Analysing Long-Range Dependence in Teletraffic

Modelling and Measuring LRD

Richard Clegg (richard@manor.york.ac.uk

Networks and Nonlinear Dynamics Group,

Department of Mathematics,

University of York

Slides prepared using the Prosper package and LaTeX

What are we doing?

- Looking at the statistical nature of internet traffic.
- Attempting to isolate the statistical phenomenon of Long-Range Dependence (LRD).
- We want to find the root cause of LRD in networks.
- Measurements on real network data.
- Modelling the situation with Markov Chains.
- Simulation using ns.

Long-Range Dependence

- LRD (Long Memory, “The Joseph Effect”) was first discovered by Hurst
- A typical process (finite Markov chain, Poisson process or finite ARIMA) has an exponentially decaying ACF tail.
- That is $R(k) \sim a^k$ where $(0 < a < 1)$.
- For LRD $R(k) \sim k^{-\alpha}$ where $(0 < \alpha < 1)$.
- Has an unsummable ACF $\sum_{k=1}^{\infty} R(k) = \infty$.
- Characterised by the Hurst parameter $\frac{1}{2} < H < 1$ where $H = \frac{1+\alpha}{2}$.

Describing Long-Range Depedence

- We can think of LRD in a number of ways:
  1. A significant level of correlation over all time scales.
  2. A process which is *bursty over any time scale we consider.*
  3. *A process with a pole (usually at 0) in the frequency spectrum.*
  4. *An AR(∞) process.*
- LRD is related to statistical self-similarity.
- LRD processes are problematic statistically (for example, convergence of mean estimates is slow).

LRD in Networks

- In 1993 Leland, Taqqu, Willinger and Wilson measured LRD in a time series of packet per unit time in Ethernet data.
- The correlation structure induced by LRD can cause significant problems for queuing systems.
- A stream of data where the packets/unit time exhibits LRD may well have significantly worse queuing performance than one.
- There is now a significant body of research into LRD in networks (several hundred papers in the last ten years).

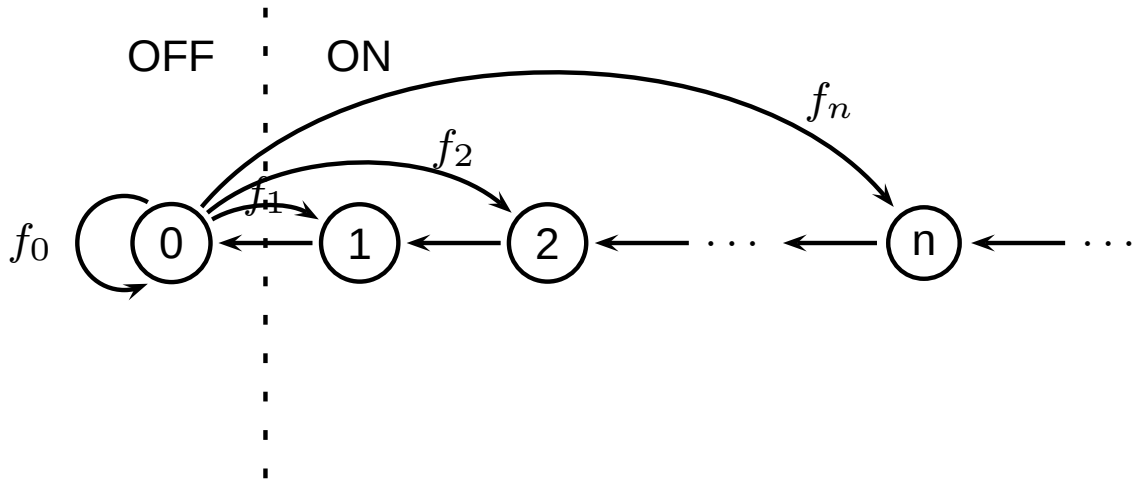
Models with LRD

Several models are commonly used in the telecomms field to investigate LRD.

- Fractional Brownian Motion: fBM
- Fractional Arima: FARIMA
- Iterations of the Double Intermittency Map
- Frequency Domain Techniques (for example Wavelet based reconstruction)

Do we *really* need another model?

The Infinite Markov Model



$$\mathbf{P} = \begin{bmatrix} f_0 & f_1 & f_2 & \dots & f_n & \dots \\ 1 & 0 & 0 & \dots & 0 & \dots \\ 0 & 1 & 0 & \dots & 0 & \dots \\ 0 & 0 & 1 & \dots & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \end{bmatrix}$$

Markov Basics

- When is this model *ergodic* (*irreducible*, *aperiodic* and *recurrent non-null*)?
- Irreducible (all states can be reached from any state) iff $\forall i \exists j > i : f_j > 0$.
- Aperiodic if $f_0 > 0$.
- An irreducible and aperiodic chain is recurrent non null (the mean recurrence time of any state is finite) iff $\sum_{i=0}^{\infty} i f_i < \infty$.
- An ergodic chain has *equilibrium probabilities* π_j for each state j.

Recurrent Non Null

$$Pr[\text{First return to } j \text{ is after } n \text{ steps}] = r_j(n)$$

Now we wish to calculate the mean return time.

$$\left(M_j = \sum_{n=1}^{\infty} n r_j(n) \right) < \infty \Leftrightarrow \text{recurrent non null}$$

$$M_0 = \sum_{i=0}^{\infty} (i+1) f_i = \sum_{i=0}^{\infty} f_i + \sum_{i=0}^{\infty} i f_i = 1 + \sum_{i=0}^{\infty} i f_i$$

For an irreducible, aperiodic chain if one state is recurrent non null, then all states must be $\rightarrow$ ergodicity.

A Finite Model

Create a finite model - states $0 \rightarrow N$

Define the transition probabilities g_i^N.

$$g_i^N = \begin{cases} f_i & 0 < i < N \\ \frac{1}{N} \sum_{i=N}^{\infty} i f_i & i = N \\ 1 - \sum_{i=1}^{N} g_i^N & i = 0 \end{cases}$$

Call the equil. prob. of the ith state π_i^N. We can easily show:

$$\pi_i^N = \pi_0^N \sum_{j=i}^{N} g_j^N$$

Calculating the system ACF

- For binary system Y_t then:

$$R(k) = Pr[Y_t = 1 | Y_{t+k} = 1] + Pr[Y_t = 0 | Y_{t+k} = 0] + 1$$

$$Pr[Y_{t+k} = 0 | Y_t = 0] = \left| [1 \;\; 0 \;\; \ldots 0] P^k [1 \;\; 0 \ldots 0]^T \right|$$

- And obviously $Pr[Y_t = 0] = Pr[X_t = 0] = \pi_0$.
- We can create a similar equation for the ON states.
- *BUT* P^k is intractable analytically.

Inducing LRD Correlation Structure

- Unbroken runs of k 0s will clearly decay exponentially with k. The f_i values set the decay of unbroken runs of 1s.
- Part of a run of k or more if $X_t \geq k$.
- Control decay of $\sum_{i=k}^{\infty} \pi_i$.
- For LRD $\sum_{i=k}^{\infty} \pi_i \sim k^{-\alpha}$ (some hand-waving here).
- Strict condition $\sum_{i=k}^{\infty} \pi_i = Ck^{-\alpha}$ for $k > 0$.
- Since $\pi_0 = 1 - \sum_{i=1}^{\infty} \pi_i$ then $C = 1 - \pi_0$.

Generating the Correlation Structure

- This system is trivially solved and we can calculate the values of f_k.
- For $k > 0$ we have (note problems with some values):

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

- The attractive thing about this series is that it is telescoping. For example.

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

Directly Using The Infinite Chain

- We can directly use the infinite chain in calculations if we use a simple algorithm. First define $F(j,k) = \sum_{i=j}^{k} f_i$ where $(j \leq k)$.
- We can see that if X is the next state moved to by the chain after state 0 then we have $Pr\{X \in [i,j]\} = F(j,k)$.
- The telescoping property makes $F(j,k)$ easy to calculate. For $j > 0$ and $k < \infty$ we have:

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

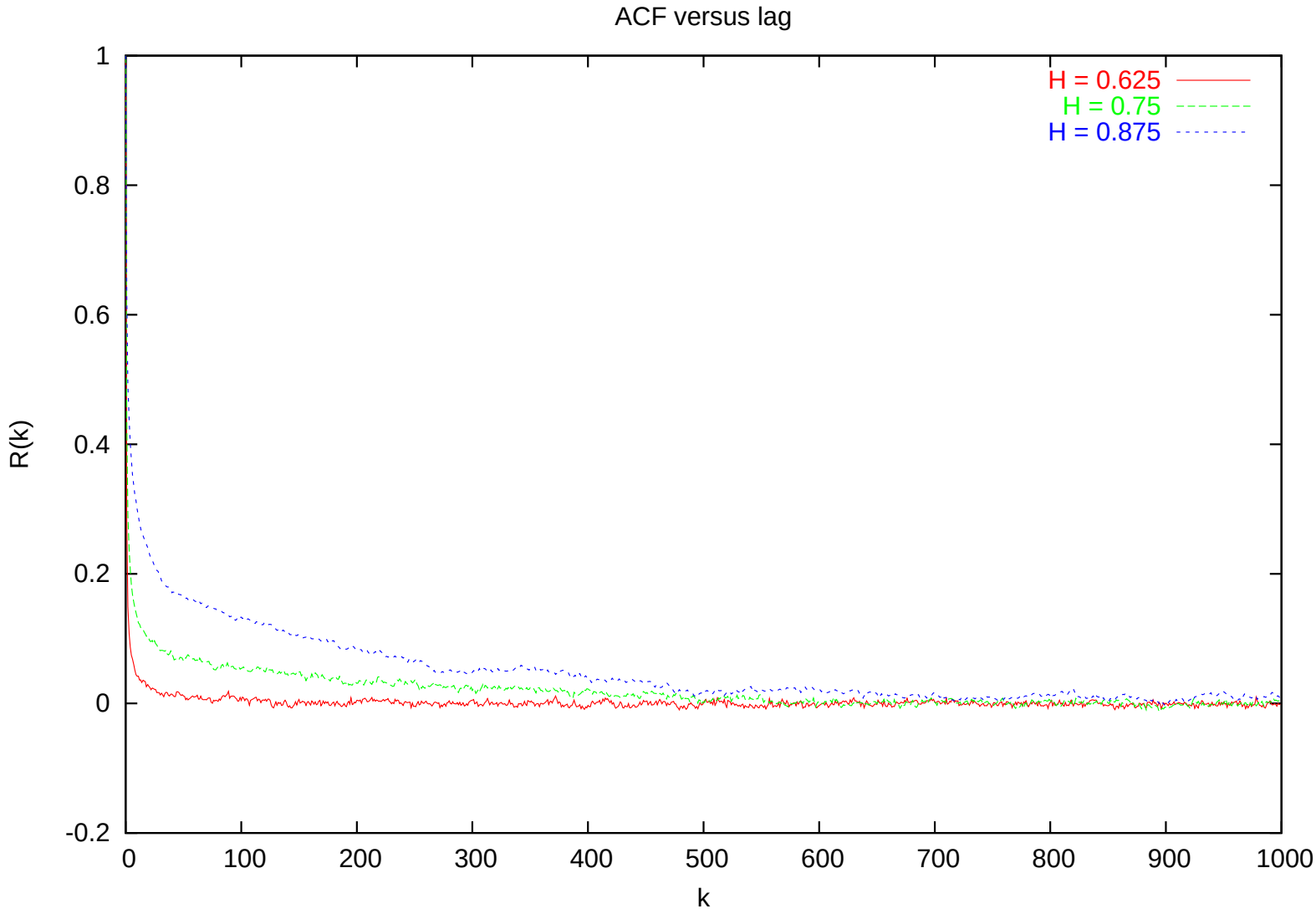
Algorithm for N state Finite Chain

1. If the state is not zero then reduce the state by one. This is the new state. Exit here.
2. Choose a new random number R in the range $[0, 1]$.
3. Set $j = 1$.
4. If $R < F(j, N)$ then the new state is $j - 1$. Exit here.
5. Increase j by 1. If $j > N$ the new state is N. Exit here.
6. Go to step 4.

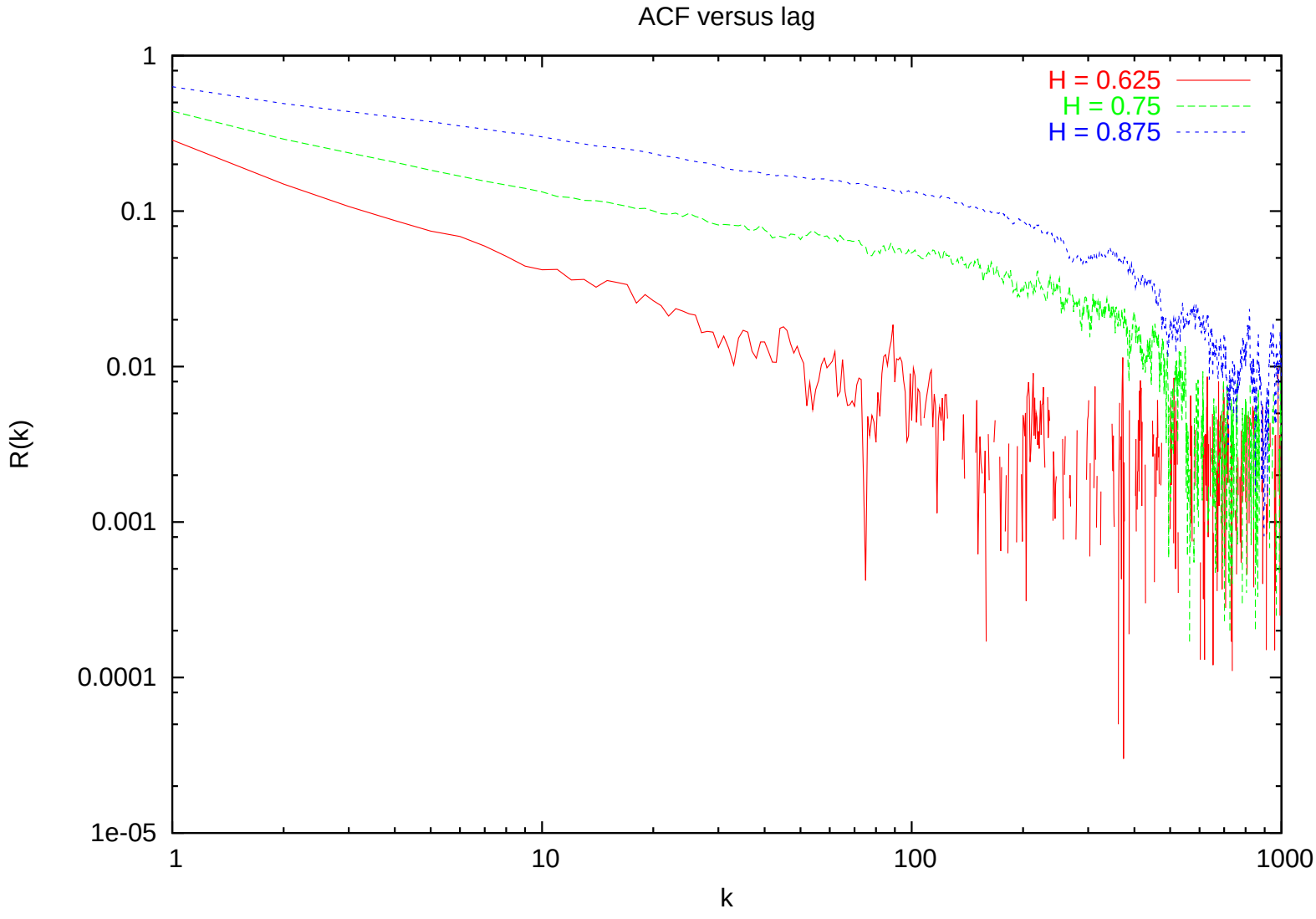
Algorithm for Infinite Chain

1. Explicitly calculate if $X \in [0, N-1]$ (where N is a small integer) using a single random no as previously.
2. Generate a new random number R in the range $[0, 1]$.
3. Calculate $Pr\{X \in [N, 2N-1] | X \in [N, \infty]\}$ if R is less than or equal to this probability then X is in the required range.
4. If X is in the required range then refine down by generating a new random number and use a binary search until X is found.
5. Otherwise increase the value of N to $2N$ and go to step 2.

ACF of process



Logscale ACF of process



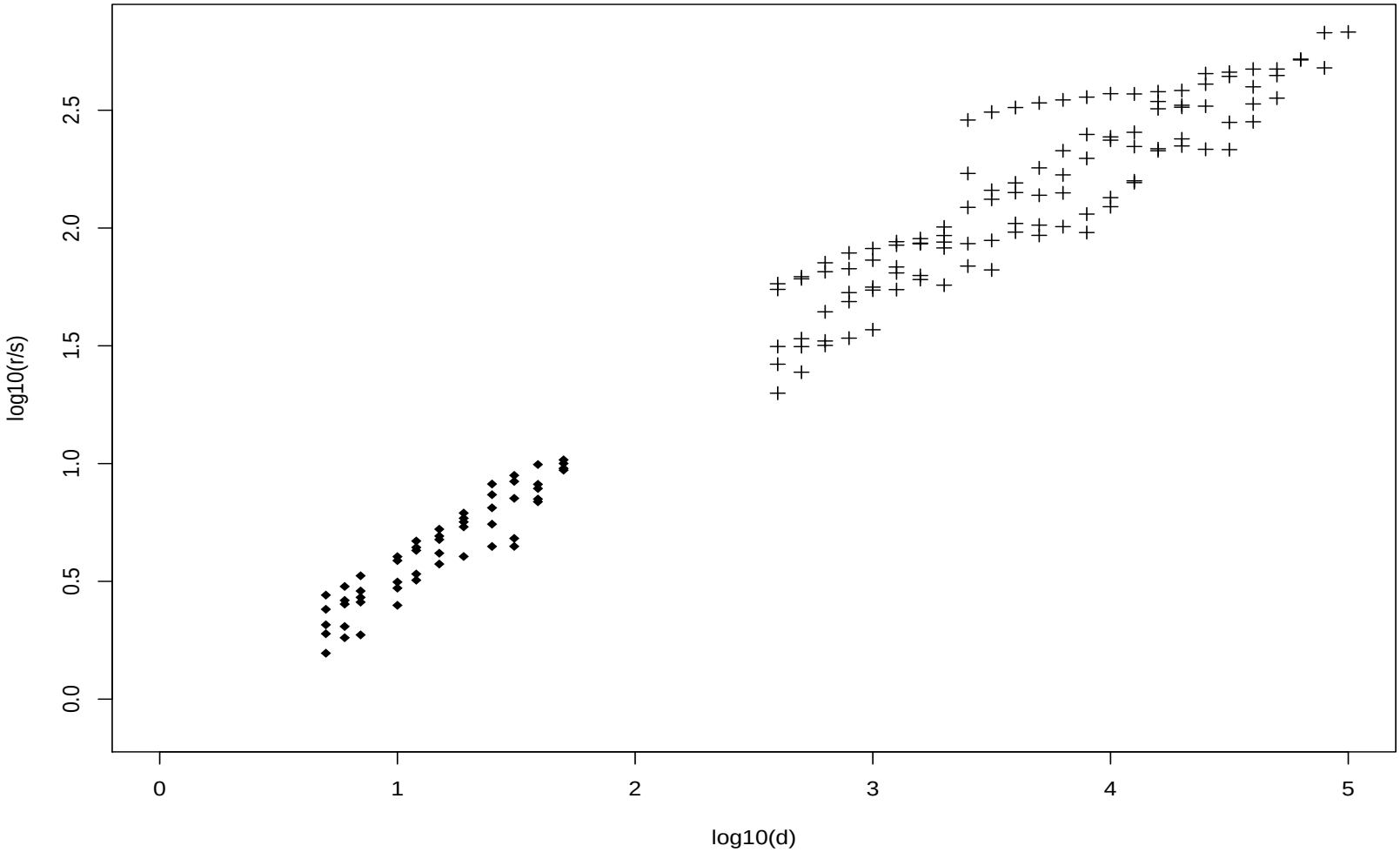
Methods for Estimating LRD

- R/S plot — the rescaled adjusted range. The oldest method for measuring H.
- Aggregated Variance — logarithm of variance versus logarithm of aggregation level.
- Periodogram — log periodogram (estimate of spectral density) versus frequency.
- Whittle's Estimator — an approximate MLE.
- Local Whittle — semi-parametric approximate MLE (parametric at frequencies near the origin).
- Wavelet based — frequency domain technique.

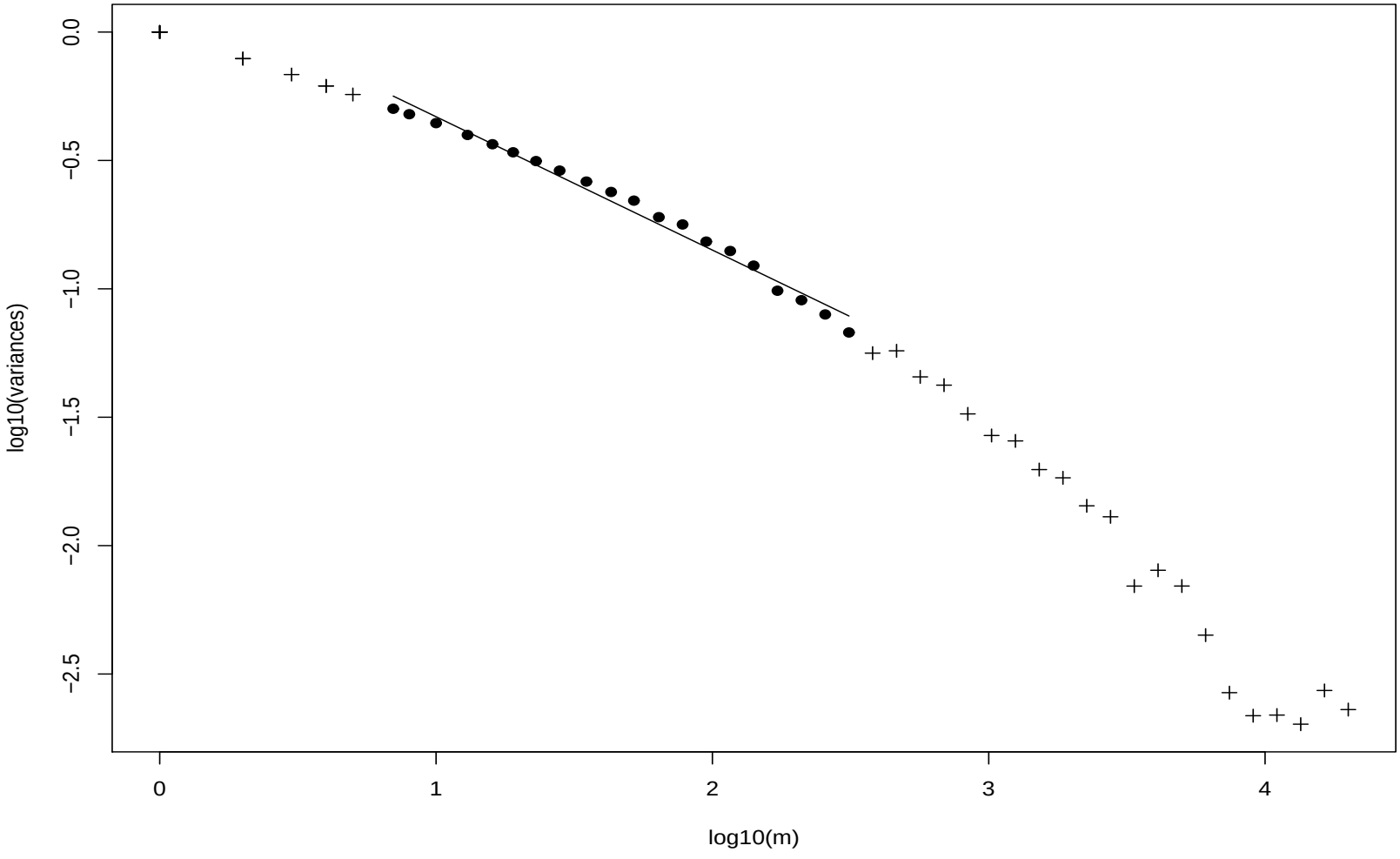
LRD Estimation Problems

- 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

R/S example — H=6.25



Variance example — H=6.25



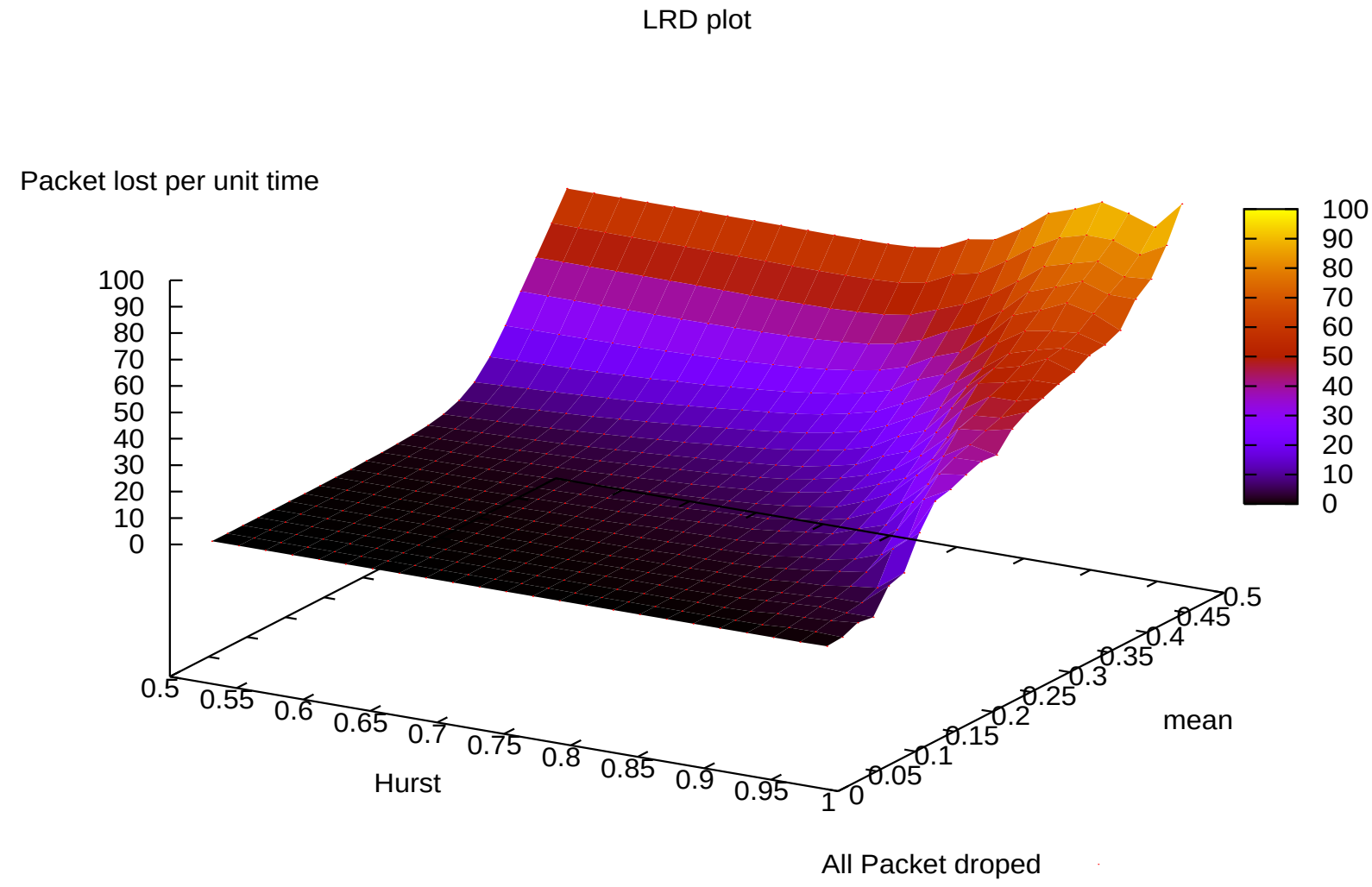
Results from Various Estimators

Data	Actual H	R/S	Var	Whit.	L.W.
FGN	0.625	0.62	0.63	0.61	0.63
FGN	0.75	0.71	0.73	0.74	0.77
FGN	0.875	0.80	0.81	0.86	0.90
Markov	0.625	0.64	0.58	0.63	0.69
Markov	0.75	0.64	0.70	0.76	0.80
Markov	0.875	0.73	0.74	0.84	0.88

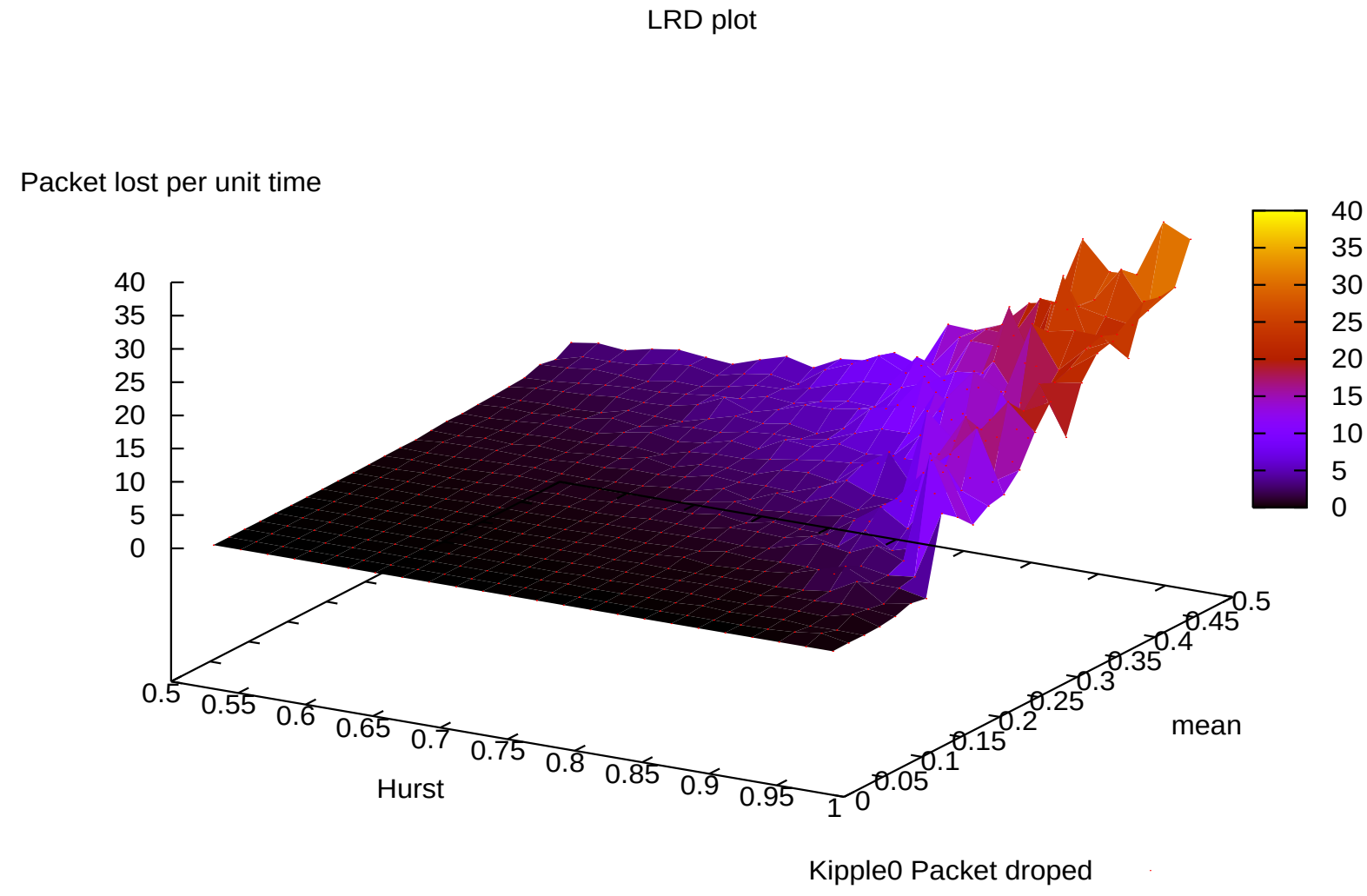
Simulation Results using ns

- ns simulator http://www.isi.edu/nsnam/ns/
- The Markov sources have been added to this simulation.
- The setup is four sources feeding gradually down to a single line.
- The mean and Hurst parameter have been varied to produce surface plots.
- The network is set so a mean of 0.5 will overload it totally.

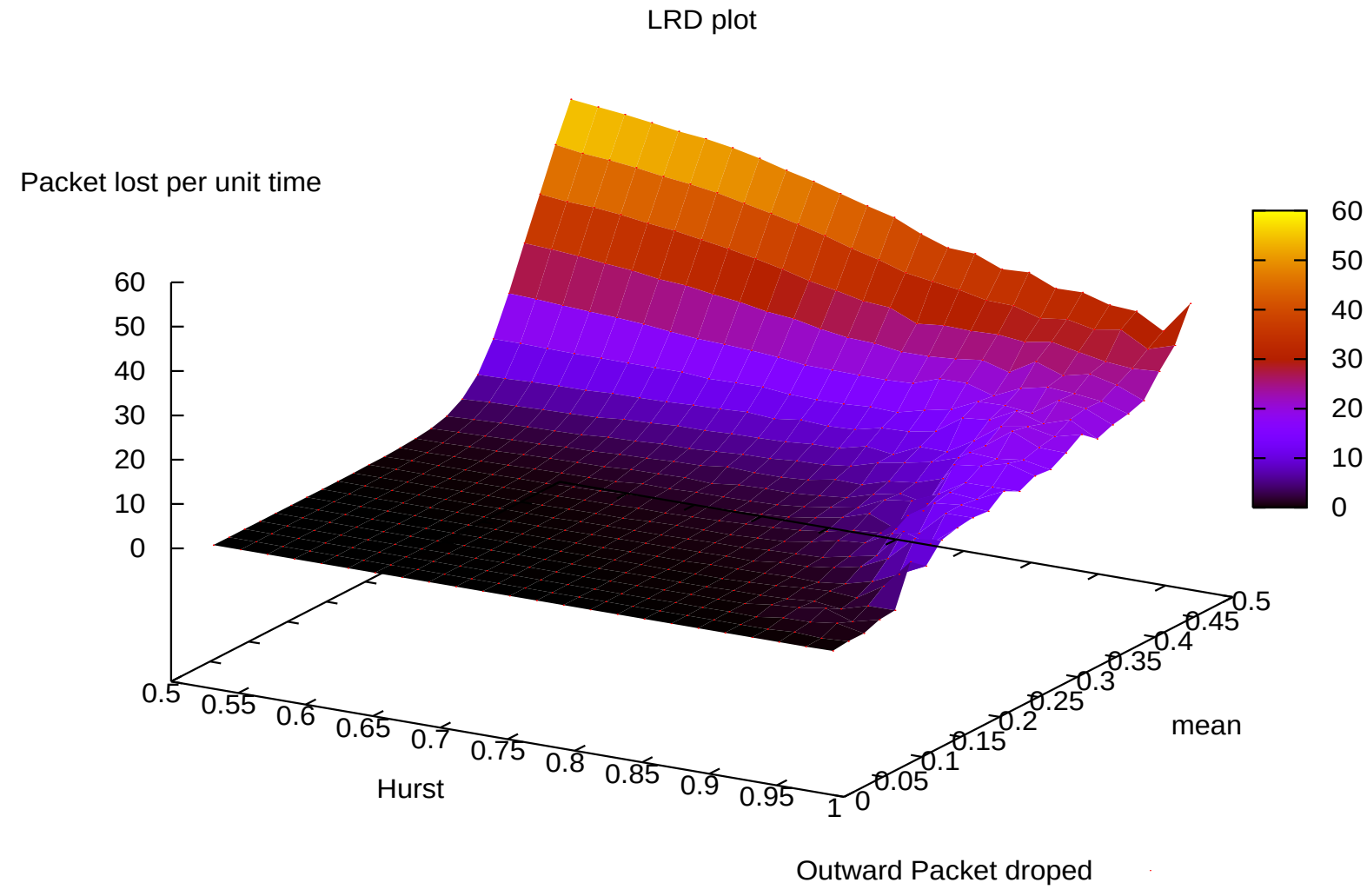
Simulated Packet Loss



Packet Loss at First Nodes



Packet Loss at Output



Tue Mar 04 13:57:05 2003

Moving to the Real Data

- Four causes of LRD in teletraffic data are suggested in the literature.
  1. Aggregation of heavy tailed sources produces LRD.
  2. TCP control feedback produces LRD.
  3. Traffic outbound from users is LRD at source.
  4. Aggregation of traffic in the network produces LRD.
- Real data can help identify the relative importance of these causes.

Some Real Data

- Data collected at incoming/outgoing pipe at University of York.
- 8.23 GB of data in 13.6 million packets — 67 minutes of data.
- 7.81 GB of this data is TCP. 0.6MB of data is ICMP. 0.4GB of data UDP.
- Outgoing data: 1.95GB of data in 6.0 million packets (av size: 323 bytes).
- Incoming data: 6.29GB of data in 7.7 million packets (av size: 821 bytes).

Disaggregating the data

- In addition to inbound and outbound we can break up traffic by port number.
- Ports are usually associated with particular services.
- Port 80 HTTP (5.78GB)
- Port 25 SMTP (226MB)
- Port 21 and 20 FTP (230MB)
- Port 53 DNS (33MB)

Results from Various Estimators

Data	R/S	Var	Whit.	L.W.
Total	0.75	0.88	0.97	0.98
In	0.73	0.89	1.32	0.97
Out	0.75	0.67	Error	1.00
http	0.80	0.89	1.33	0.98
ftp	0.83	0.93	0.93	0.99
smtp	0.72	0.68	0.72	1.02

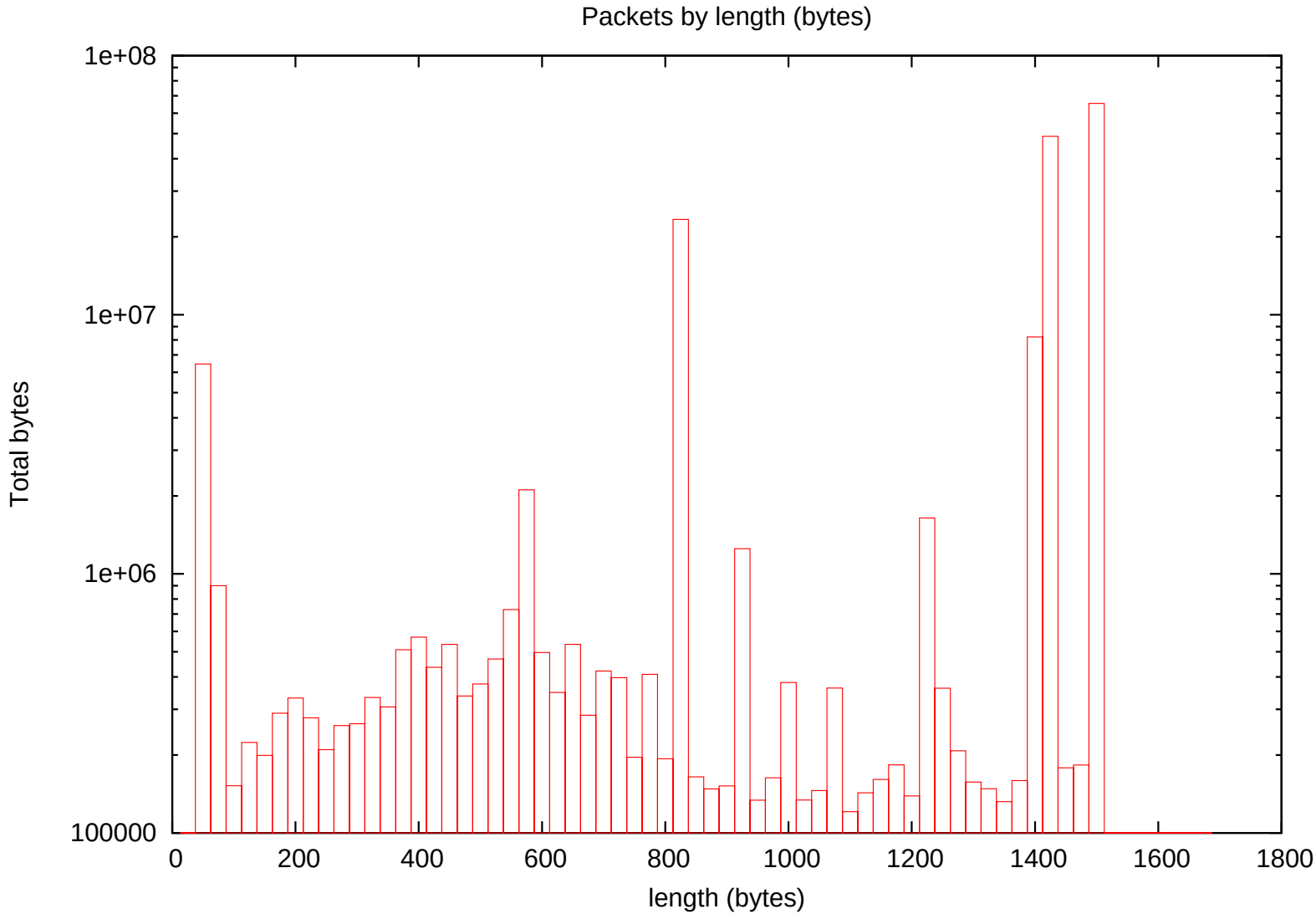
0.1 sec results

Data	R/S	Var	Whit.	L.W.
Total	0.82	0.92	Error	0.88
In	0.83	0.93	0.95	0.87
Out	0.75	0.78	0.97	0.88
http	0.79	0.93	0.96	0.89
ftp	0.67	0.93	Error	Error
smtp	0.78	0.75	0.81	1.06

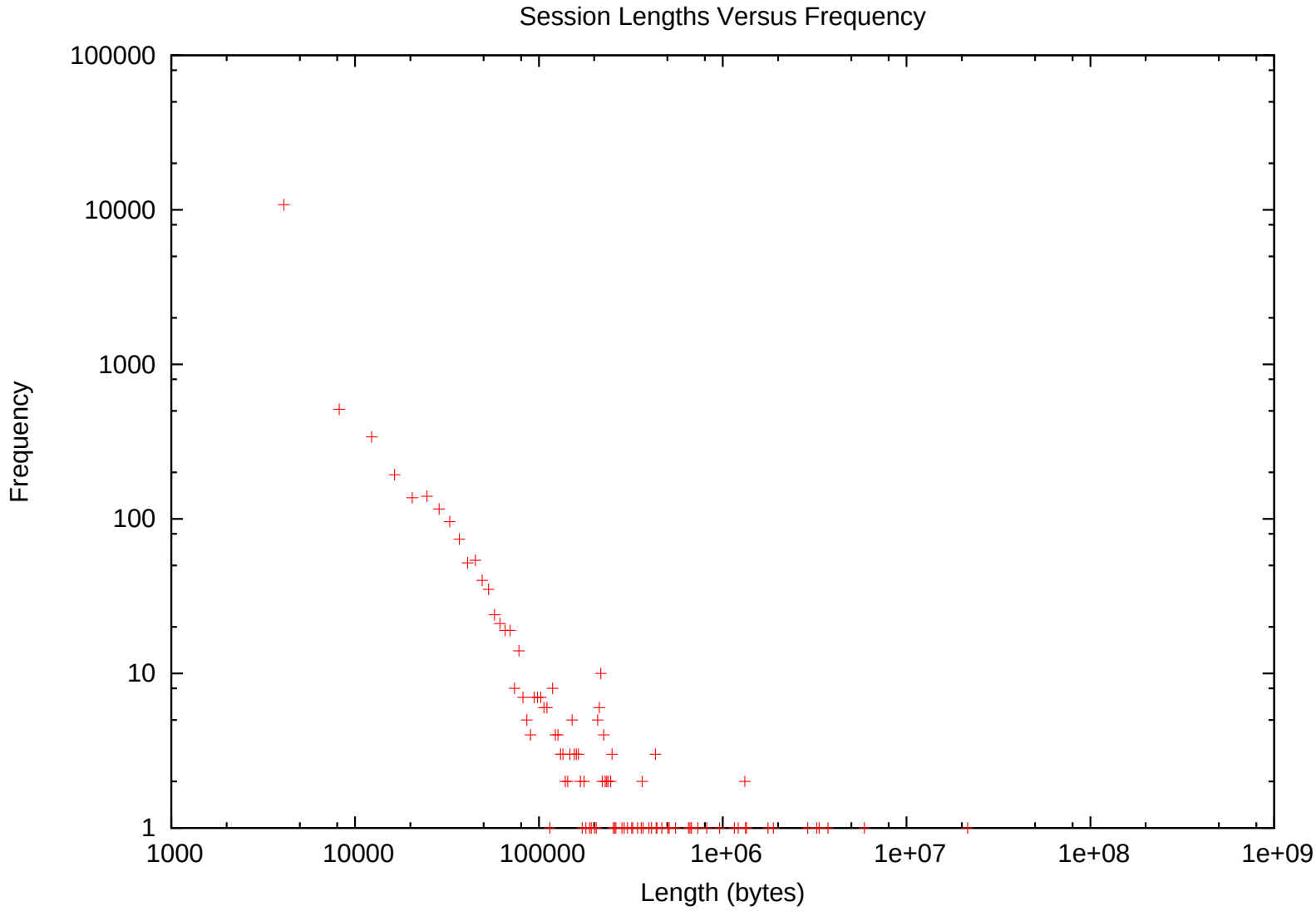
Smaller sample Results

Data	R/S	Var	Whit.	L.W.
Total	0.79	0.89	0.77	0.88
In	0.78	0.90	0.66	0.89
Out	0.66	0.60	0.60	0.78
http	0.78	0.90	0.79	0.88
ftp	0.19 ?	0.62	0.59	0.96
smtp	0.61	0.67	0.53	0.61

TCP Packet Lengths (bytes)



TCP Session Lengths (bytes)



Conclusions

- LRD is an important topic for study in telecomms networks.
- A new model has been presented which introduces LRD in a simple way.
- Modelling with this LRD generating source shows it has a considerable effect on queuing performance.
- Measuring LRD in real data is extremely difficult.
- Four well known methods produce inconsistent results.
- More work remains to be done on this project.

For More Information

- The Networks and Nonlinear Dynamics Group (NNDG) at York is new to the field of telecomms research (only four years of experience).
- We welcome collaboration with other groups and the benefit of their research experience.
- Email richard@manor.york.ac.uk
- Web http://gridlock.york.ac.uk
- This work was undertaken with the help of our colleagues at Queen Mary, University of London, BTexact and Nortel Networks.