Self-similarity, correlations and networks.

(The internet is fractal. We wish it wasn't.)

Richard Clegg (richard@richardclegg.org)

Networks and Nonlinear Dynamics Group,

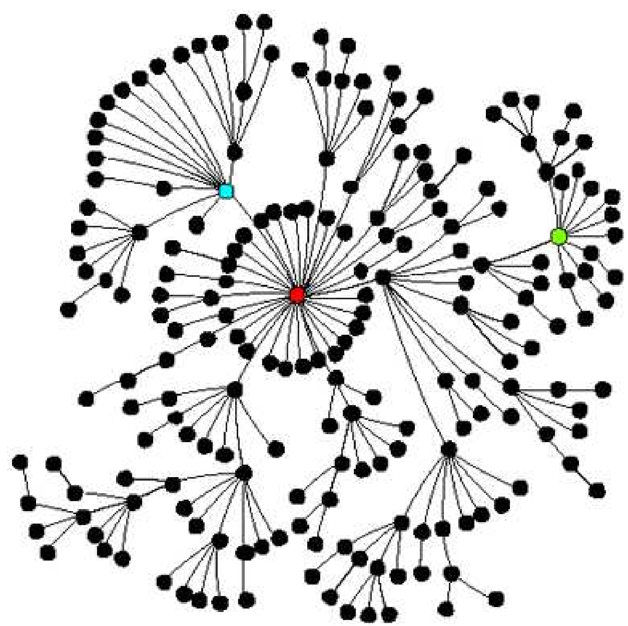
Department of Mathematics,

University of York

Slides prepared using the Prosper package and LaTeX

Talk plan

- What are scaling properties?
- Why should we be interested in them?
- Why are internet engineers interested in LRD?
- How does LRD arise in the internet?

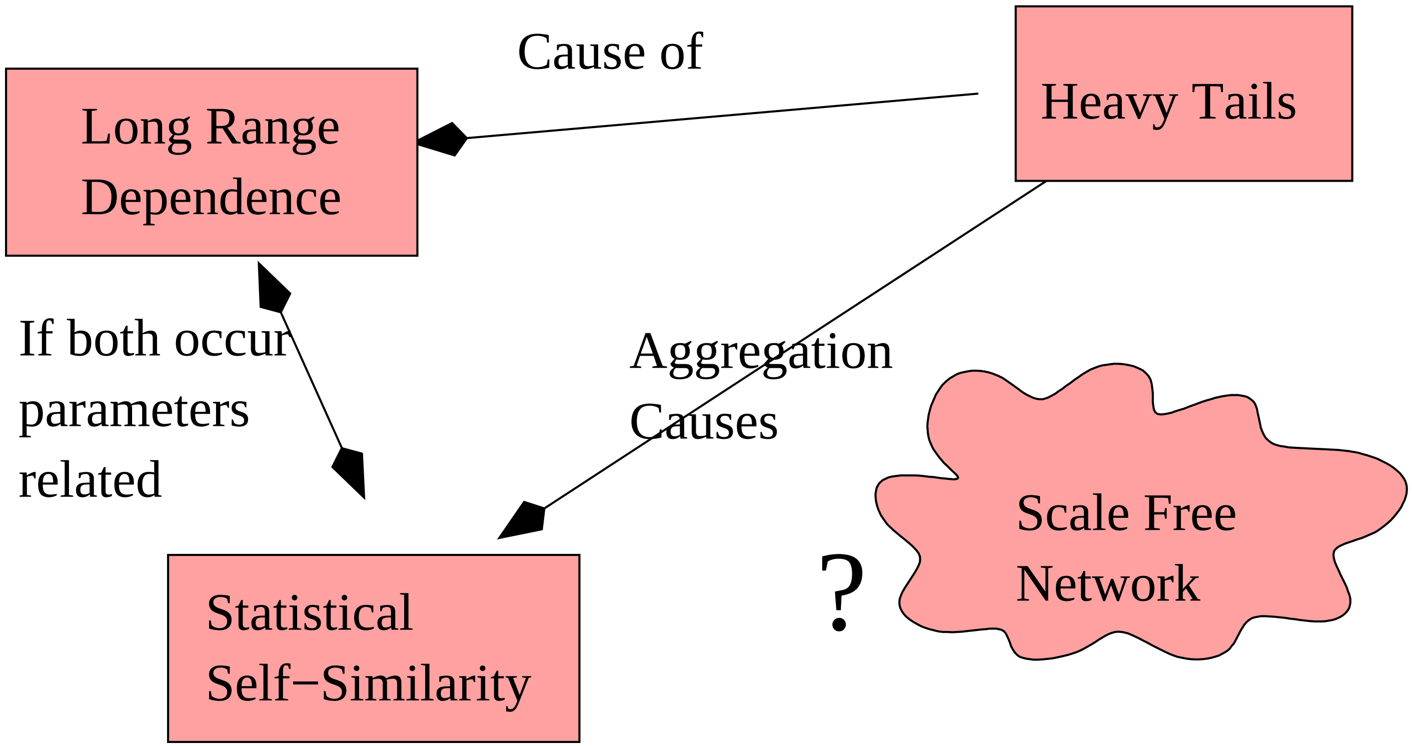


Some Scaling Properties Roughly Defined

Scaling laws are everywhere.

- Statistically Self-Similar: The distribution of a process is the *same* after stretching $Y_t \stackrel{d}{=} c^{-H} Y_{ct}$. Examples: coastlines, tree-bark, internet traffic traces.
- Long-Range Dependent: A process has significant correlations even over long time scales. $\rho(k) \sim k^{-\alpha}$ for $\alpha \in (0,1)$. Examples: global temperature, internet traffic traces, Nile river minima.
- Heavy Tailed: Distribution where extreme events still have a significant likelihood. $\mathbb{P}\left[X > x\right] \sim x^{-\beta}$ for $\beta \in (0,2)$ Examples: heights of trees, frequencies of words, lengths of file in the internet.
- Scale Free Network: k is number of connections. $\mathbb{P}\left[k\right] \sim k^{-\lambda}$. (Internet connectivity, airport "hubs", STD transmission).
- So many power laws — how do they all interact?

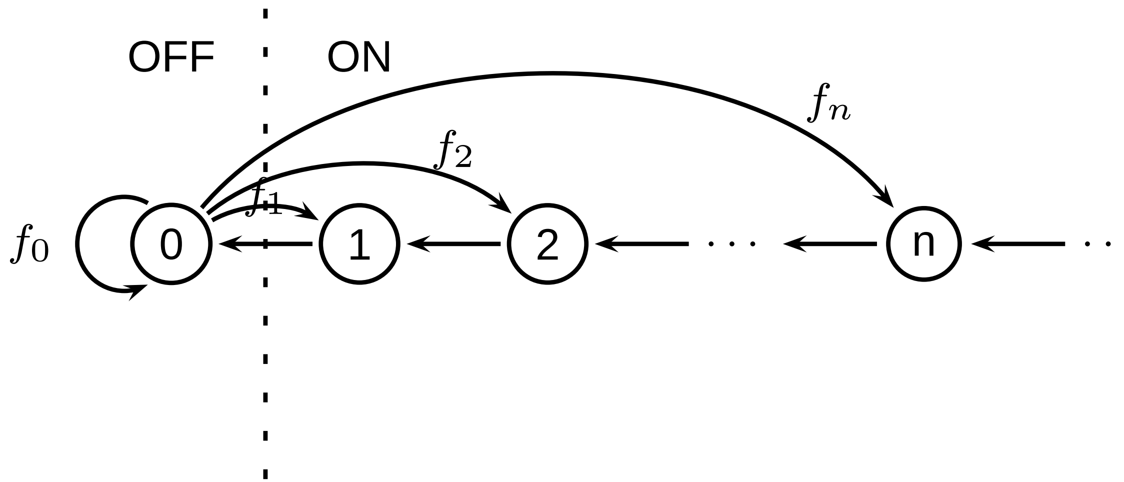
Some Inter-relations



If Y_t is stat. self similar $1/2 < H < 1$ with stationary increments $X_t = Y_t - Y_{t-1}$ then X_t has LRD and same Hurst H [Beran, 1994, page 51].

Aggregation of many heavy-tailed processes is a self-similar process with related parameter [Taqqu et al., 1997].

A Markov Chain Exhibiting Scaling



For an ON-OFF series with LRD, with parameter $\alpha = 2 - 2H$ and mean $1 - \pi_0$.

$$f_k = \frac{1-\pi_0}{\pi_0} \left[k^{-\alpha} - 2(k+1)^{-\alpha} + (k+2)^{-\alpha} \right],$$

for $k > 0$ and,

$$f_0 = 1 - \sum_{i=1}^{\infty} f_i = 1 - \frac{1-\pi_0}{\pi_0} \left[1 - 2^{-\alpha} \right].$$

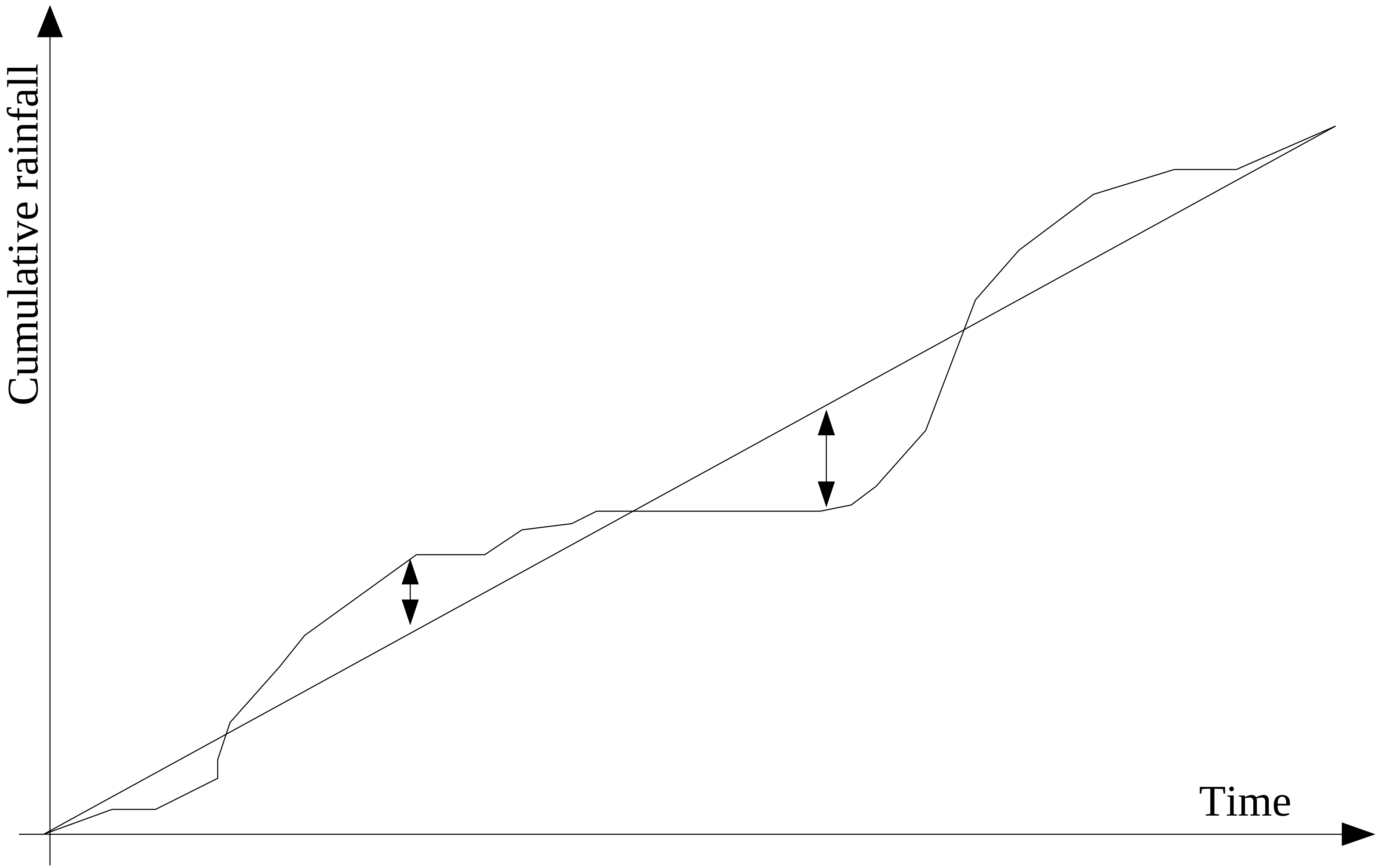
Tracing the Source of the Nile

- [Hurst, 1951] Investigated minima in the Nile river between 622 and 1281 A.D. His goal was to investigate the idea of designing the ideal reservoir.
- The Nile river data is now thought of as a classic “LRD” data set.
- This data (and LRD data in general):
  - Overall appears stationary.
  - Contains long high and low periods.
  - Cycles of a number of frequencies seem to appear but in a random order.
- Mandelbrot refers to LRD as “The Joseph effect” (after the “Seven fat years and seven lean years”).
- LRD is also known as long memory. It is characterised by the Hurst parameter $H \in (1/2, 1)$.

The Horrible Properties of LRD

- Computationally, LRD is a nightmare to work with.
- Consider $\rho(k)$ — the effect we are looking for is at large k we only have many samples for small k. Standard estimators for $\rho(k)$ are biased for large k.
- The sample mean converges at a rate proportional to n^{2H-2} not n^{-1}.
- The sample variance S^2 is no longer an unbiased estimate of the variance σ^2.
- If we take standard techniques for confidence intervals then, as $n \to \infty$ a statistic will be outside a given confidence interval a.s. no matter how small that confidence interval.
- Only investigate LRD if you have a “large” data set (hundreds are good, thousands are better, millions are nice).

How to Dam the Nile



The R/S statistic

- Take $R(t,k)$ (the range beginning at t for time k) and normalise it with $S(t,k)$.
- How does this rescaled range change as k increases?
- Given certain conditions [Mandelbrot, 1975]

$$\frac{R(t,k)}{S(t,k)} \xrightarrow{d} \varepsilon k^H,$$

where H is the Hurst parm. and ε is an r.v.

- For large k a log log plot of R/S vs k is straight line of slope H.
- Actually this is a *terrible* measure of H (biased).
- Local Whittle and Wavelet based estimators are a better alternative.

Measuring LRD

- Measuring the ACF is not a good way to establish the presence of LRD.
- LRD is detected in the slope at high lags. ACF is only accurate at low lags. (ACF estimator is biased in presence of LRD).
- Some biased estimators with poor convergence performance.
- All are vulnerable to some extent to non-stationarities in the data.
- Periodicity and trends in particular can be a problem.
- While some estimators give confidence intervals, often results from different estimators do not agree even within 95% intervals.
- More information: http://math.bu.edu/people/murad/methods/index.html

LRD and the Internet

- In 1993 LRD (and self-similarity) was found in a time series of bytes/unit time [Leland et al., 1993] measured on an Ethernet LAN.
- This finding has been repeated a number of times by a large number of authors (however recent evidence suggests this may not happen in the core).
- A higher Hurst parameter often increases delays in a network. Packet loss also suffers.
- If buffer provisioning is done using the assumption of Poisson traffic then the network will be underspecifed.
- The Hurst parameter is a dominant characteristic for a number of packet traffic engineering problems.

Sources of LRD

(1) Data is LRD at Source

- Claim arises from measurements on VBR video traffi c.
- Pictures are updated by sending changes.
- A still scene is few changes, a cut or pan is a lot of changes.

(2) Data arise from aggregation of heavy tailed ON-OFF sources.

- It can be shown [Taqqu et al., 1997] that ON/OFF sources with heavy-tailed train lengths leads to self-similarity.
- It has been observed that the sizes of fi les transferred on the internet follow a heavy-tailed distribution.

Sources of LRD (continued)

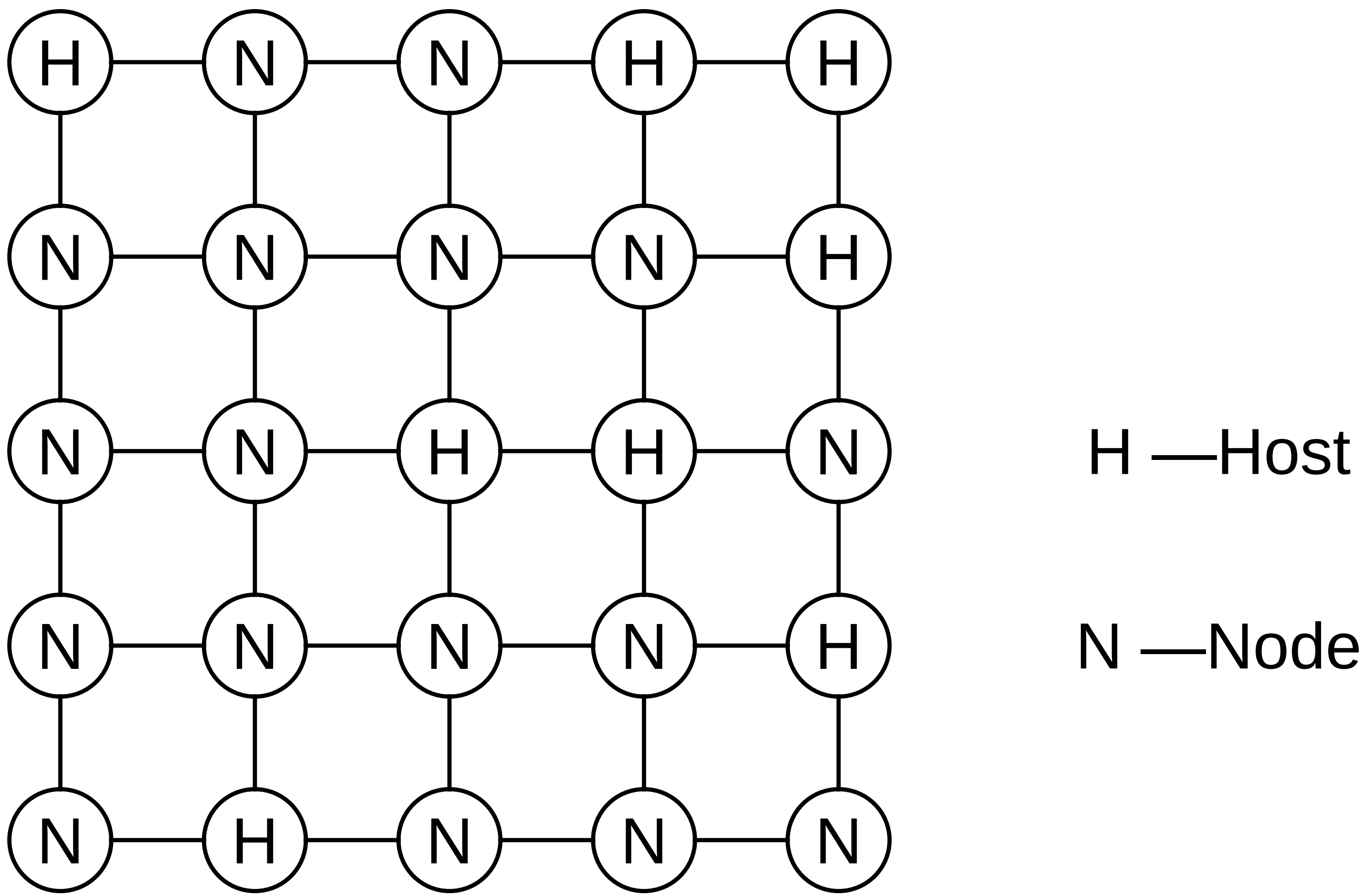
(3) LRD arises from feedback mechanisms in the TCP protocol.

- This claim comes from Markov models of TCP timeout and retransmission.
- A Markov model is used to show that the doubling of timeouts can cause correlations in timeseries of transmitted data.
- Modelling shows that this can lead to LRD over certain timescales (“local” LRD).

(4) LRD arises from network topology or routing.

- Consider a simulation on a Manhattan network with randomly distributed sources and sinks.
- The sources produce Poisson traffic.
- Packets find their shortest route to the sink (accounting for the traffic on the next hop).
- In this simple situation the aggregated traffic shows LRD.

Simple experimental network



Simulation Model

- Manhattan network with randomly dispersed hosts.
- Hosts may produce Poisson or LRD traffic which is sent to a randomly selected host.
- Packets route based upon a “least hops to destination” algorithm.
- However, when routes are equal hops, the least congested hop is chosen.
- Alternatively, a “fixed route” algorithm may be used.
- Congestion is all at nodes — nodes send one packet per simulation step.
- Routing seems to increase the amount of LRD. Hurst increases or becomes present.
- Without routing, there should be no LRD present. This seems to be the case (but is hard to be sure).

Queuing and the Hurst Parameter

$$B([s,t)) = A([s,t)) + Q(s) - Q(t) = A([s,t)) + \Delta Q(s,t)$$

Notation $\nu_X = \text{var}\,(X(s))$ and $\nu_X(x) = \text{var}\,(X[s,s+x))$.
Assume $\text{E}\,[\Delta Q(s,t)] = 0$ and $\nu_A(x) = \text{var}\,(A([s,s+x))) \sim \sigma^2 x^{2H}$.

How does queuing affect variance and Hurst?

$$\begin{aligned}
|\nu_B(x) - \nu_A(x)| &= |\text{var}\,(A([s,s+x) - \Delta Q(s,s+x)) - \text{var}\,(A([s,s+x))\,| \\
&= |2\text{cov}\,(A([s,s+x)), \Delta Q(s,s+x)) + \text{var}\,(\Delta Q(s,s+x))\,| \\
&\le 2\nu_A(x)^{1/2}(4\nu_Q)^{1/2} + 4\nu_Q \qquad (\text{Note}:\text{var}\,(\Delta Q(s,s+x)) \le 4\nu_Q) \\
&\sim 4\sigma x^H \nu_Q^{1/2} + 4\nu_Q.
\end{aligned}$$

If we assume that the queue has a finite second moment then $\nu_B \sim \nu_A$ since $4\sigma x^H \nu_Q^{1/2} + 4\nu_Q$ is negligible compared to $\sigma^2 x^{2H}$.

Conclusions and sources of info

- Scaling laws are a ubiquitous phenomenon in nature and engineered systems.
- This subject is of particular concern to internet traffic engineers.
- Real-life effects of such (seemingly obscure) properties can be a real concern.
- More info
  - This talk online `www.richardclegg.org/pubs`.
  - Mathematics of LRD [Beran, 1994].
  - Heavy-Tails (collection of research papers) [Adler et al., 1998].
  - LRD (collection of research papers) [Doukhan et al., 2003].
  - LRD (intro. in context of teletraffic) [Clegg, 2004, chapter 1].

Bibliography

References

[Adler et al., 1998] Adler, R. J., Feldman, R. E., and Taqqu, M. S. (1998). *A Practical Guide to Heavy Tails*. Birkhäuser.

[Beran, 1994] Beran, J. (1994). *Statistics For Long-Memory Processes*. Chapman and Hall.

[Clegg, 2004] Clegg, R. G. (2004). *The Statistics of Dynamic Networks*. PhD thesis, Department of Mathematics, University of York.

[Doukhan et al., 2003] Doukhan, P., Oppenheim, G., and Taqqu, M. S. (2003). *Theory and Applications of Long-Range Dependence*. Birkhäuser.

[Hurst, 1951] Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, pages 770–808.

[Leland et al., 1993] Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1993). On the self-similar nature of Ethernet traffi c. In Sidhu, D. P., editor, *Proc. ACM SIGCOMM*, pages 183–193, San Francisco, California.

[Mandelbrot, 1975] Mandelbrot, B. B. (1975). Limit theorems of the self-normalised range for weakly and strongly dependent proceses. *Z. Wahr. verw. Geb.*, pages 271–285.

[Taqqu et al., 1997] Taqqu, M. S., Willginer, W., and Sherman, R. (1997). Proof of a fundamental result in self-similar traffi c modeling. *Comp. Comm. Rev.*, 27:5 – 23.