N}\sum_{t=pN}^{(p+1)N} X_t\right|^2 \sim N^{-\gamma}, \quad N \to \infty.$$

- Question: Practical estimation of LRD or self-similarity?

Long-Range Dependence (or Long Memory)

One model (among others): Fractional Brownian motion

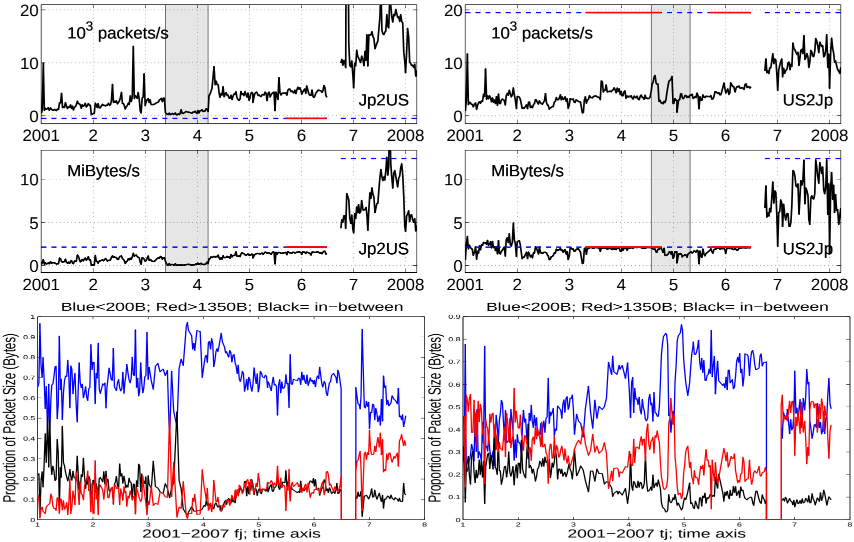
- Self-similar, Gaussian and with stationary increments



- Question: Practical estimation of LRD or self-similarity?

Longitunal study of MAWI backbone dataset

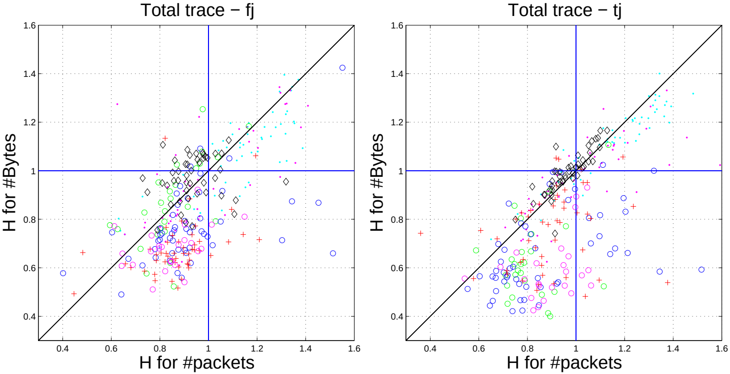
[Borgnat et al. INFOCOM 2009]



Is the LRD the same for packet and byte counts ?

H-parameter estimated without Sketches

Scatter plots of $H(B)$ (byte) vs. $H(P)$ (packet)



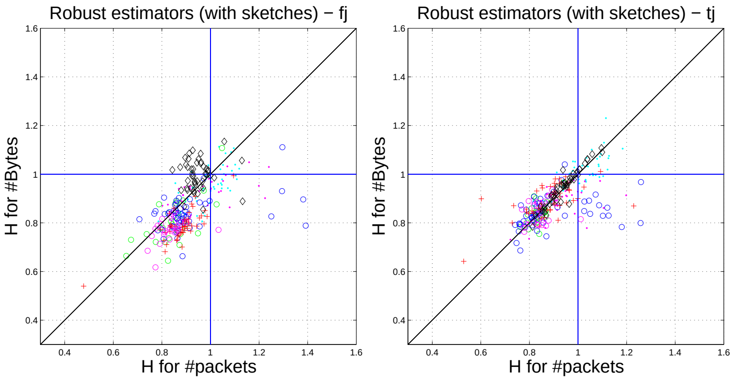
Global estimates. Symbols are: **o**: **B** without congestion; • : **B** with congestion; **+**: **B** anomaly (US2Jp) and restricted traffic (Jp2US); ⋄: **F**.

(Left: Jp2US; Right: US2Jp).

Is the LRD the same for packet and byte counts ?

H-parameter estimated with Sketches

Scatter plots of $H(B)$ (byte) vs. $H(P)$ (packet)



Median-sketch estimates. Symbols are: **o**: **B** without congestion; • : **B** with congestion; **+**: **B** anomaly (US2Jp) and restricted traffic (Jp2US); ◇: **F**.

(Left: Jp2US; Right: US2Jp).