Understanding the Long-Term Self-Similarity of Internet Traffic

Steve Uhlig and Olivier Bonaventure

InfoNet group

University of Namur, Belgium

E-mail : {suhlig,obonaventure}@info.fundp.ac.be

URL : http://www.infonet.fundp.ac.be/

The Traffic Trace

Measurement study

- Collect Netflow records for a Belgian ISP during 6 days.
- Netflow $flow$: total volume between $start$ and end (for layer-4 flows)
- All the incoming traffic (interdomain).

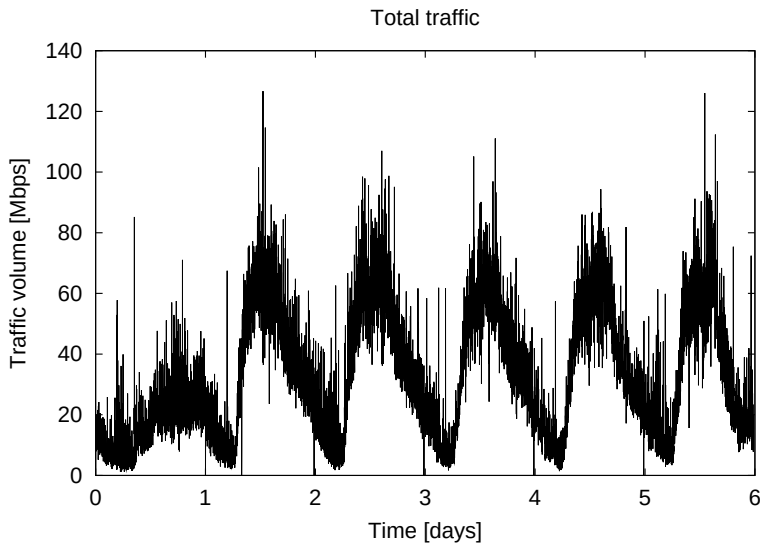
Studied ISP

- BELNET : research and government ISP (`http://www.belnet.be`)
  - high bandwidth links to two transit ISPs, E3 link to SURFNET/AMS-IX, OC-3 link to BNIX, 1.5 DS3 links to TEN-155
  - Main user : University attached to E3 backbone

Total Traffic

Total Traffic

- Granularity of the traffic records = 1 minute
- Represents 2.1 Tbytes of traffic
- Average Incoming traffic : 32 Mbps [97.5% TCP]
- 42 million flows



Self-Similarity

Definition :

Let $X = \{X_i, i \geq 1\}$ be a stationary sequence (our sample)

Define the *m-aggregated* sequence :

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \quad k = 1, 2, ...$$

Then the sequence $X^{(m)} = \{X_k^{(m)} : k = 1, 2, ...\}$ is said *asymptotically self-similar* if

$$X \stackrel{d}{=} m^{1-H} \; X^{(m)} \;\; as \;\; m \to \infty$$

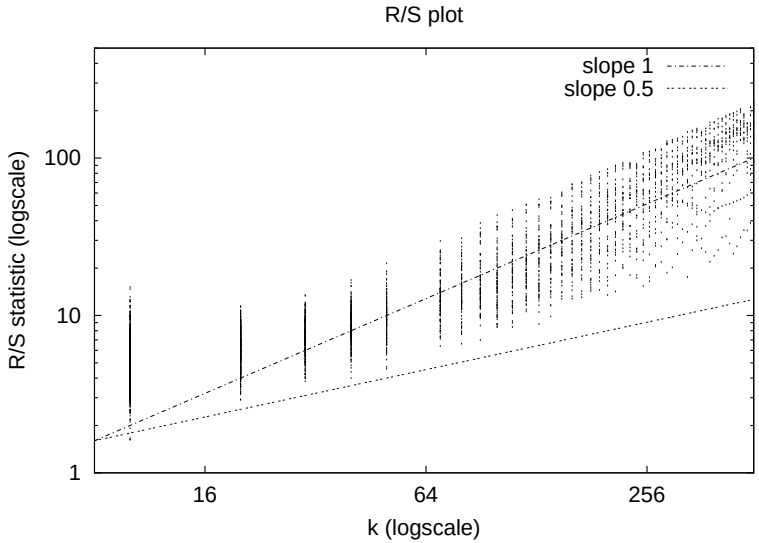
Estimators for Self-Similarity

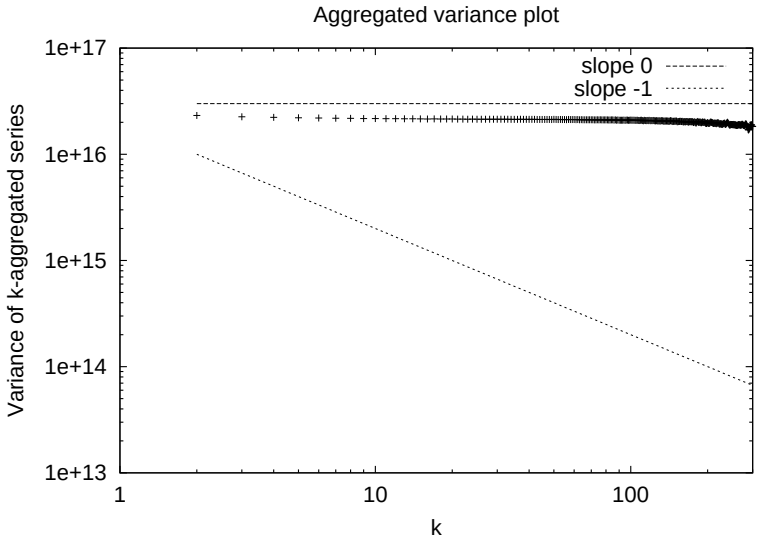
Used estimators

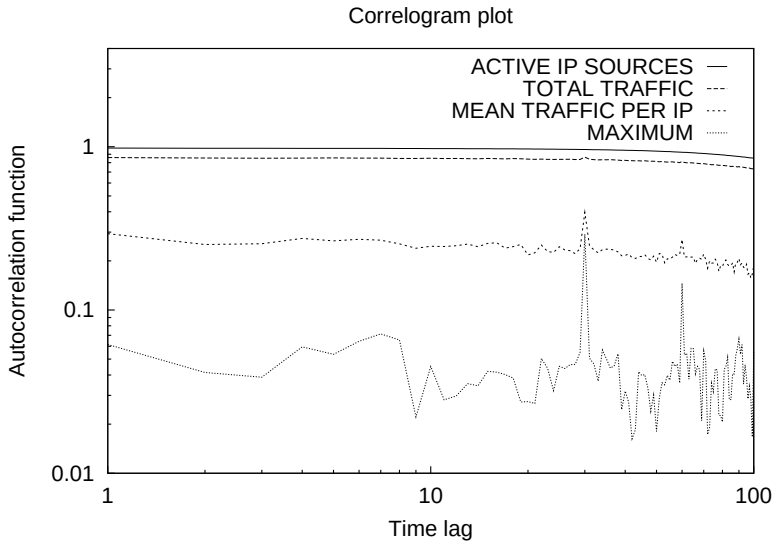
- R/S statistic : log-log plot gives slope $\sim H$
- aggregated variance : log-log plot gives slope $\sim 2H - 2$
- correlogram : log-log plot gives slope $\sim 2H - 2$
- periodogram : log-log plot near origin gives slope $\sim 1 - 2H$

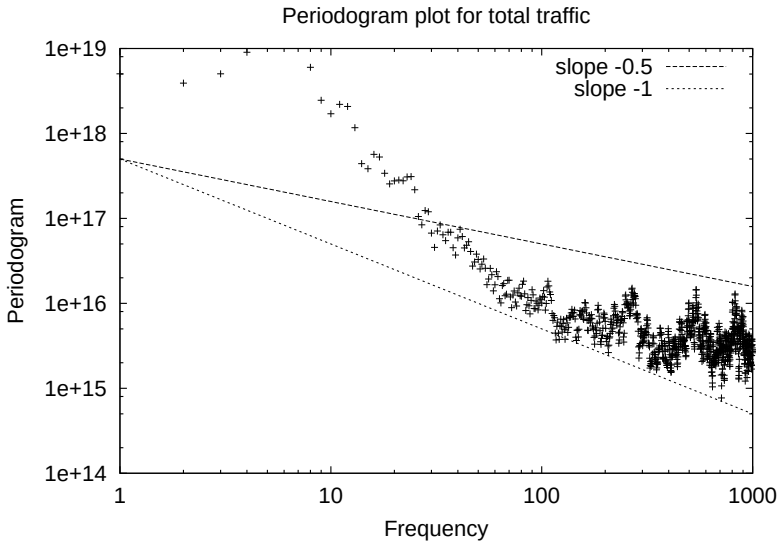
Self-Similarity is asymptotic $\Rightarrow$ estimating H is tricky.

Total Traffic Self-Similarity









Who's who ?

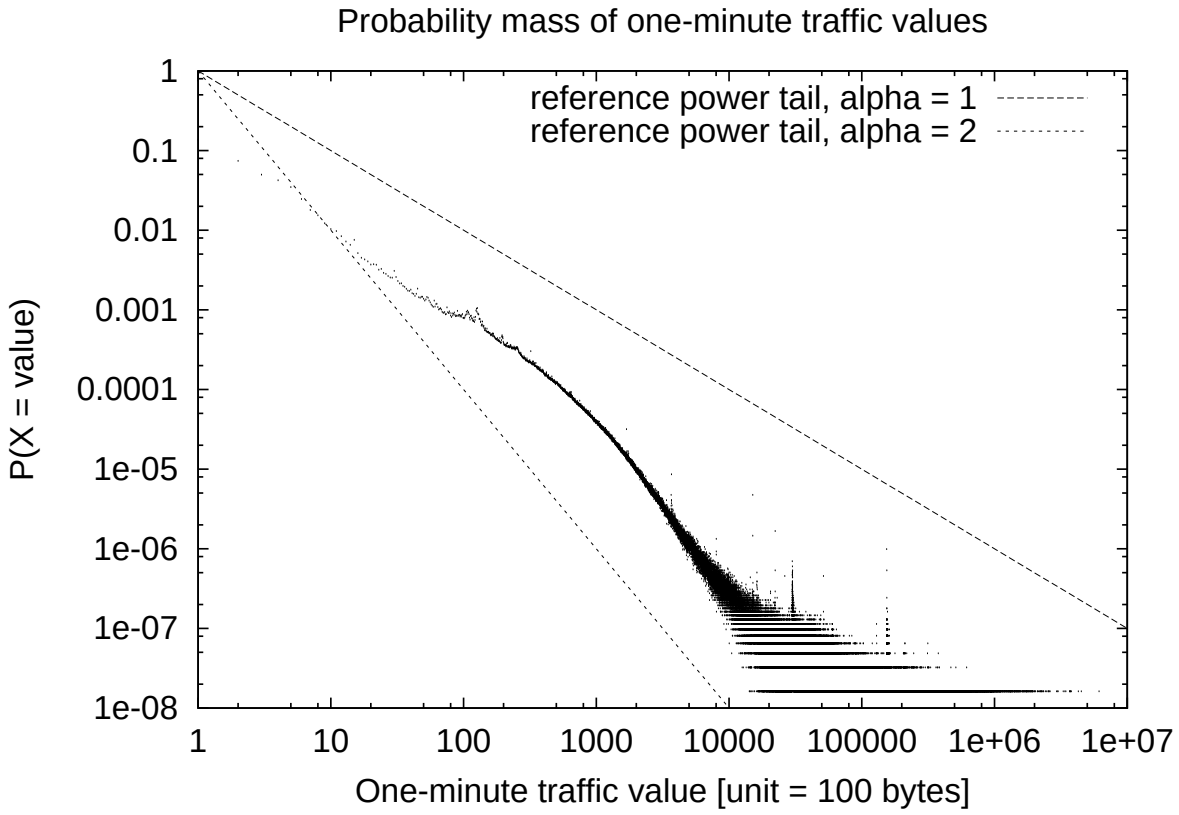
Several factors can explain total traffic self-similarity :

- heavy-tails in flows sizes (or length) : proved to be able to generate self-similarity (Crovella and Bestavros 1996)
- number of IP sources sending traffic : possible factor (proof via results in stochastic processes).
- ...

Heavy-tails often considered in the literature as THE factor for self-similarity.

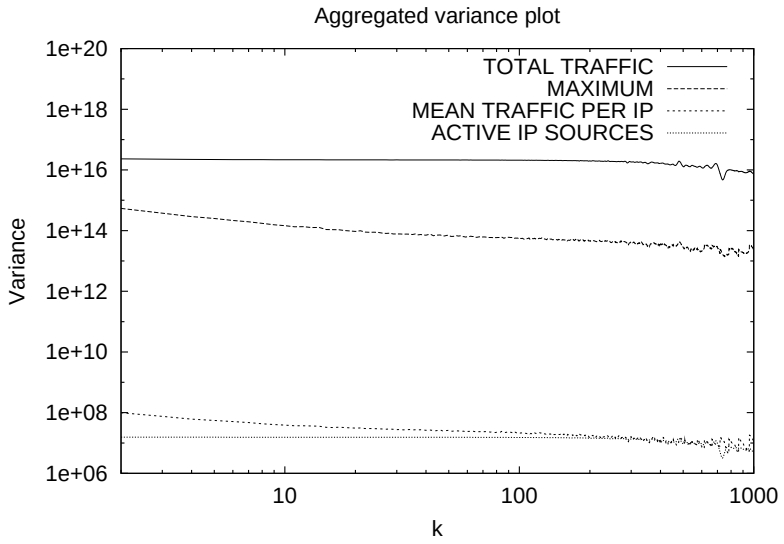
Heavy-Tails

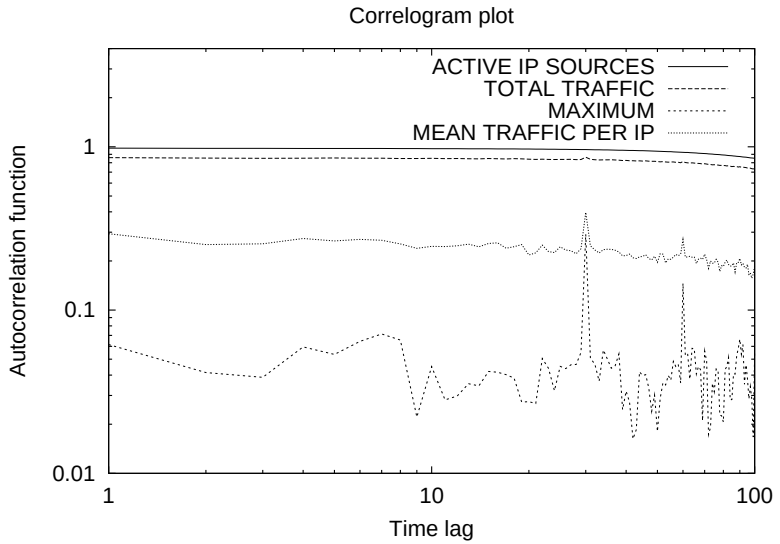
Heavy-tailed distribution (persistence of large values) :
$P[X > x] \sim x^{-\alpha},\ as\ x \to +\infty.$



Dynamics of Traffic Sources

Looking at the evolution of number of 1-minute IP addresses (same for prefixes and ASs) during the week...





Damn, it's self-similar too !

The Role of Heavy-Tails

So the question is : to what extent are those heavy-tails important ?

- sufficient condition for self-similarity
- but to what extent are those large traffic volumes important ?

Let's try the following experiment (or "get rid of these bursts !"):

- Determine total amount of traffic T_k (in bytes) for minute k and the number N_k of IP addresses sending traffic during that minute.
- For each minute k, generate an approximation of the exponential distribution with mean T_k/N_k so that the simulated traffic corresponds to a total of about T_k bytes and a number of points of about N_k points by relying on the exponential distribution formula $P(X = x) = \frac{N_k}{T_k} e^{-(N_k/T_k)x}$.

Experiment (1)

Principle :

For each minute of the week, generate an (discrete) exponential distribution with N_k values (IP sources) for a total of T_k bytes (1-minute traffic volume).

We can do that because exponential distributions are cool : their mean (T_k/N_k) gives it all...

foreach minute k {
    foreach *value* = 0 to $\max_k$ {
        // Attributing to *value* its frequency of occurence
        frequency(value) = $(N_k^2/T_k) * e^{-(N_k/T_k)value}$
        // Attributing to *value* its traffic volume
        volume(value) = $value * (N_k^2/T_k) * e^{-(N_k/T_k)value}$
    }
}

Experiment (2)

Approximations due to discrete distribution:

- cutting the tail of the 1-minute distribution ($\max_k$) :

$$\frac{N_k^2}{T_k}\ e^{-(N_k/T_k)\max_k} \geq \frac{1}{N_k}$$

- deviation for total traffic :

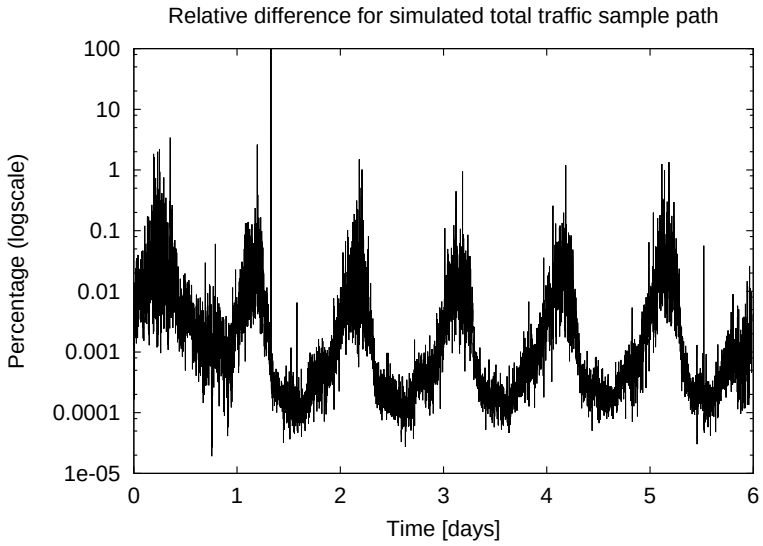
$$\mid (T_k - \sum_{value=0}^{\max_k} volume(value))/T_k \mid$$

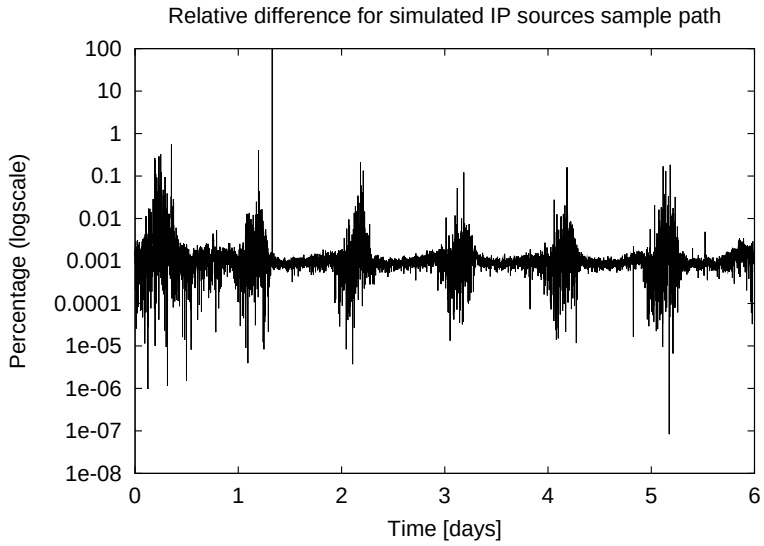
- deviation for IP sources :

$$\mid (N_k - \sum_{value=0}^{\max_k} frequency(value))/N_k \mid$$

Experiment (3)

Evolution of the discretization error:



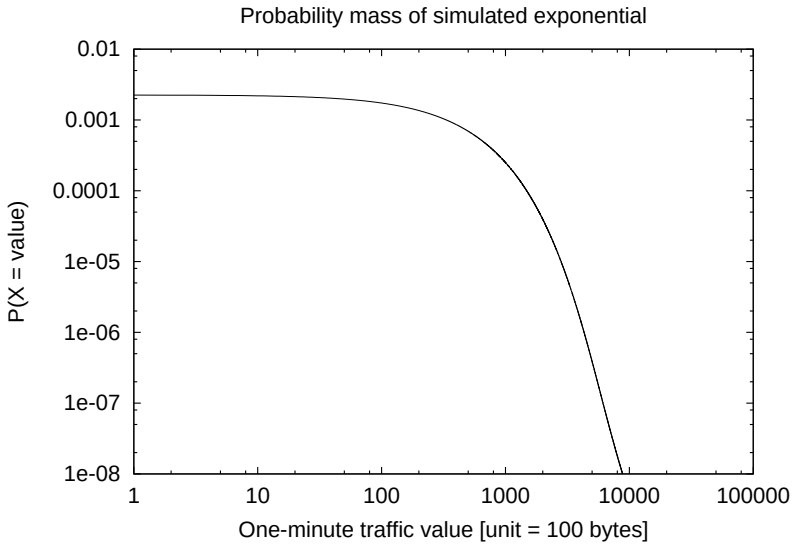


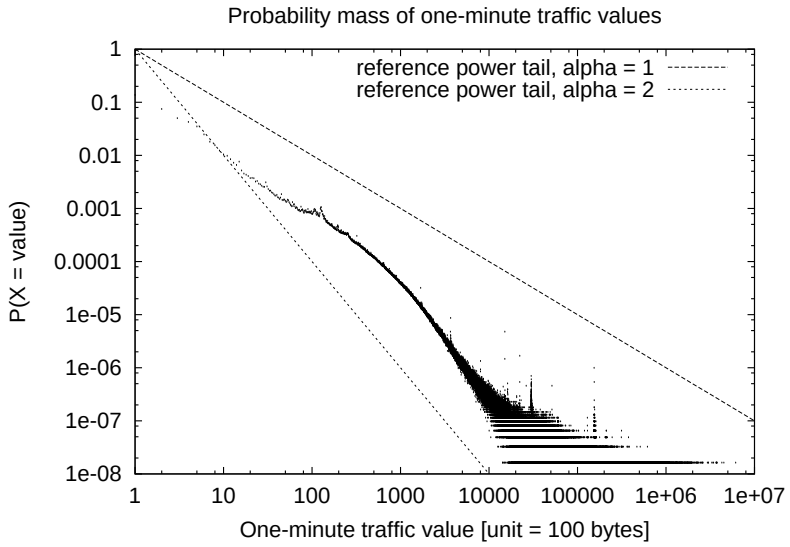
Difference in total traffic ~ 0.01 % and in number of IP sources ~ 0.001 % on average.

A better precision would allow to reduce this already small error.

Experiment (4)

Simulated traffic values for IP sources vs. original values for IP sources :





We hence managed to prevent the large bursts to occur while self-similarity has not changed at all.

$\Rightarrow$ heavy-tails have a limited role in Internet traffic self-similarity on the long-term

Conclusions

- Interdomain traffic is self-similar on the long-term.
- 1-minute IP sources (also prefixes and ASs) sending traffic are self-similar too.
- Changing relative traffic volume (limiting bursts) without changing source dynamics leaves self-similarity unchanged.
- Heavy-tails in volume sent by IP hosts are not THE most important aspect of the traffic self-similarity.
- The problem is probably stochastic, with traffic sources driving the long-term self-similarity.