Self-similar traffic

![Graphs and charts showing self-similar traffic patterns.](image1)

Self-similarity

![Fractal images illustrating self-similarity.](image2)
Aggregate traffic - exact self-similarity

Intuition: self-similar processes “look the same” at all (i.e., over a wide range of) time scales

Def.: A stationary process \( X = (X_k : k \geq 1) \) is called exactly self-similar (with self-similarity parameter \( H \), \( 0 < H < 1 \)), if for all \( m \geq 1 \),

\[
X = m^{1-H} X^{(m)}
\]

[LTWW94] LAN traffic is consistent with exact self-similarity
Variance time plot

Network topology 1989
Network topology 1992
Self-similarity

- Just a mathematical concept?
- What does it mean?

Self-similarity via heavy tails

Math:

Superposition of independent ON/OFF sources is **self-similar**, if durations of periods are **heavy-tailed** with **infinite variance**

Superposition of independent ON/OFF sources is **short-range dependent**, if durations of periods are **light-tailed**
Covariance

Given two random variables $x$, $y$ with means $\mu_x$ and $\mu_y$, their covariance is:

$$\text{Cov}(x, y) = \sigma_{xy}^2 = E[(x - \mu_x)(y - \mu_y)] = E[xy] - E(x)E(y)$$

Their correlation coefficient is the normalized covariance

$$\text{Cor}(x, y) = \rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$
**Short-Range Dependence**

- A stationary process \( X = (X_k : k \geq 1) \) with mean \( y \), variance \( \rho^2 \) and autocorrelation function \( r(k) \), \( k \geq 1 \), is said to exhibit short-range dependence (SRD) if there exists \( 0 < \rho < 1 \) and \( \tau > 0 \) with

\[
    r(k) \tau \rho^{-k} \to 0 \quad \text{as} \quad k \to \infty
\]

- Important feature: Autocorrelations decay (at least) exponentially fast for large lags \( k \)

---

**Poisson process: a SRD processes**
Short-range dependence

- The aggregated process $X^{(m)} = (X^{(m)}(k); k \geq 1)$ tends to second-order white noise, as $k \to \infty$
  
  \[ r^{(m)}(k) \to 0 \quad \text{as} \quad k \to \infty \]

  for all $k \geq 1$, where $r^{(m)}$ denotes the autocorrelation function of $X^{(m)}$

- The variance-time function, i.e., the variance of the sample mean, as a function of $m$, satisfies:

  \[ \text{var}(X^{(m)}) \sim cm^{-1} \quad \text{as} \quad m \to \infty \]

Key features
- Short range dependence = finite correlation length
- Fluctuations over narrow range of time scales
- Plotting $\text{var}(X^{(m)})$ vs. $m$ on log-log scale shows linear relationship for large $m$, with slope $-1$
**Light-tailed distributions**

- $X$ random variable with distribution function $F$.
- $F$ is said to be light-tailed if there exists $c > 0$

\[
(1 - F(X))e^{cx} \to 0 \quad \text{as} \quad x \to \infty
\]

- Important feature: tails decay exponentially fast for large $x$; i.e.,

\[
P[X > x] = 1 - F(X) \sim e^{-x} \quad \text{as} \quad x \to \infty
\]

**Light-tailed distributions**

- Examples: Exponential, Normal, Poisson, Binomial
- Key features:
  - $F$ has limited variability
  - $F$ is tightly concentrated around its mean
  - $F$ has finite moments
  - $P[X > x]$ vs. $x$ on log-linear scale is linear for large $x$
Summary of light-tails and SRD

- Distributional assumptions
  - Light-tails imply limited variability in space
- Assumptions about temporal dynamics
  - SRD implies limited variability over time
- Common characteristics of traditional traffic processes
  - Limited burstiness (in time and space)

Long-range dependence

- A stationary process $X = (X_k : k > 1)$ with mean $y$, variance $\rho^2$ and autocorrelation function $X r(k), k > 1$, is said to exhibit long-range dependence (LRD) if for some $1/2 < H < 1$ and
  $$ r(k) \sim ck^{2H-2} \quad \text{as} \quad k \to \infty $$

  $H$ is called the Hurst parameter

- Important features of LRD
  - Infinite correlation length
  - Fluctuations over all time scales
  - No characteristic time scale
Long-range dependence

- The aggregated process $X^{(m)} = (X^{(m)}(k); k \geq 1)$ tends to non-degenerate limiting process, for for $m, k$ sufficiently large
  $$r^{(m)}(k) \to r(k) \quad \text{as} \quad k \to \infty$$

- The variance-time function satisfies:
  $$\text{var}(X^{(m)}) \sim cm^{2H-2} \quad \text{as} \quad m \to \infty$$

Heavy-tailed distributions

- $X$ random variable with distribution function $F$
- $F$ is said to be heavy-tailed if there exists $c > 0$
  $$1 - F(X) = P[X > x] \sim cx^{-\alpha} \quad \text{as} \quad x \to \infty$$

- Important features:
  1. $1 < \alpha < 2$, $X$ has finite mean but infinite variance
  2. Heavy-tailed implies high variability
  3. Tail decays like a power, hence power-law dist.
  4. Plotting $P[X > x]$ vs. $x$ on log-log scale is linear for large $x$ with slope $\alpha$
Detour
Characteristics of modem calls (~1999)

Interarrival times of modem calls
Durations of modem calls

What about pkts from modem calls
Detour
Characteristics of Web
(~2000)

General characteristics of WWW transfers

TCP connections

time in seconds
General characteristics of WWW transfers

TCP connections

General characteristics of WWW transfers

Number of HTTP requests

0 5 10 15 20

midnight 5am 10am 3pm 8pm

Time in hours

1 hour 5 minutes 25 seconds
# of TCP connections per session

Flow durations
Why is LAN traffic self-similar

Possible explanations:
- Network?
- User behavior?

User behavior:
- Examine characteristics of individual src-dst pairs
- Clustering of packets between src-dst pairs
- Define clusters as ON/OFF periods
- Distribution of ON/OFF periods

SRC/DST traffic matrix
Texture plot
Grouping IP packets into flows

- Group packets with the “same” address
  - Application-level: single transfer web server to client
  - Host-level: multiple transfers from server to client
  - Subnet-level: multiple transfers to a group of clients

- Group packets that are “close” in time
  - 60-second spacing between consecutive packets

ON/OFF periods
ON/OFF periods are heavy-tailed

Self-Similarity via heavy tails

Math:
Superposition of independent ON/OFF sources is **self-similar**, if durations of periods are **heavy-tailed with infinite variance**

Statistical analysis of LAN traffic traces:
- Users are ON/OFF
- ON periods are heavy-tailed (file sizes)
- OFF periods are heavy-tailed (think times)
- Distributions of ON/OFF-periods show heavy tails with infinite variance
Wide area network traffic

How are WANs different from LANs

- Network effects matter: roundtrip delays, queuing, flow control
- Many more source destination pairs (not continuously active)

WAN traffic is not exactly self-similar [PF95, FGWK98]

- Generalize notion of self-similarity
- Examine nature of traffic at application/connection layer
- Beyond self-similarity (where are the network effects)
Asymptotic self-similarity

Def.: A stationary process $X = (X_k : k \geq 1)$ is called asymptotically self-similar (with self-similarity parameter $H$, $0 < H < 1$), if for all large enough $m$:

$$X \approx m^{1-H} X^{(m)}$$

Observations:
- Asymptotic self-similarity is equivalent to long-range dependence of infinite correlation length.
- Asymptotic self-similarity does not specify the small-time scale behavior of a process.

Structural model of WAN traffic

Cox’s construction
- M/G/oo model or birth-immigration process
  - Poisson session arrivals
  - Session durations or session sizes are heavy tailed with infinite variance (i.e., $1 \leq \alpha < 2$)
  - Traffic within session is generated at constant rate
  - The resulting process is (asymptotically second-order) self-similar with self-similarity parameter

$$H = (3 - \alpha) / 2$$
Structural model of WAN

- Telnet and FTP sessions
  - Extract session-level information from WAN traces
  - Test if arrivals are consistent with Poisson
  - Test if arrivals are consistent with independence

Dataset WAN traffic LBL/WRL
Test for Poisson arrivals

Test for heavy tail
Implications (shaded 2%, black 0.5%)

Self-similar?
Self-similar?

Mathematical results

LAN:
- Superposition of independent ON/OFF sources
- ON/OFF periods are heavy-tailed with infinite variance
- Packets per unit time is exactly self-similar

WAN:
- Sessions arriving in a Poisson manner
- Sizes (# packets) are heavy-tailed with infinite variance
- Packets per unit time is asymptotically self-similar
Statistical analysis of WEB

Before Web (1994):
- **Self-similarity at packets per time unit**
  - Poisson arrivals at application layer (FTP, Telnet)
  - Heavy-tailed session durations/sizes

Since Web (1995)?
- Arrivals of User session
- # of Web requests per session
- Dist. of # of bytes, pkts, duration per request?

Web client trace analysis 1995

- Modified Web browser (Mosaic)
- Population: students at BU
- Duration: 21 Nov 94 to 8 May 95

<table>
<thead>
<tr>
<th>Sessions</th>
<th>4,700</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>591</td>
</tr>
<tr>
<td>URLs Requested</td>
<td>575,775</td>
</tr>
<tr>
<td>Files Transferred</td>
<td>130,140</td>
</tr>
<tr>
<td>Unique Files Requested</td>
<td>46,830</td>
</tr>
<tr>
<td>Bytes Requested</td>
<td>2,713 MB</td>
</tr>
<tr>
<td>Bytes Transferred</td>
<td>1,849 MB</td>
</tr>
<tr>
<td>Unique Bytes Requested</td>
<td>1,088 MB</td>
</tr>
</tbody>
</table>
What about WEB traffic

Durations of WEB transfers???
File size of WEB transfers???

Unique files vs. files transferred?
What about the available files

What about off times?

Users

Web page

HTTP Request 1

HTTP Request 2

HTTP Request 3

HTTP Request 4

HTTP Request 4

TCP 1

TCP 2

TCP 3

TCP 4
What about the WEB

Interarrival times of URL requests
Statistical analysis of WAN traffic Traces

Before Web (1994):
- **Self-similarity at packets per time unit**
  - Poisson arrivals at application layer (FTP, Telnet)
  - Heavy-tailed session durations/sizes

Since Web (1995):
- **Self-similarity at # of TCP connections per time unit**
  - Poisson arrivals of User session (modem session)
  - Heavy-tailed # of TCP connections per session