Internet Traffic Variability
(Long Range Dependency Effects)

Dheeraj Reddy

CS8803 Fall 2003

Self-similarity and its evolution in Computer Network Measurements

- Prior models used ‘Poisson-like’ models
  - Origins in telecommunication voice traffic
  - Smoother aggregate traffic.
  - Simpler and traditional (tried and tested) congestion management (OR principles)
- Self-Similarity – An orthogonal approach
  - The Ethernet measurements at Bellcore and their subsequent statistical analysis.

Prior Models

- Poisson processes or Markov-modulated Poisson processes
- Multi-state Markov models
  - Voice traffic (two state Markov model)
    - Silence and Talking states
  - Video traffic (multi-state Markov model)
    - Variable bit-rate video traffic (I, B and P frames)
- IDC, Peak-to-Mean Ratio, coefficient of variation are inadequate to illustrate the burstiness

On the Self-Similar Nature of Ethernet Traffic

Leland, Taqqu, Willinger, Wilson. IEEE/ACM ToN, Vol. 2, pp 1-15, 1994

- Establish self-similar nature of Ethernet traffic
- Illustrate the differences between self-similar and standard models
- Serious implications of self-similar traffic for design, control and performance analysis of packet-based communication systems

Self similarity

- Definition
  - If X be a covariance stationary stochastic process with μ and σ^2 and auto-correlation function *r*

    $$X = \left(X_t : t = 0,1,2\ldots\right) \qquad r(k) \sim k^{-\beta} L_1(k), k \to \infty$$

  - For each m=1,2,3…. If X^m denotes a new time series obtained by averaging the original series over non-overlapping blocks of m

    $$X_k^{(m)} = 1/m(X_{km-m+1} + \ldots\ldots + X_{km})$$

  - Process X is self similar with self-similarity parameter H = 1-β/2 if

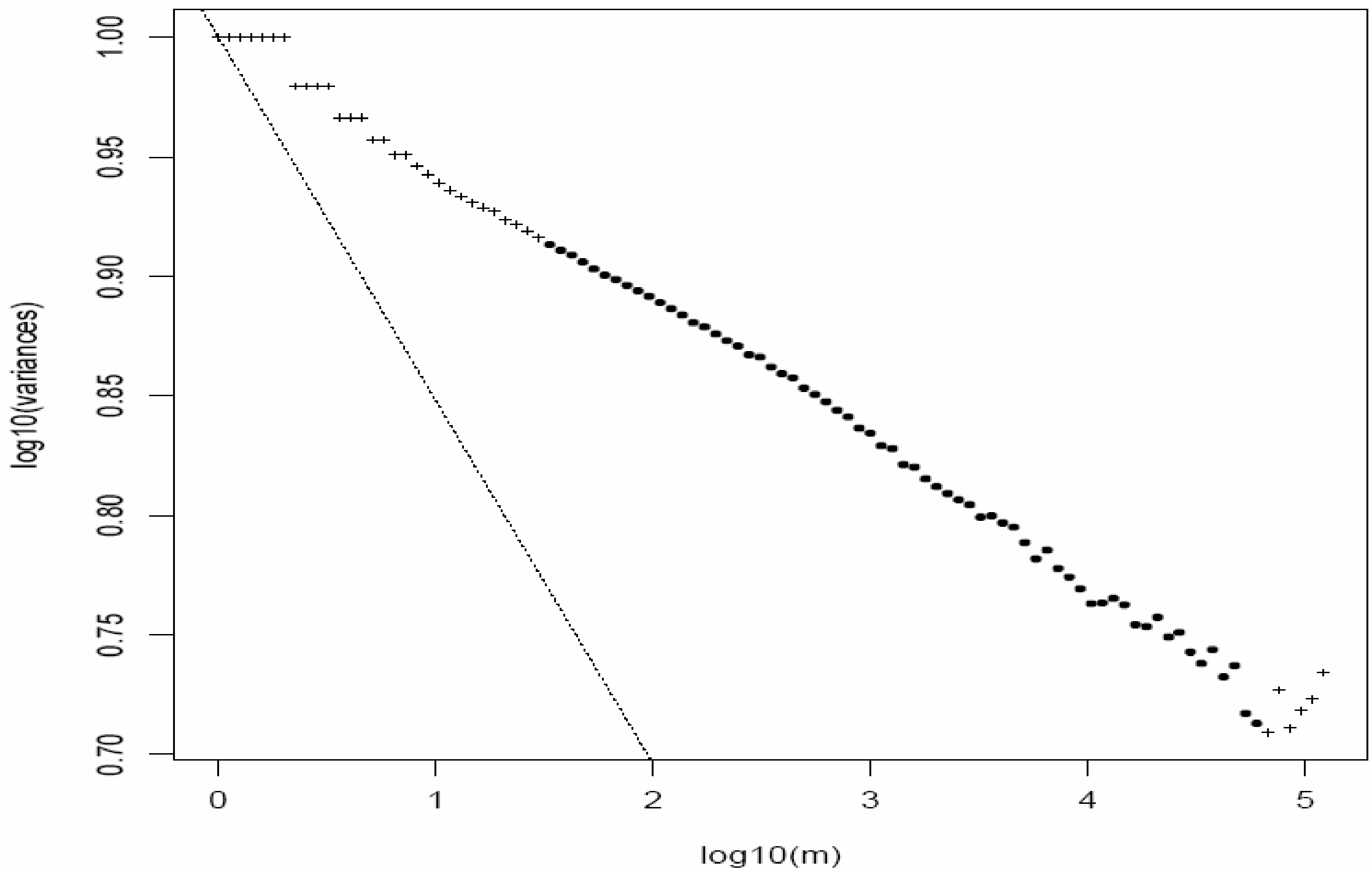
    $$r^{(m)}(k) = r(k), m = 1,2,3\ldots$$

Traces of Ethernet Traffic Measurements					
Measurement Period		data set	total number of bytes	total number of packets	Ethernet utilization
AUGUST 1989	total (27.45 hours)		11,448,753,134	27,901,984	9.3%
Start of trace: Aug. 29, 11:25am	low hour (6:25am-7:25am)	AUG89.LB AUG89.LP	224,315,439	652,909	5.0%
End of trace: Aug. 30, 3:10pm	normal hour (2:25pm-3:25pm)	AUG89.MB AUG89.MP	380,889,404	968,631	8.5%
	busy hour (4:25pm-5:25pm)	AUG89.HB AUG89.HP	677,715,381	1,404,444	15.1%
OCTOBER 1989	total (20.86 hours)		14,774,694,236	27,915,376	15.7%
Start of trace: Oct. 5, 11:00am	low hour (2:00am-3:00am)	OCT89.LB OCT89.LP	468,355,006	978,911	10.4%
End of trace: Oct. 6, 7:51am	normal hour (5:00pm-6:00pm)	OCT89.MB OCT89.MP	827,287,174	1,359,656	18.4%
	busy hour (11:00am-12:00am)	OCT89.HB OCT89.HP	1,382,483,551	2,141,245	30.7%
JANUARY 1990	total (40.16 hours)		7,122,417,589	27,954,961	3.9%
Start of trace: Jan. 10, 6:07am	low hour (Jan. 11, 8:32pm-9:32pm)	JAN90.LB JAN90.LP	87,299,639	310,038	1.9%
End of trace: Jan. 11, 10:17pm	normal hour (Jan. 10, 9:32am-10:32am)	JAN90.MB JAN90.MP	182,636,845	643,451	4.1%
	busy hour (Jan. 11, 10:32am-11:32am)	JAN90.HB JAN90.HP	711,529,370	1,391,718	15.8%
FEBRUARY 1992	total (47.91 hours)		6,585,355,731	27,674,814	3.1%
Start of trace: Feb. 18, 5:22am	low hour (Feb. 20, 1:21am-2:21am)	FEB92.LB FEB92.LP	56,811,435	231,823	1.3%
End of trace: Feb. 20, 5:16am	normal hour (Feb. 18, 8:21pm-9:21pm)	FEB92.MB FEB92.MP	154,626,159	524,458	3.4%
	busy hour (Feb. 18, 11:21am-12:21am)	FEB92.HB FEB92.HP	225,066,741	947,662	5.0%

Table 1. Qualitative description of the sets of Ethernet traffic measurements used in the analysis in Section 4.2.

Statistical tests for Self-Similarity

- Variance-time plots
  - Analysis of the variances of the aggregated processes
- R/S Statistic
  - Time-domain analysis of the data
- Periodogram Analysis
  - Frequency domain analysis (~ FFT)



Variance-time plot of sequence AUG89.MB. The asymptotic slope is clearly larger than the slope -1.0 of the dotted reference line and is readily estimated to be about -0.40. H=0.80

Significance of self-similarity

- Nature of traffic generated by individual Ethernet users. Aggregate traffic study provides insights into traffic generated by individual users.
- Commonly used measures of “burstiness” like IDC, peak-to-mean ratio etc. are not meaningful for self-similar traffic and can be replaced by Hurst parameter.
- Nature of congestion produced by self-similar models differs drastically from that predicted by standard formal models

Why is Ethernet traffic self-similar ?

- Plausible physical explanation of self-similarity in Ethernet traffic
- Convergence results for processes that exhibit high variability (i.e., infinite variance)

Willinger, Taqqu, Sherman and Wilson: Self similarity through high variability: Statistical Analysis of Ethernet LAN traffic at Source Level
ACM SIGCOMM 1995

Mathematical Result

- Superposition of many ON/OFF sources whose ON-periods and OFF-periods exhibit the Noah-effect (i.e., have high variability or infinite variance) produces aggregate network traffic that features the Joseph effect (i.e., is self-similar or long-range dependent).
  - Terminology attributed to Mandelbrot

Idealized ON/OFF Model

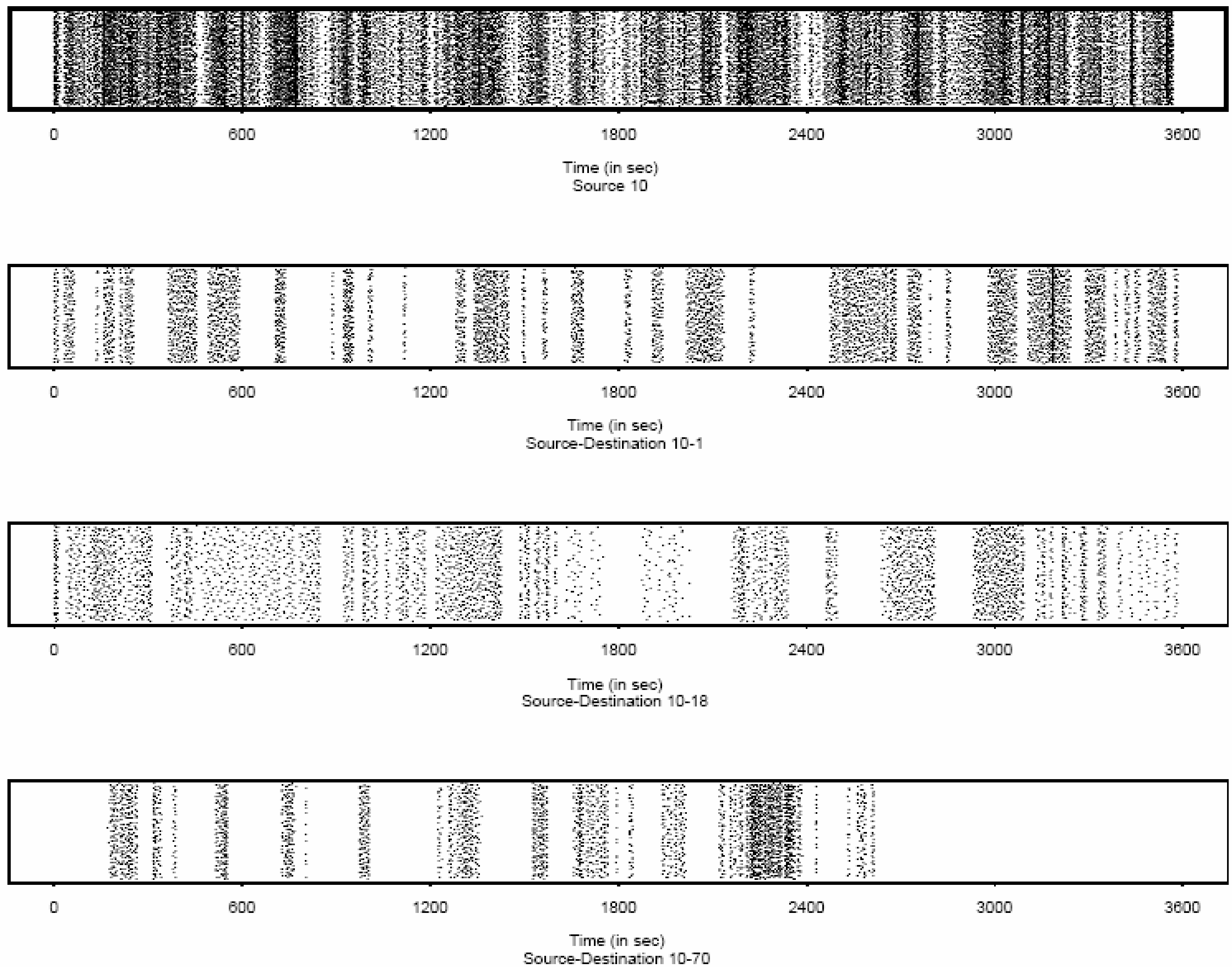
- **Theorem 1**. For large enough source Number M and Block aggregation size b, the cumulative load $\{W^*_{M,b}(j),\ j \geq 0\}$ behaves statistically as

$$\frac{1}{2}bM + b^H M^{1/2} G_{H,\sigma}(j)$$

where $H = \dfrac{3-\alpha}{2}$ and $\sigma^2 = \dfrac{1}{4E(U)2(\alpha-1)(2-\alpha)(3-\alpha)}$

Measurements and Analysis

- Two sets of Ethernet measurements from the Bellcore measurements. (89 & 94)
- Unlike previous studies, data has been classified according to source-destination pairs by looking at headers to verify
  - The ON/OFF traffic model assumption
  - The Noah Effect for the corresponding ON and OFF periods



Textured plots of a source-cumulative and individual source-destination pairs

Checking for the Noah Effect

- Complementary distribution plots

$$\log(P(U > u)) \sim \log(c) - \alpha \log(u), \text{as } u \to \infty$$

- Hill's estimate
  - Let U_1, U_2,…, U_n denote the observed ON-(or OFF-)periods and write $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$ for the corresponding order statistics

$$\hat{\alpha}_n = \left(\frac{1}{k} \sum_{i=0}^{i=k-1} (\log U_{(n-1)} - \log U_{(n-k)}) \right)^{-1}$$

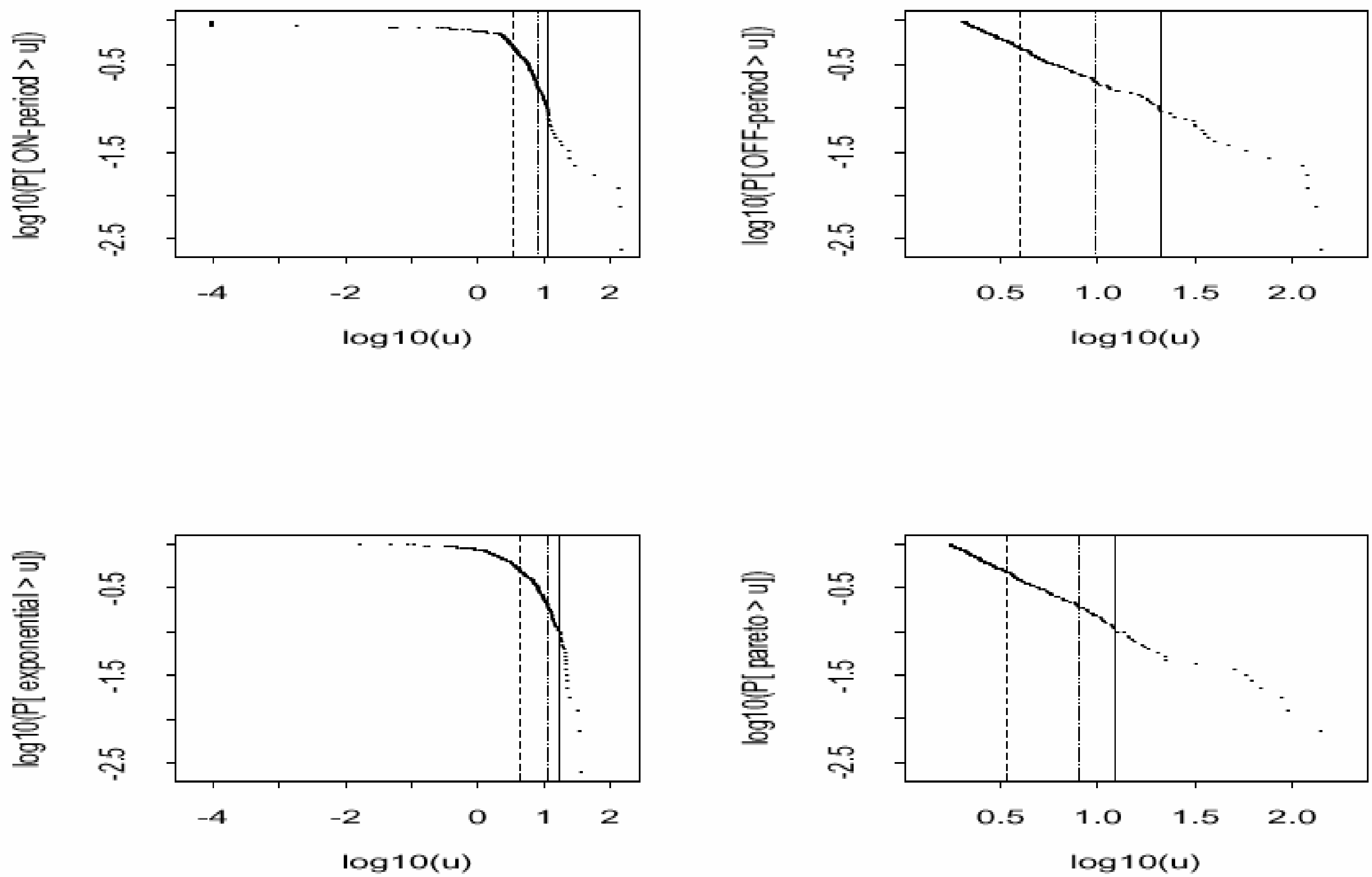


Fig. 2. Complementary distribution plots for *ON-periods* (top left) and *OFF-periods* (top right) for the source-destination pair 10-18, using a threshold value of $t = 2s$; for a sample from an *exponential distribution* that matches the ON-periods (lower left), and for a sample from a *Pareto distribution* that matches the OFF-periods (lower right). (The vertical solid, dotted and dashed lines indicate that 10%, 20% and 50% of all data points are to the right of the respective lines.)

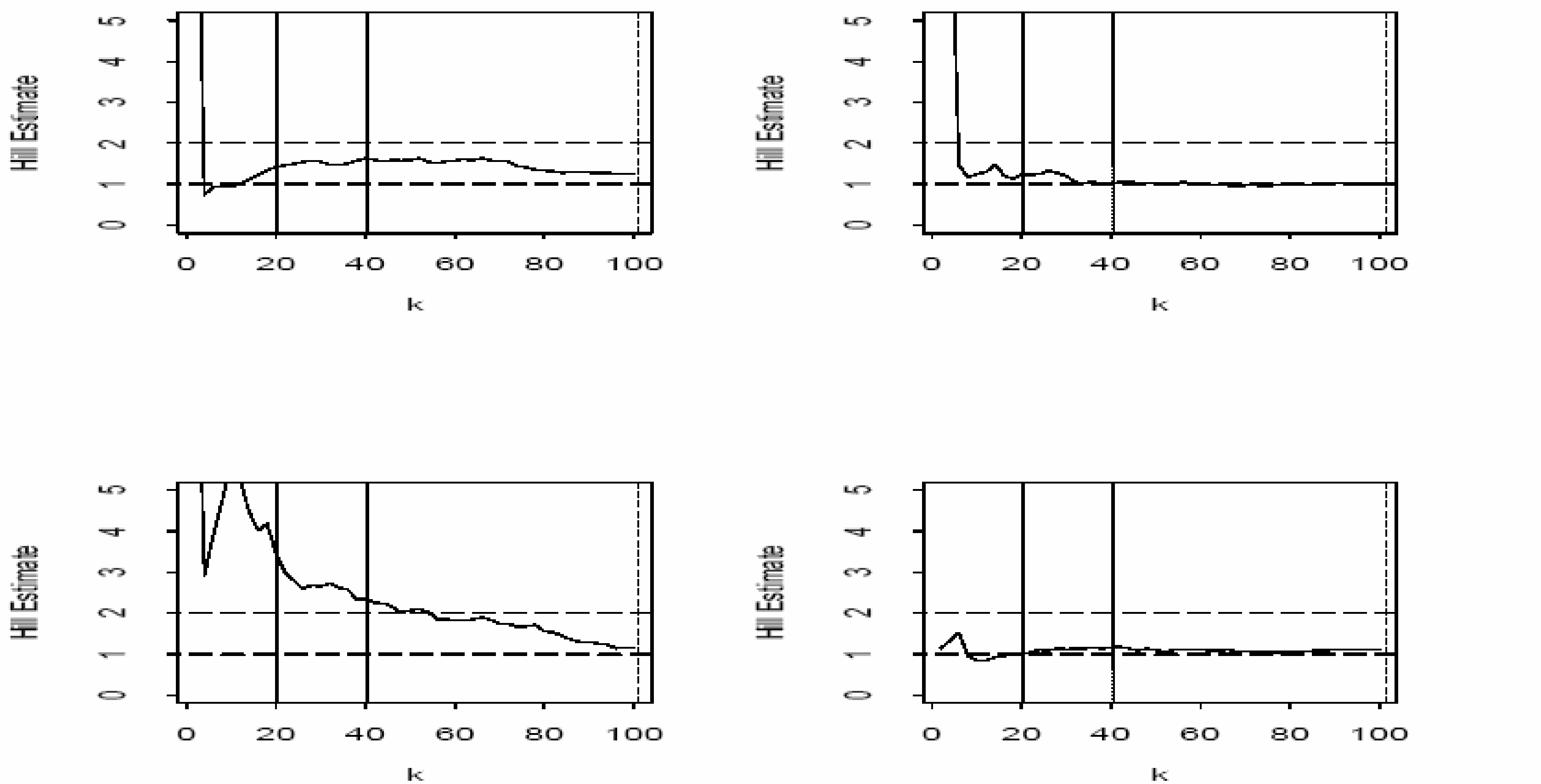


Fig. 3. Hill estimate plots for *ON-periods* (top left) and *OFF-periods* (top right) for the source-destination pair 10-18, using a threshold value of $t = 2s$; for a sample from an *exponential distribution* that matches the ON-periods (lower left), and for a sample from a *Pareto distribution* that matches the OFF-periods (lower right). (The vertical solid, dotted and dashed lines indicate that 10%, 20% and 50% of the largest order statistics have been included in the Hill estimation calculation.)

Robustness of the Noah Effect

- As far as the Noah effect is concerned it doesn't matter how the OFF-periods or the inter-train distances (or for that matter ON-periods or the train lengths) are defined. (choice of threshold t)
- Why??
  - Distributions that satisfy the hyperbolic tail condition are scalable.

    for sufficiently large u,t and u > t

$$P(U > u \mid U > t) \sim \left(\frac{u}{t}\right)^{-\alpha}, \; 1 < \alpha < 2$$

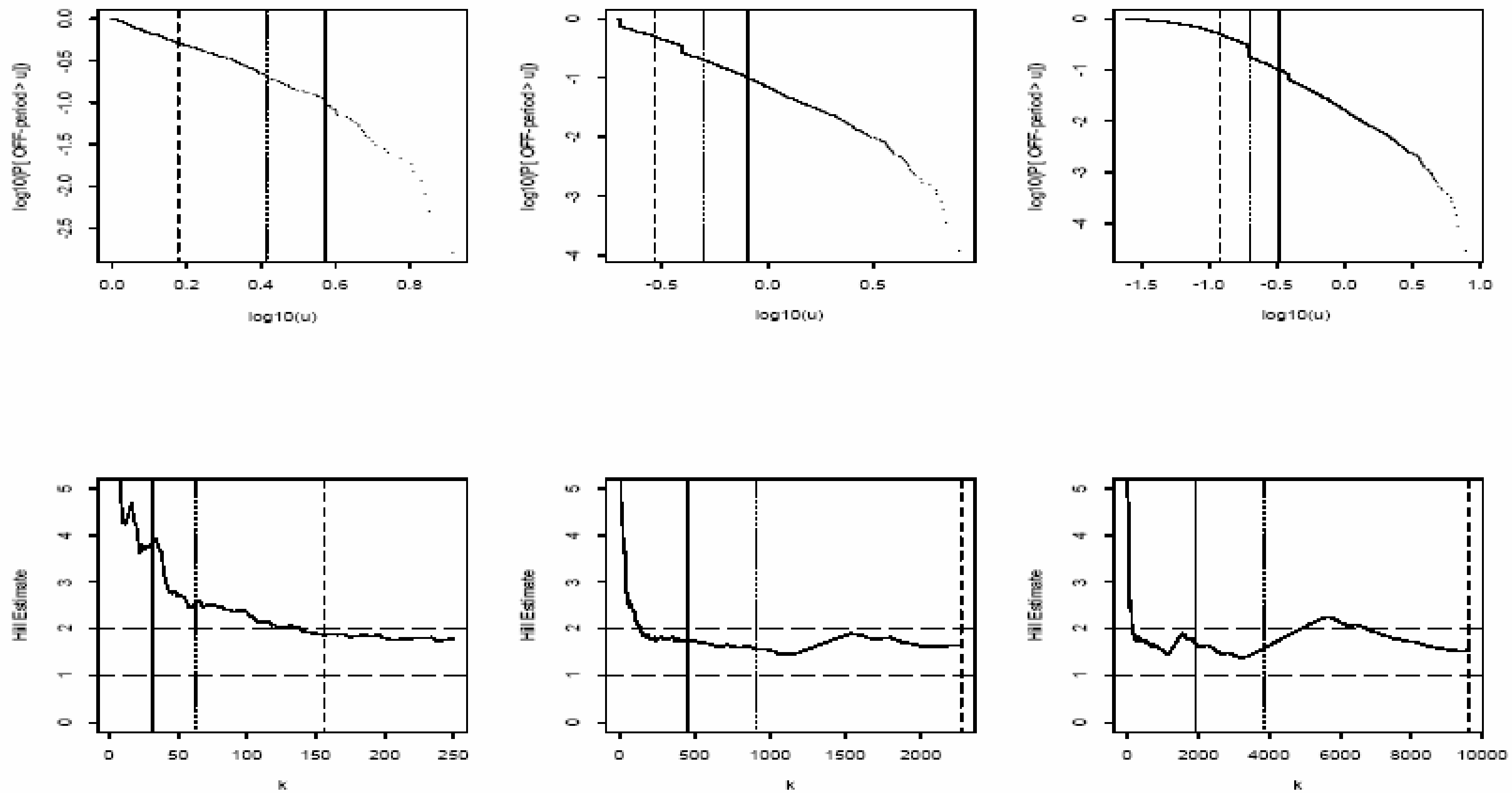
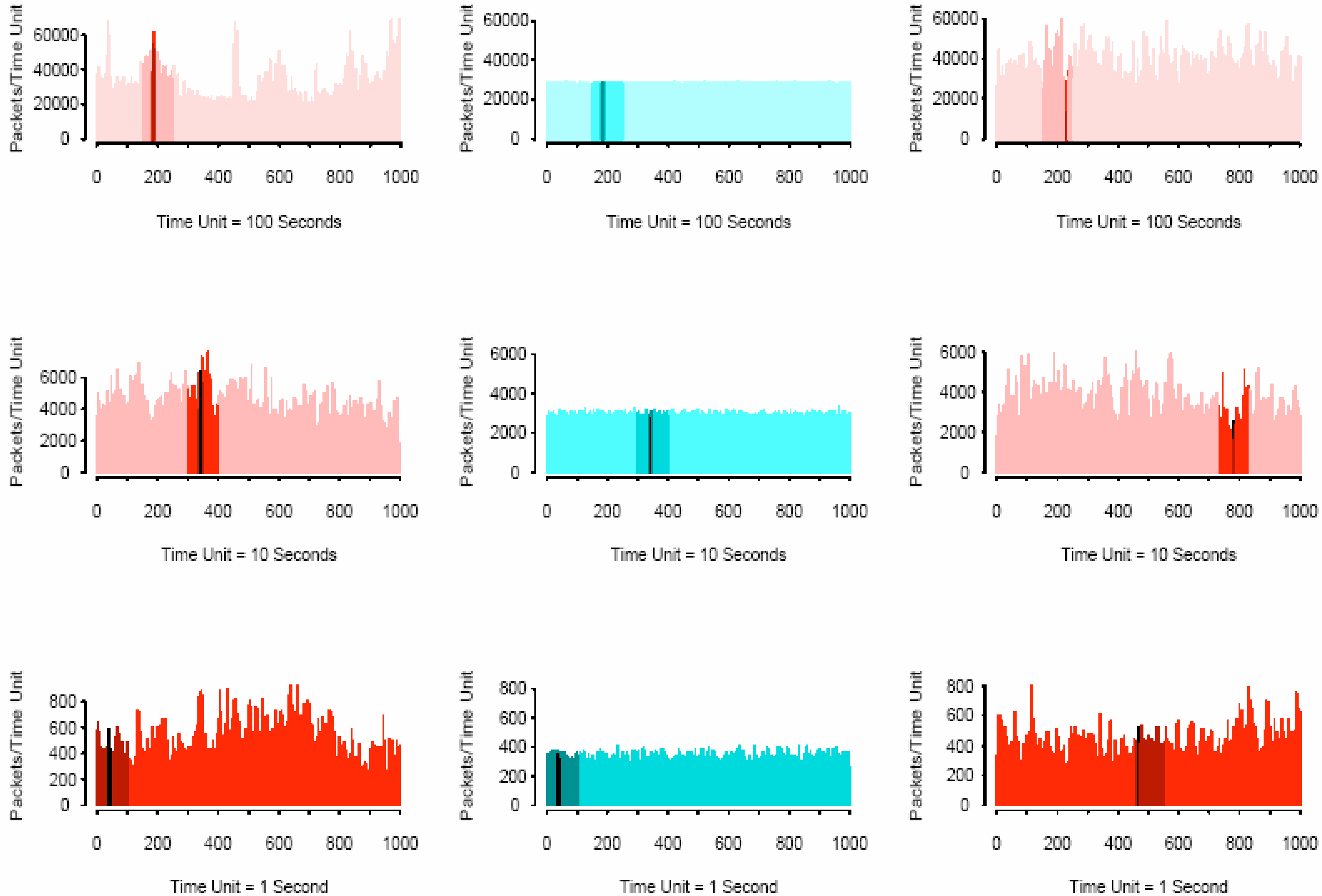


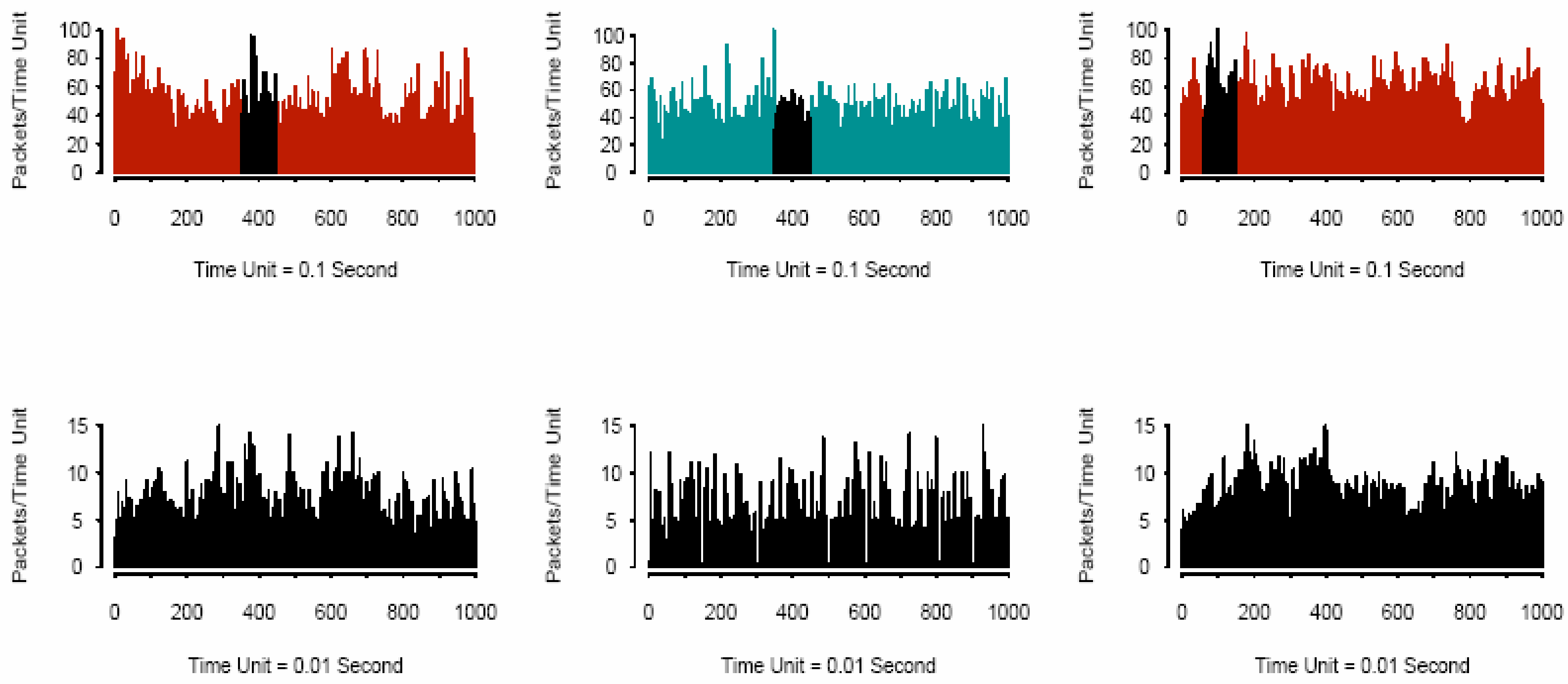
Fig. 4. An illustration of the robustness property of the Noah Effect for the OFF-periods (using source 10). For threshold values $t = 1.0s, 0.20s$ and $0.025s$, the top row gives the complementary distribution plots and the bottom row the corresponding Hill plots.

Implications of Noah Effect

- Traffic Modeling and Synthetic traffic generation
  - Parsimonious modeling is still possible despite the complexity of network traffic since a single parameter needs to be estimated.
- Performance and Protocol Analysis
  - Fewer meaningful parameters that need to be investigated



Column1: Actual trace; Column 2: Synthetic trace from appropriate traditional model; Column 3: synthetic trace from a self-similar model with one parameter.



Column1: Actual trace; Column 2: Synthetic trace from appropriate traditional model; Column 3: synthetic trace from a self-similar model with one parameter.

What is the impact of LRD on queuing in a packet network ?

- Queuing Performance
  - When incoming traffic is fractal in nature
- Conditions under which parsimonious traffic models are appropriate

A. Erramilli, O.Narayan and W. Willinger, “Experimental Queuing Analysis with Long Range Dependent Packet Traffic” IEEE/ACM Trans. Networking, vol. 4, no. 2, Apr 1996

Experimenting with Traces

- A single 30-min trace.
  - Variability of relevant traffic statistics is within confidence limits
  - Experiment with inter-arrival traces to preserve inter-arrival time distributions
- Queuing System Characteristics
  - Infinite waiting room, deterministic service time and single server

Experimenting with traces(2)

- Three sets of experiments
  - Original trace
  - QNA based approximations (two-moment approximation for the mean waiting time in a GI/G/1 queue)
  - Synthetic trace obtained by shuffling the time series of inter-arrival times
  - Vary the service time of the server to obtain different utilizations of the queue

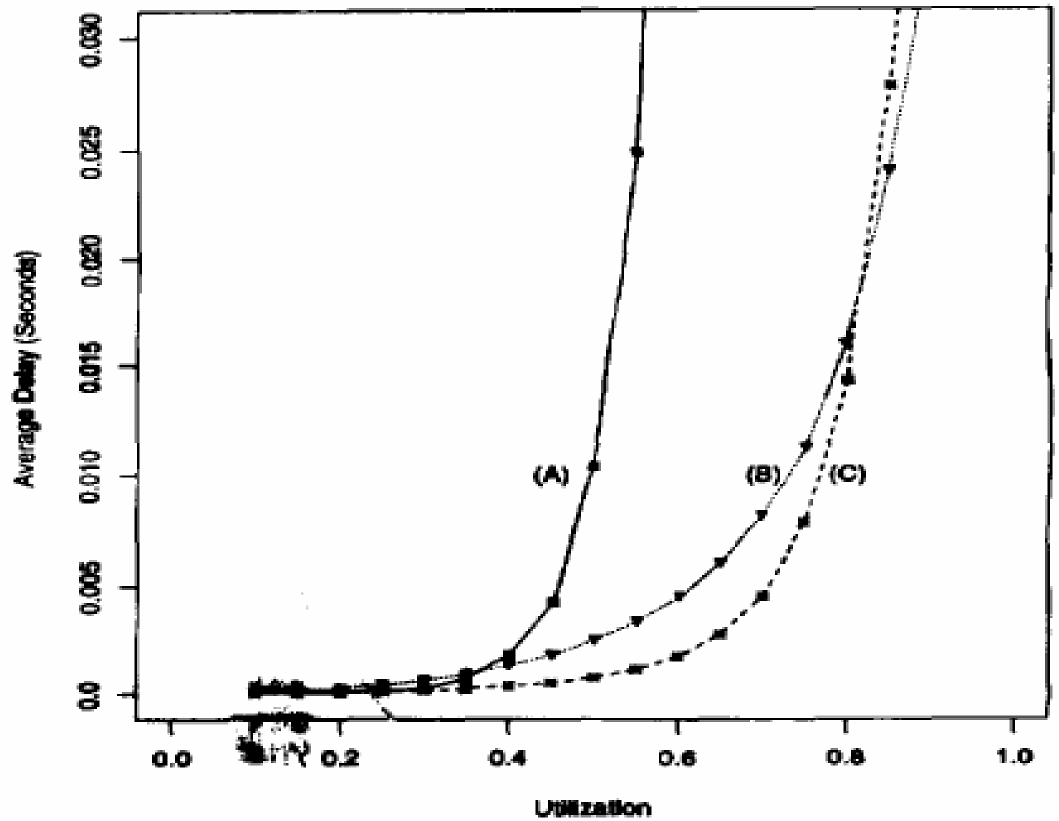


Fig. 2. Average delay (in seconds) versus utilization plot for original trace (A), QNA-based approximation (B), and fully shuffled trace (C).

Check the “knee of the curve”

Differences between A and C suggest that even the best renewal model will underestimate the delays at moderate to high utilizations

Two more experiments

- Divide the inter-arrival times Ethernet trace into blocks of size m
  - External shuffle
    - order of the blocks is shuffled
    - preserving the sequence within.
    - preserves short-range correlations
  - Internal shuffle
    - Sequence within each block is randomized
    - Order of blocks is preserved
    - Destroys short-range correlations

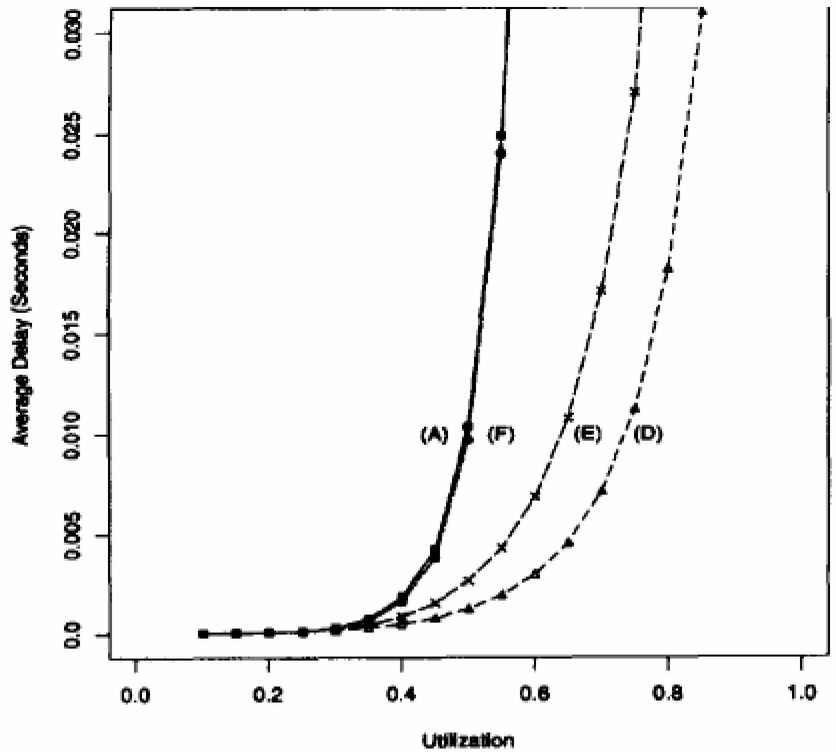


Fig. 4. Average delay (in seconds) versus utilization plot for original trace (A), trace with identical one-step correlations (D), externally shuffled trace (E), and internally shuffled trace (F).

m=25;
average block duration=76ms,
varying from 14-629ms;

- The internally shuffled trace is almost coincident with the original trace
- The LRD is not only important for queuing performance but is a dominant characteristic for determining several issues in traffic engineering

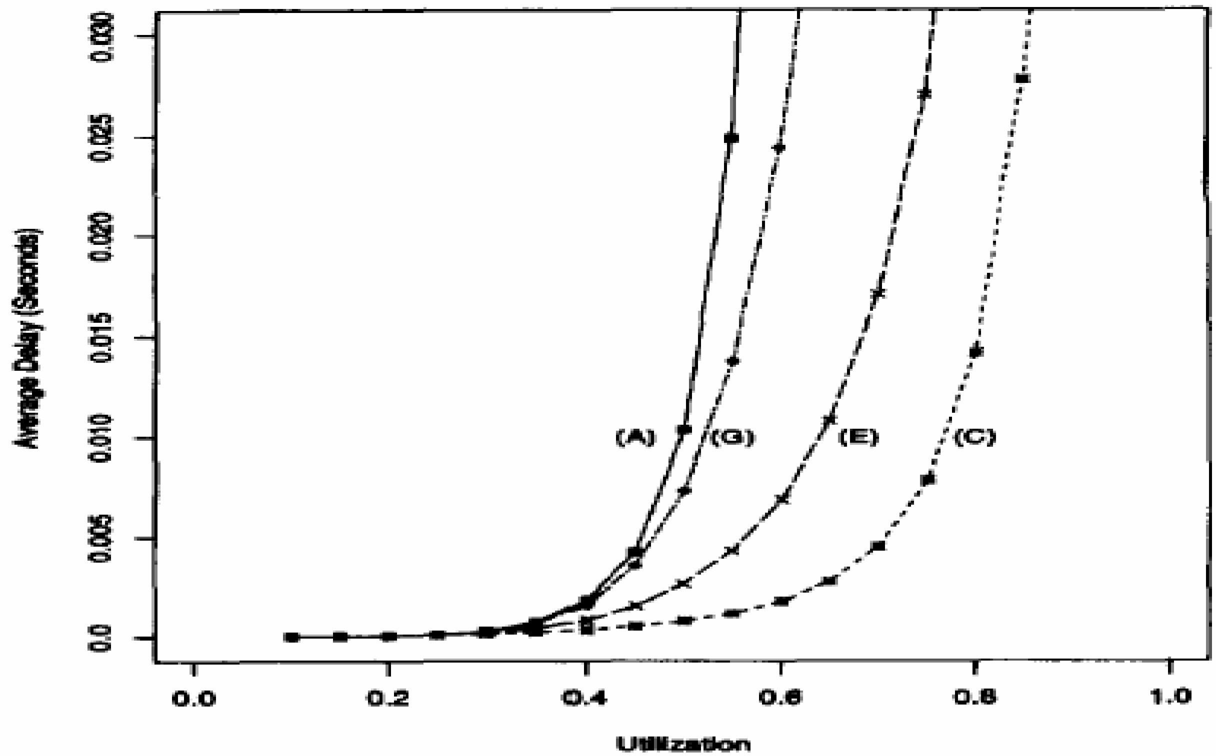


Fig. 5. Average delay (in seconds) versus utilization plot for original trace (A), fully shuffled (i.e., external shuffle with block size $m = 1$) trace (C), externally shuffled trace with block size $m = 25$ (E), and externally shuffled trace with block size $m = 500$ (G).

- Correlations over extremely long time scales in the data have measurable practical consequences
- A description in terms of arrival counts over a small time interval is adequate even though it won't include traffic characteristics below this scale

Observations

- Tails of queue length distributions obtained with actual data traces are heavier than indicated by exponential delay (due to LRD).
- Experiments with counts are more in tune with the past LRD studies. (similar results are expected from data sets with time series of counts in datasets)

Why does WWW traffic (subset of network traffic) looks self-similar ?

Mark E, Crovella, and Azer Bestavros, “Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes” IEEE/ACM Trans. Networking, vol. 5, no. 6, Dec 1997

- Different from the earlier work which decomposed the whole traffic as generated from different sources

How is this different?

- Since the focus is on only web traffic, the busiest four hours are taken
- Less busy hours do not show self similar characteristics
  - Possibly because the traffic demand is too less in the logs of data collected

Data Collection

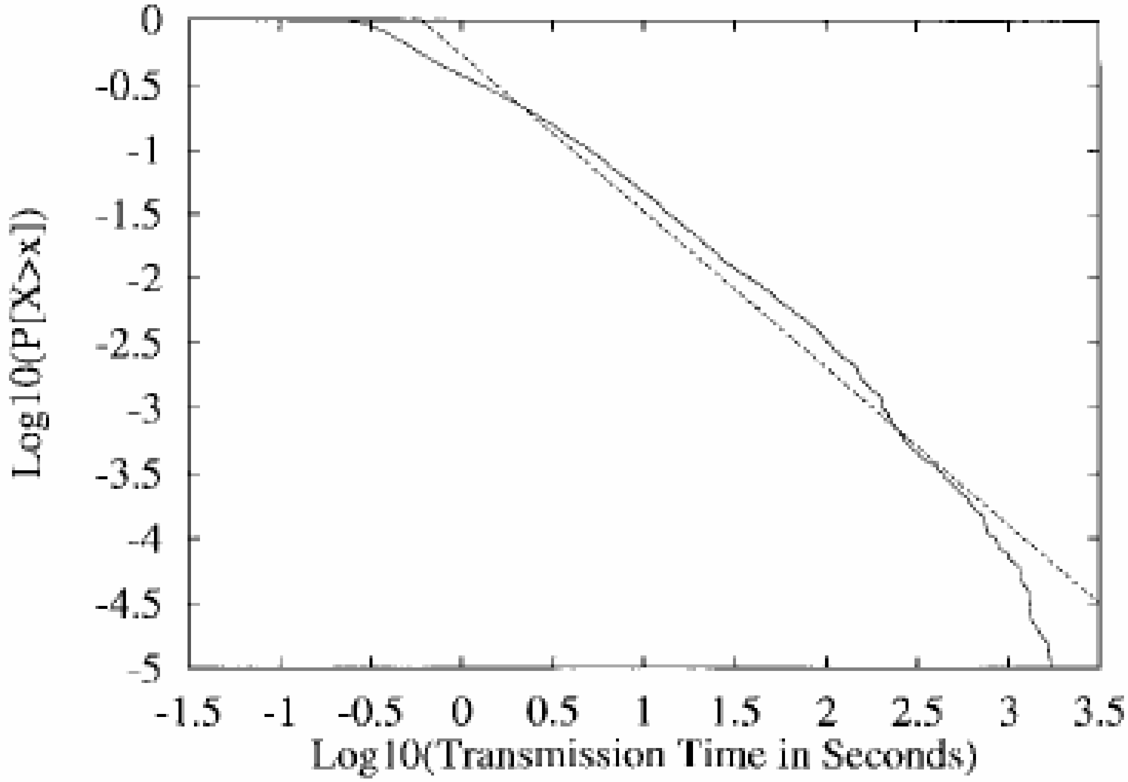
- Modified NCSA Mosaic browser
- URL, Session, User, Machine, Request Time, Document size, Object Retrieval time
- Convert logs to traffic time series, bytes transferred in each request are allocated into bins spanning the transfer duration

TABLE I
SUMMARY STATISTICS FOR TRACE DATA USED IN THIS STUDY

Sessions	4700
Users	591
URLs Requested	575,775
Files Transferred	130,140
Unique Files Requested	46,830
Bytes Requested	2713 MB
Bytes Transferred	1849 MB
Unique Bytes Requested	1088 MB

Explaining self-similarity

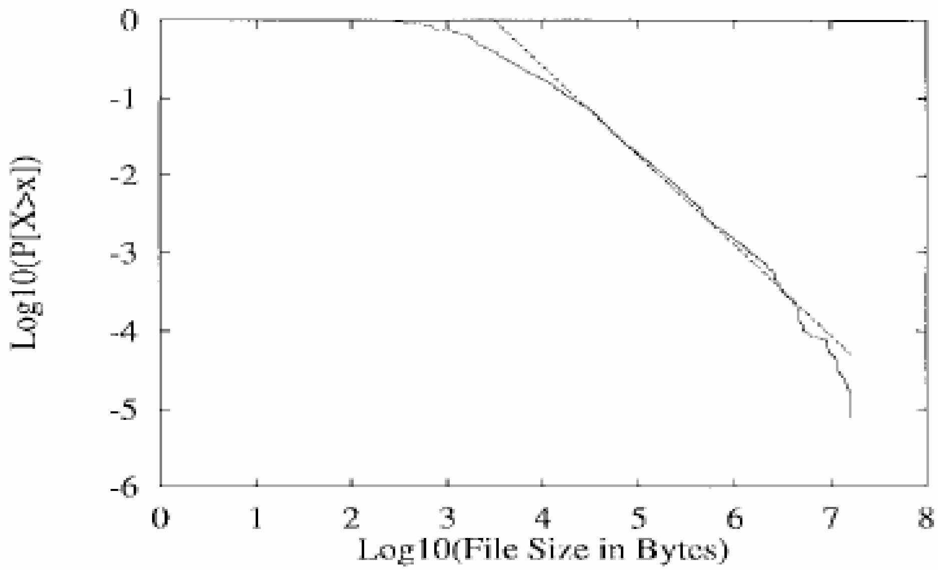
- Superposition of Heavy-Tailed Renewal Processes
  - ON times correspond to the transmission duration of individual web objects (assumption that transmission rate is constant during ON times)
  - OFF times correspond to the intervals between transmissions
- Examining Transmission Times
  - Distribution of web transmission times



The value of α using the Whittle estimator 1.2

Explaining self-similarity

– Why are transmission times variable ?
  - Size distribution of web objects (files) α=1.15



(a)

Explaining Self-Similarity

- Rather than the set of file requests made by users, the transmission times are more strongly determined by the set of available files.

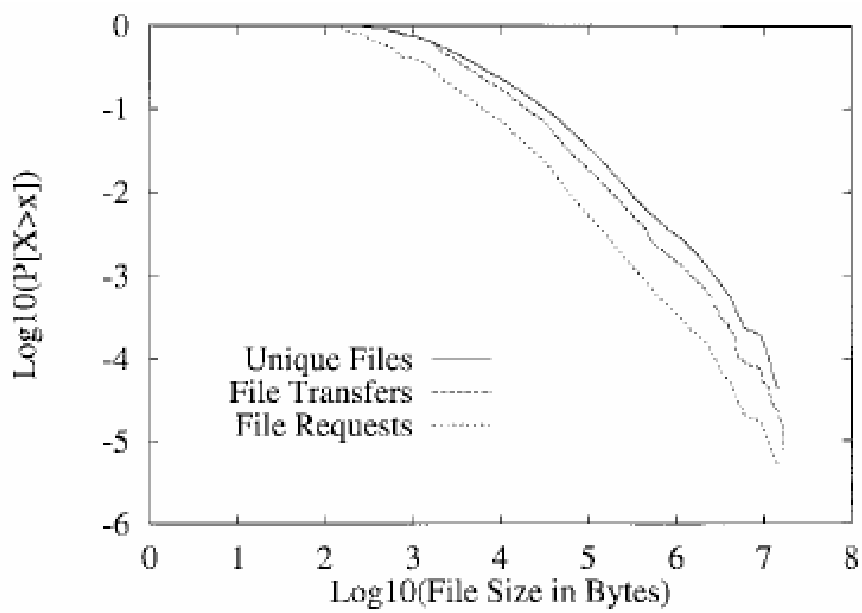
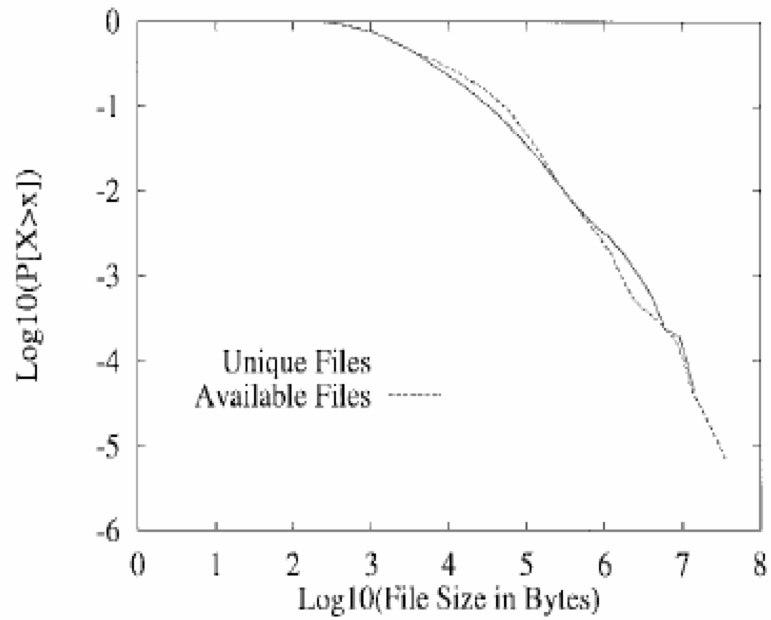


Fig. 6. LLCD plots of the different distributions.

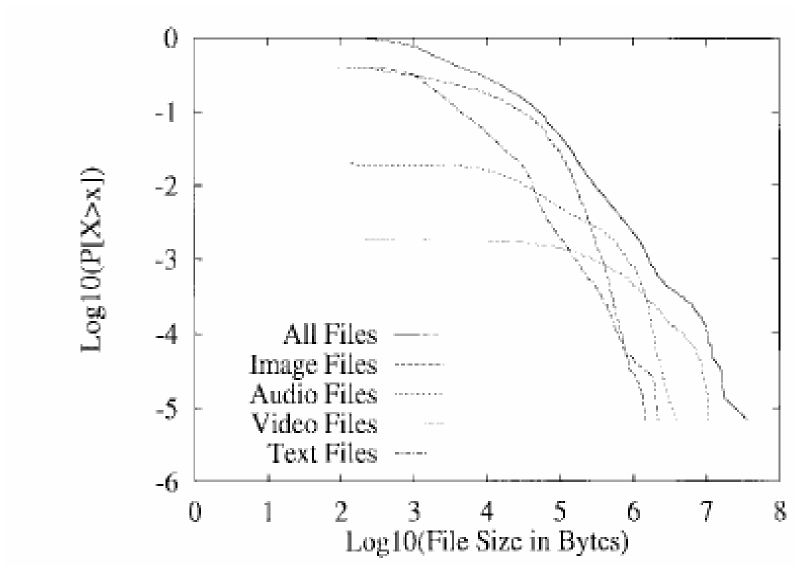
Explaining Self-Similarity

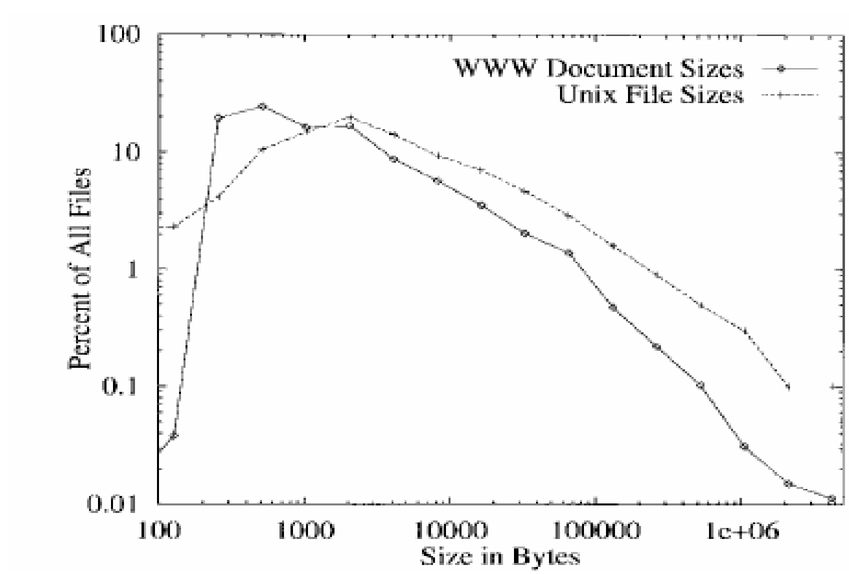
- Using the www-stat tool, file size distribution at web servers can be obtained
- This distribution closely matches Unique files distribution



Explaining Self-Similarity

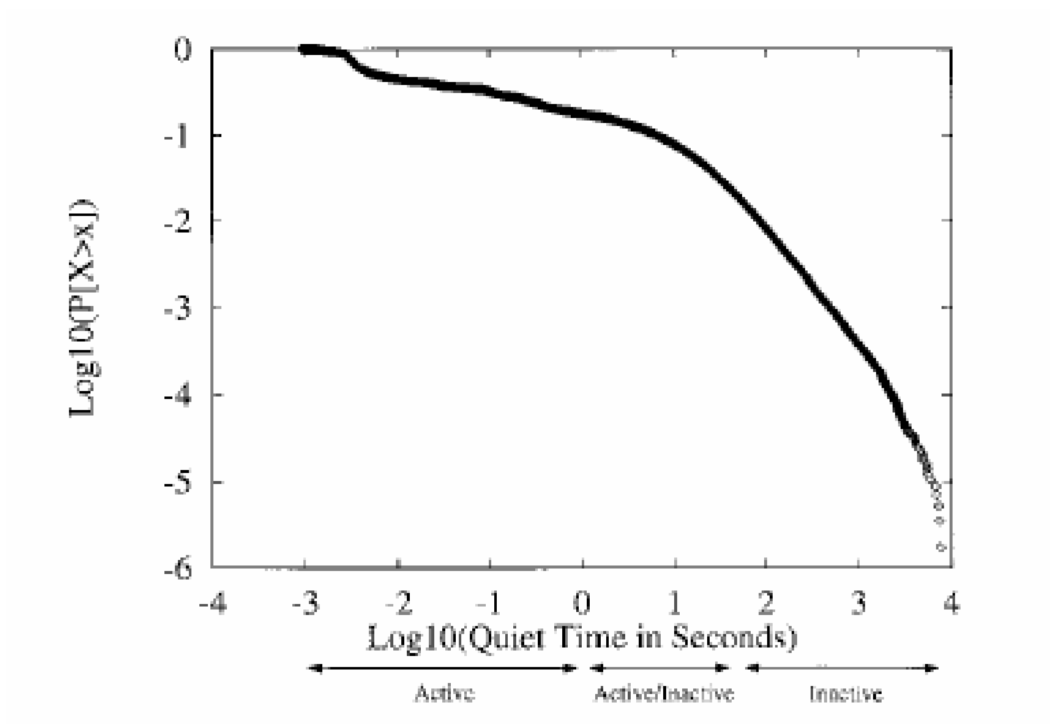
- Why are available file-sizes heavy-tailed ?
- Probably a property of most data storage systems





Examining Self-Similarity

- The Off-times' distribution(α=1.5)
- Weibull & Pareto distributions for active and inactive times



Future Work

- Parsimonious modeling is good enough but it doesn’t quantify the effects of various factors in traffic management
- Multi-resolution signal processing using Wavelets as used for quakes’ prediction may be used for traffic prediction too !!
  - Choice of wavelet and the subsequent math