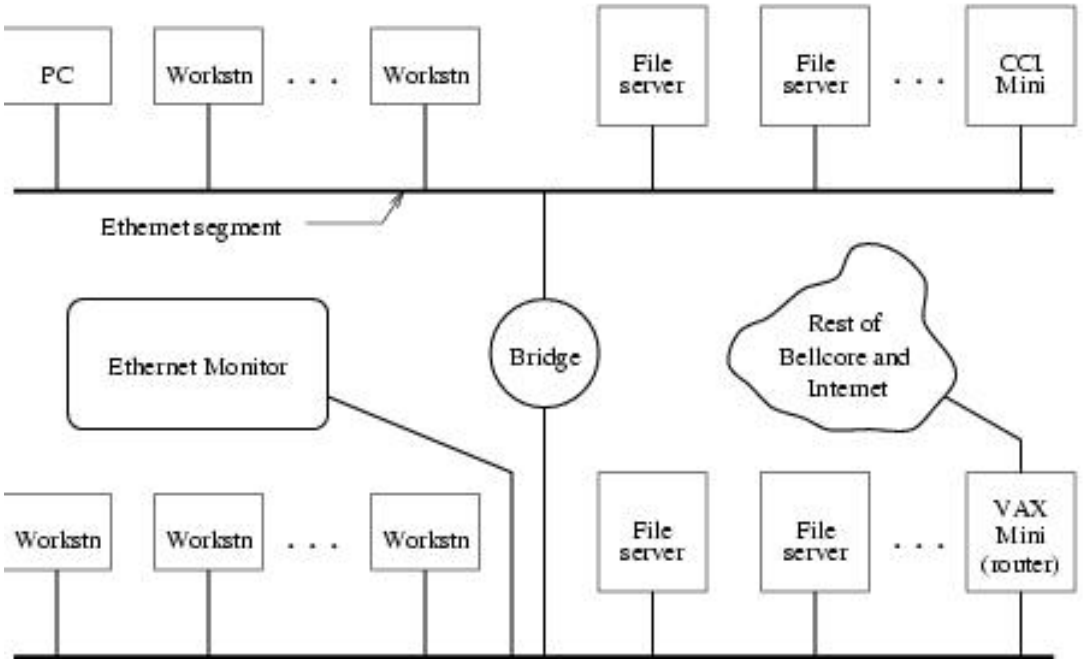
Internet Measurement & Analysis (IMA)

Topics in Internet Measurement

Modeling Network Traffic

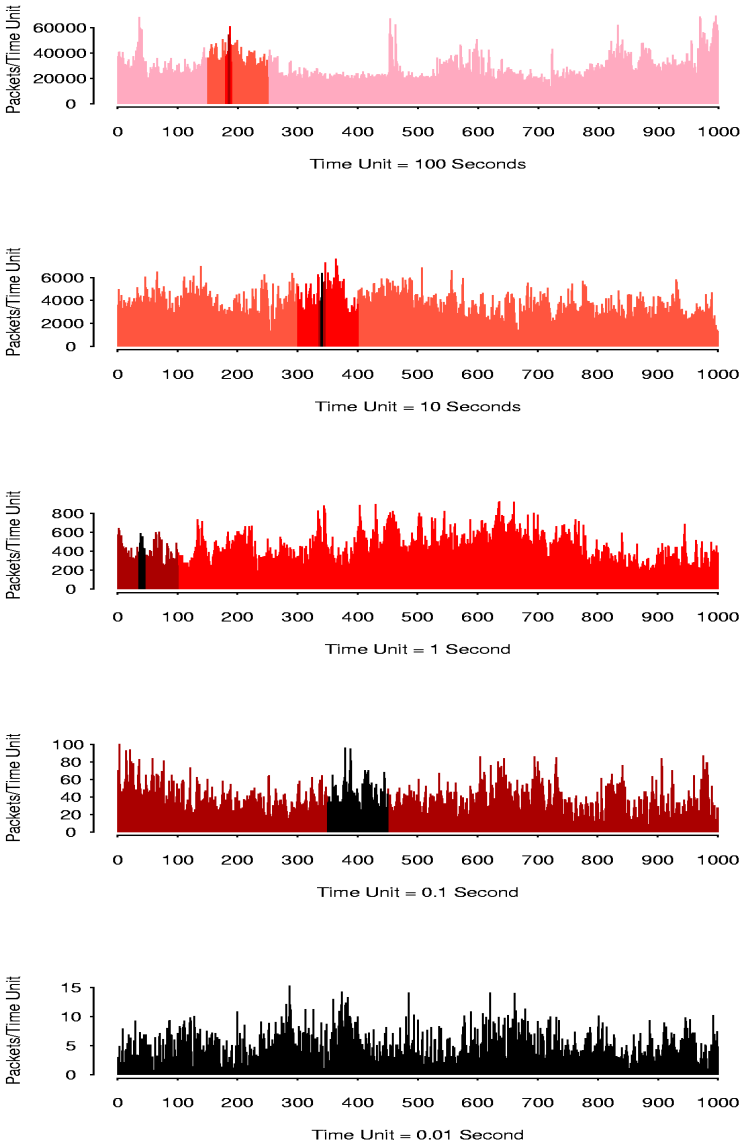
Prof. Georgios Smaragdakis, Ph.D.

Network topology 1989

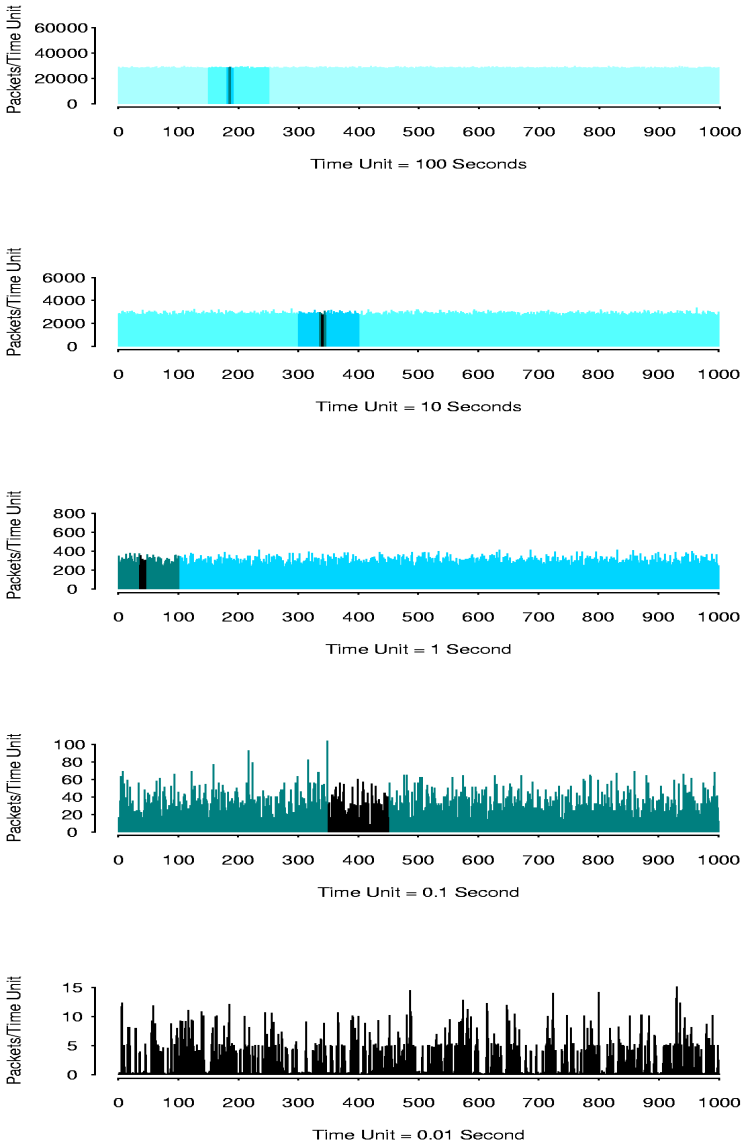


Network Traffic: Poisson?

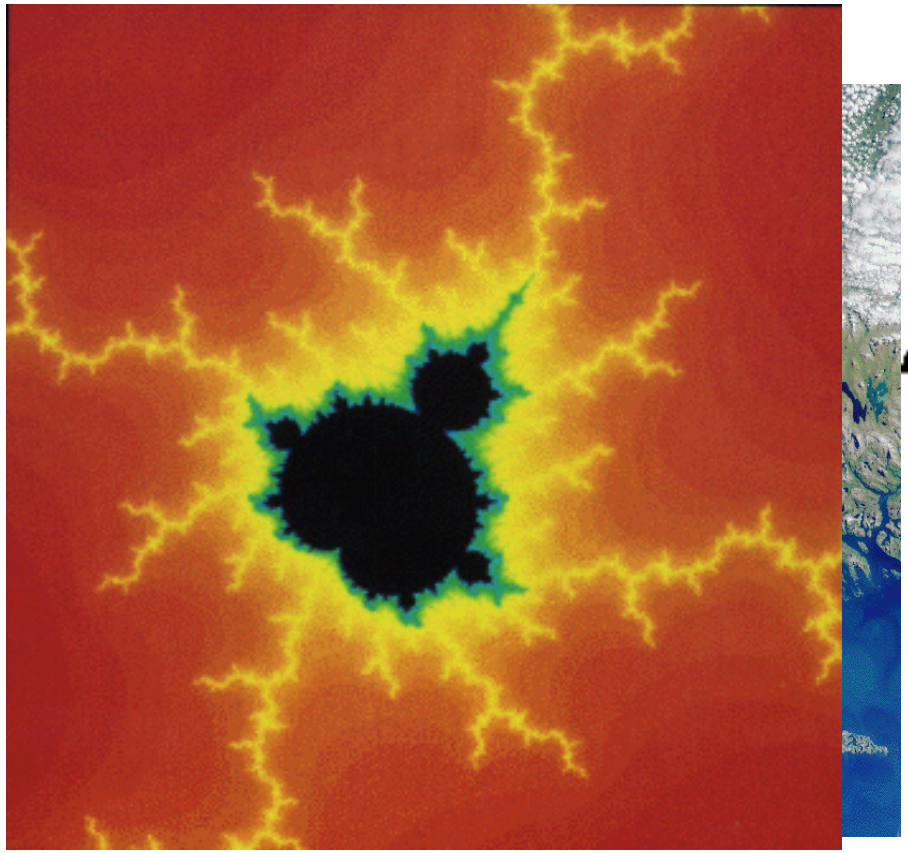
Realization

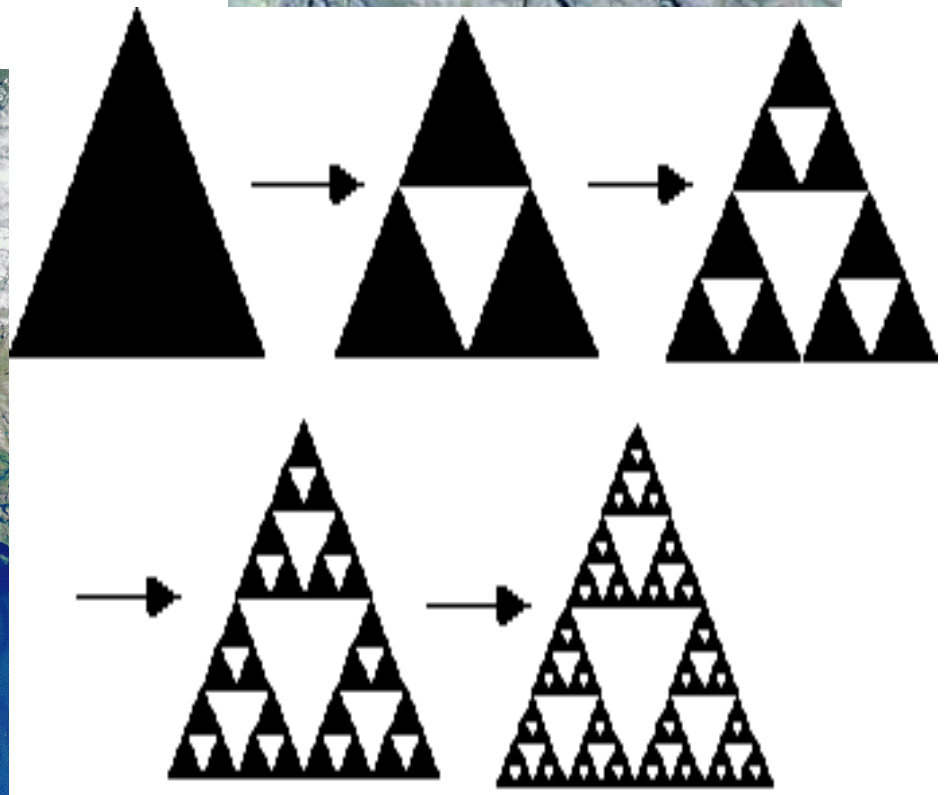


Poisson

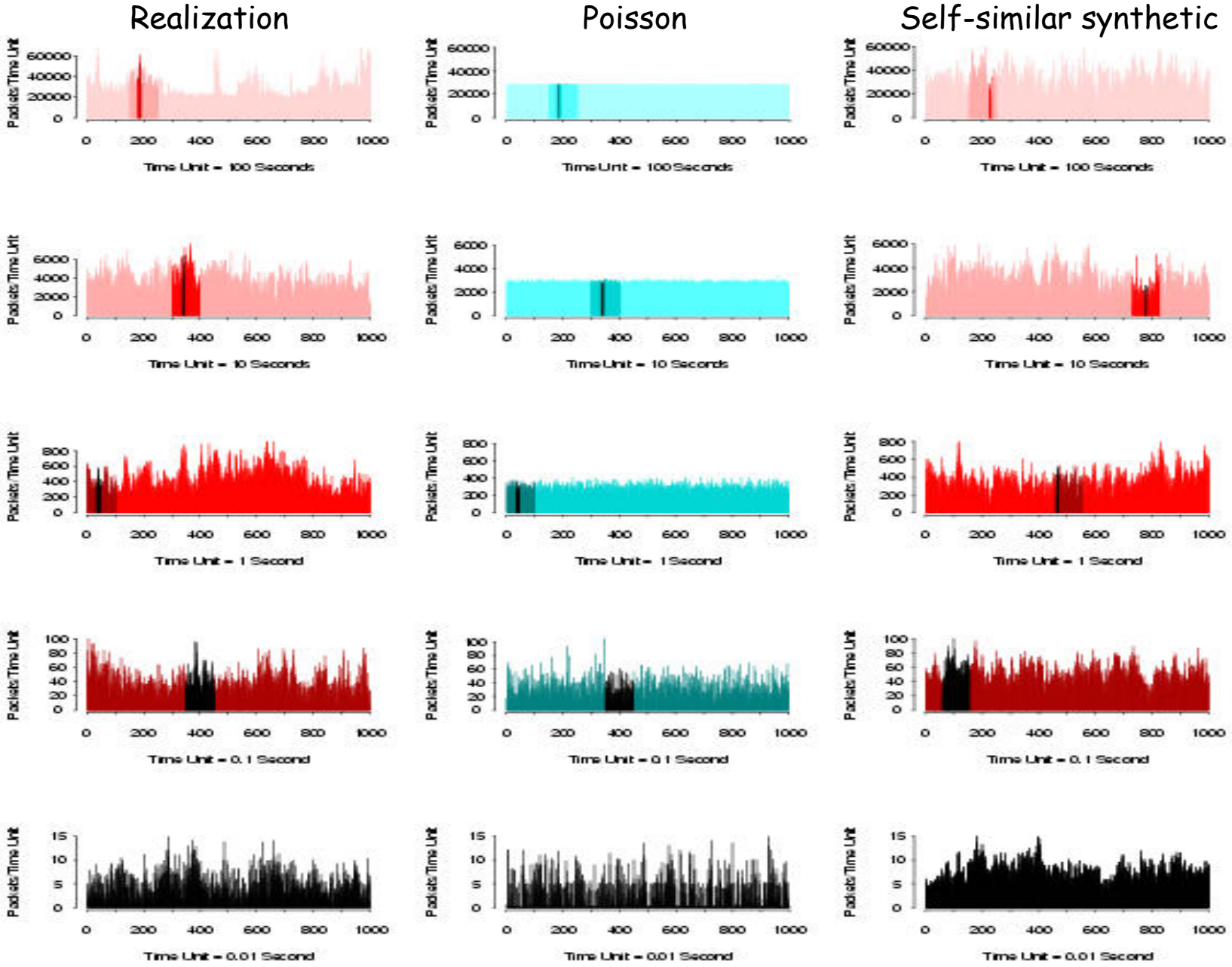


Self-similarity

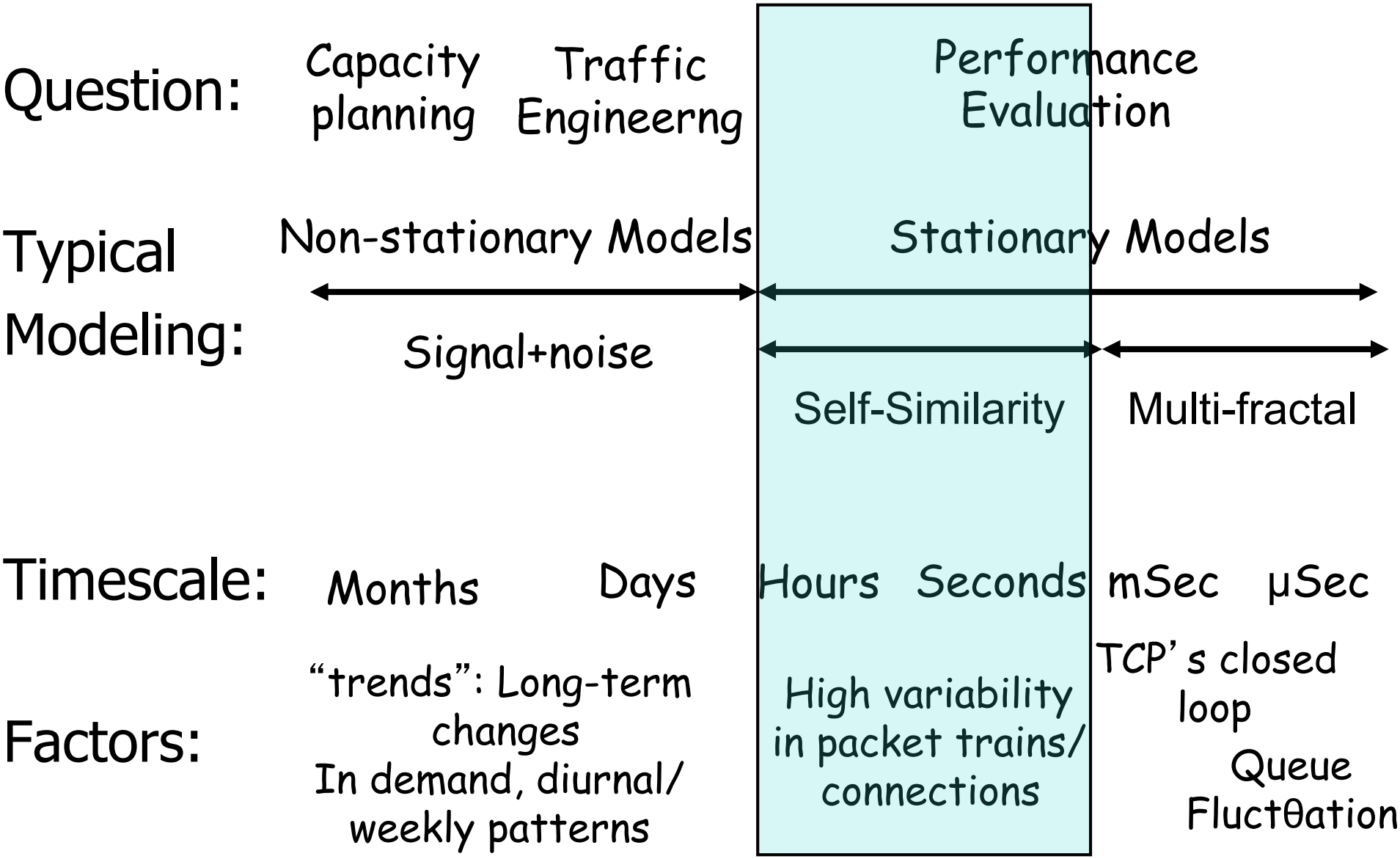




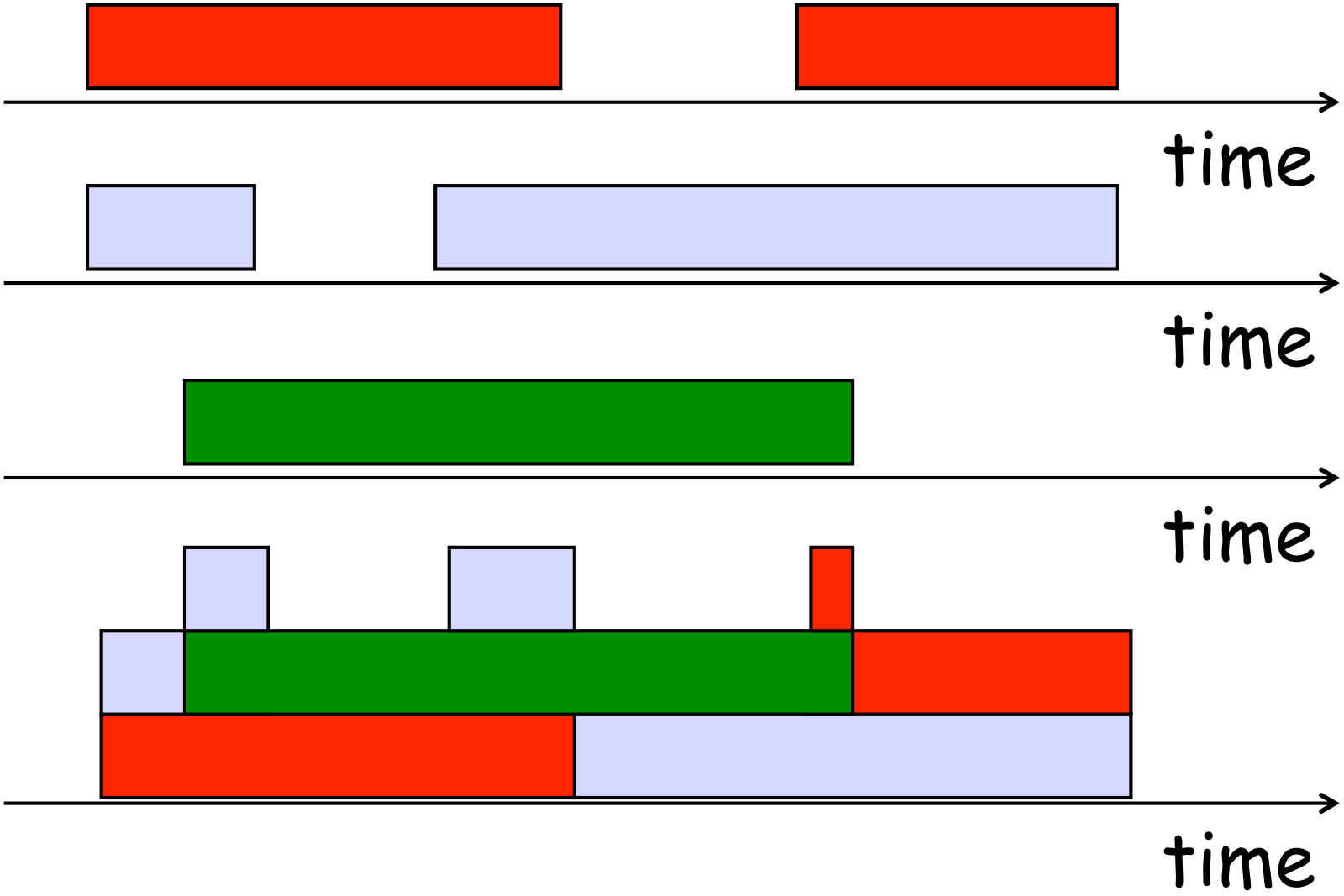
Examples: Nature (rivers, coastline)


Realization
Poisson
Self-similar synthetic
Packets/Time Unit
Time Unit = 100 Seconds
Time Unit = 10 Seconds
Time Unit = 1 Second
Time Unit = 0.1 Second
Time Unit = 0.01 Second

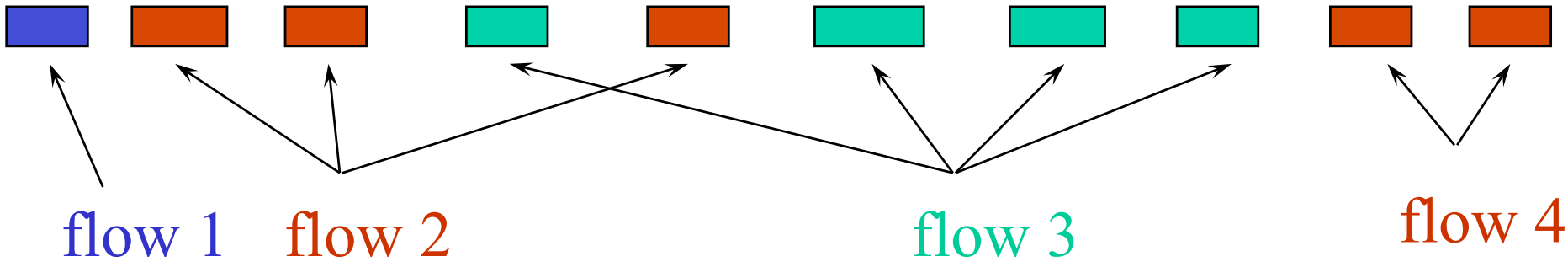
Overview of Traffic Analysis



Route Causes? Superposition of sources
(e.g., computer processes, human activity)



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

Light tails

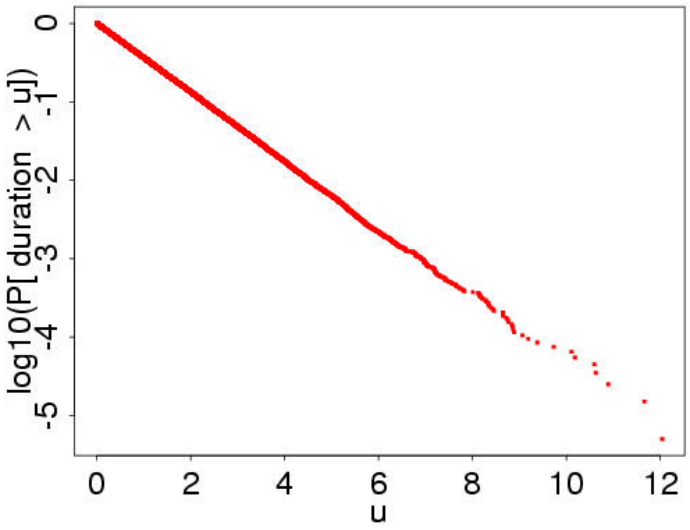
Definitions:

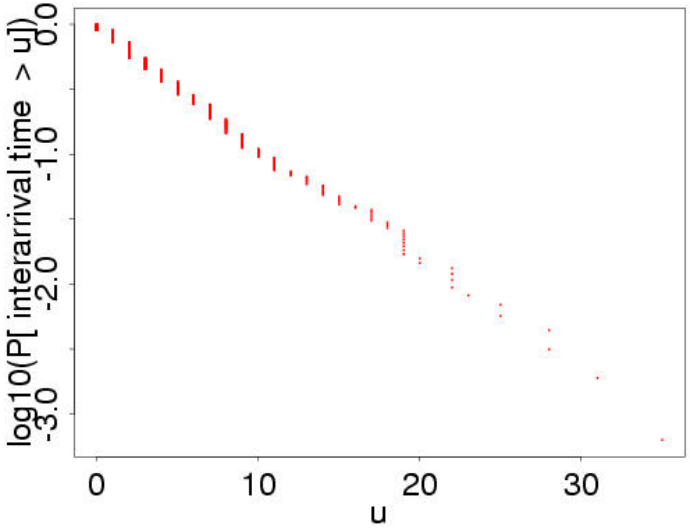
Short-range Dependent:

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

Light-tailed distributions

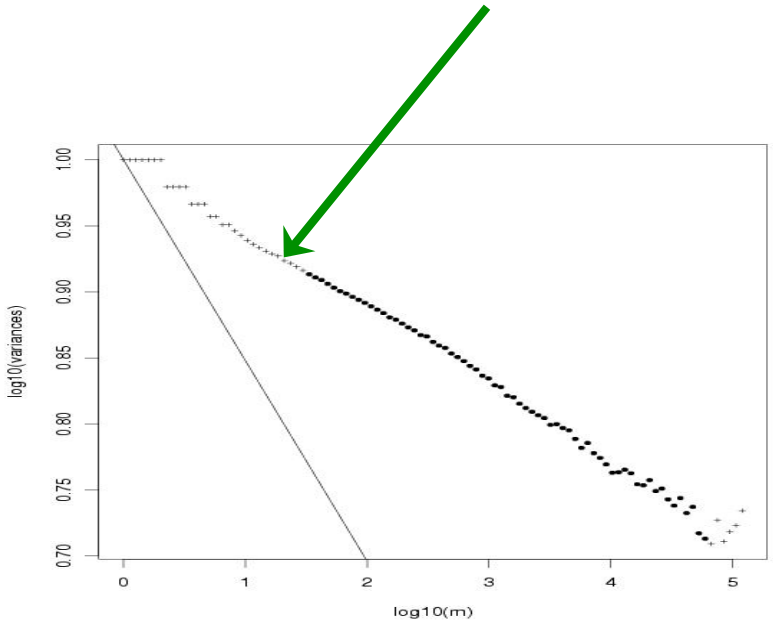
- Examples: Exponential, Normal, Poisson, Binomial
- Key features of Function (F):
  - F has limited variability
  - F is tightly concentrated around its mean
  - F has finite moments
  - P[X > x] (CCDF) vs. x on log-linear scale is linear for large x





Short-range dependence (SRD)

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



Self-similarity via heavy tails

Definitions:

Self-similar:

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

Long-range Dependent:

Superposition of independent ON/OFF sources is long-range dependent, if durations of periods are heavy-tailed

Self-Similarity via heavy tails

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

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

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)

Heavy-tailed distributions

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

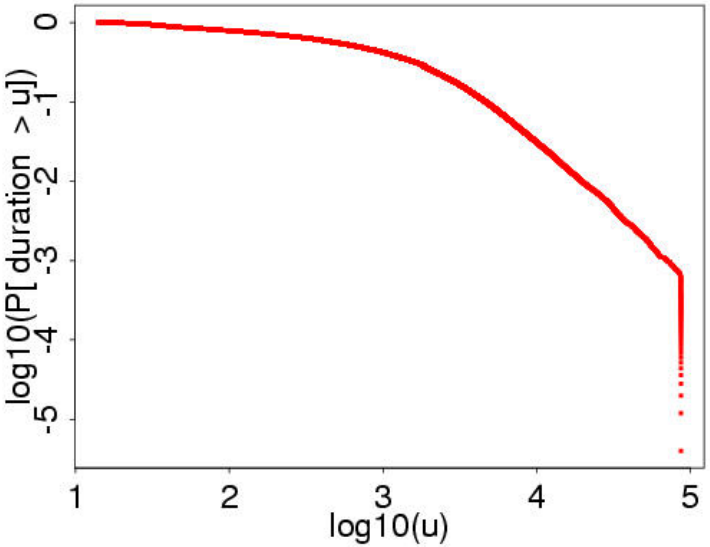
$$1 - F(X) = P[X > x] \sim cx^{-\alpha} \quad \text{as} \quad x \to \infty$$

$1 < \alpha < 2$, X has finite mean but infinite variance.

Where: $\alpha=\beta+1$

→parsimonious model (small number of parameters)

Heavy-tailed distributions



- Important features:
  - Finite mean but infinite variance
  - Heavy-tailed implies high variability
  - Tail decays like a power, hence power-law distribution
  - Plotting P[X > x] (CCDF) vs. x on log-log scale is linear for large x with slope α
- LRD is not a characteristic only of Heavy-tailed distributions

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 (self-similarity parameter H) if for all $m \geq 1$,

$$X = m^{1-H} X^{(m)}$$

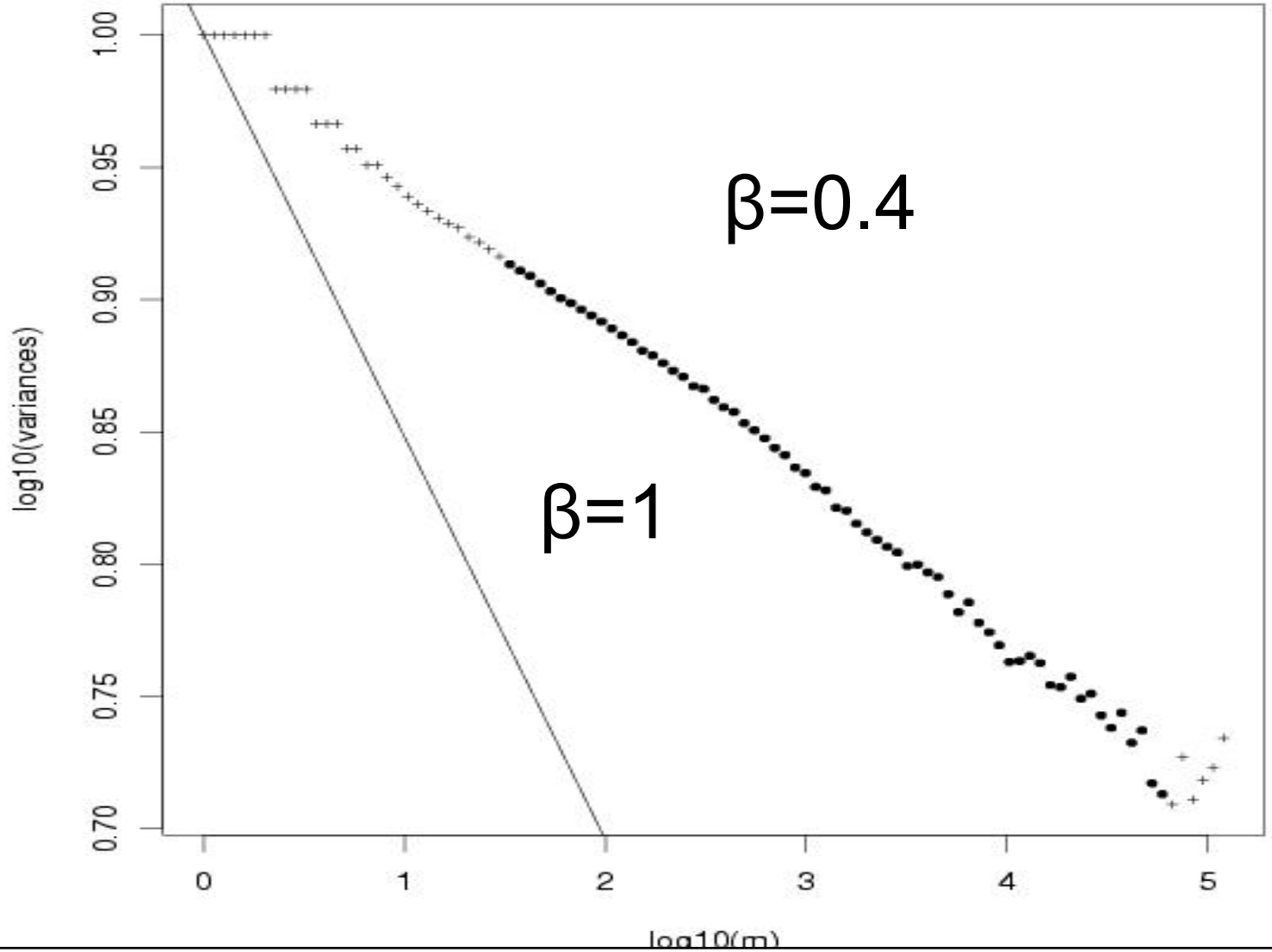
H is called the Hurst parameter. To be self-similar, 0.5 < H < 1

Test of *variance*, i.e., the measure of how far a set of samples is spread out, $var(X)=E((X - E(X))^2)$.

$$\mathrm{var}(X^{(m)}) \sim c m^{-2H-2} \quad \text{as} \quad m \to \infty$$

Variance time plot

$H=1-\beta/2 = 0.8$ in (0.5,1)



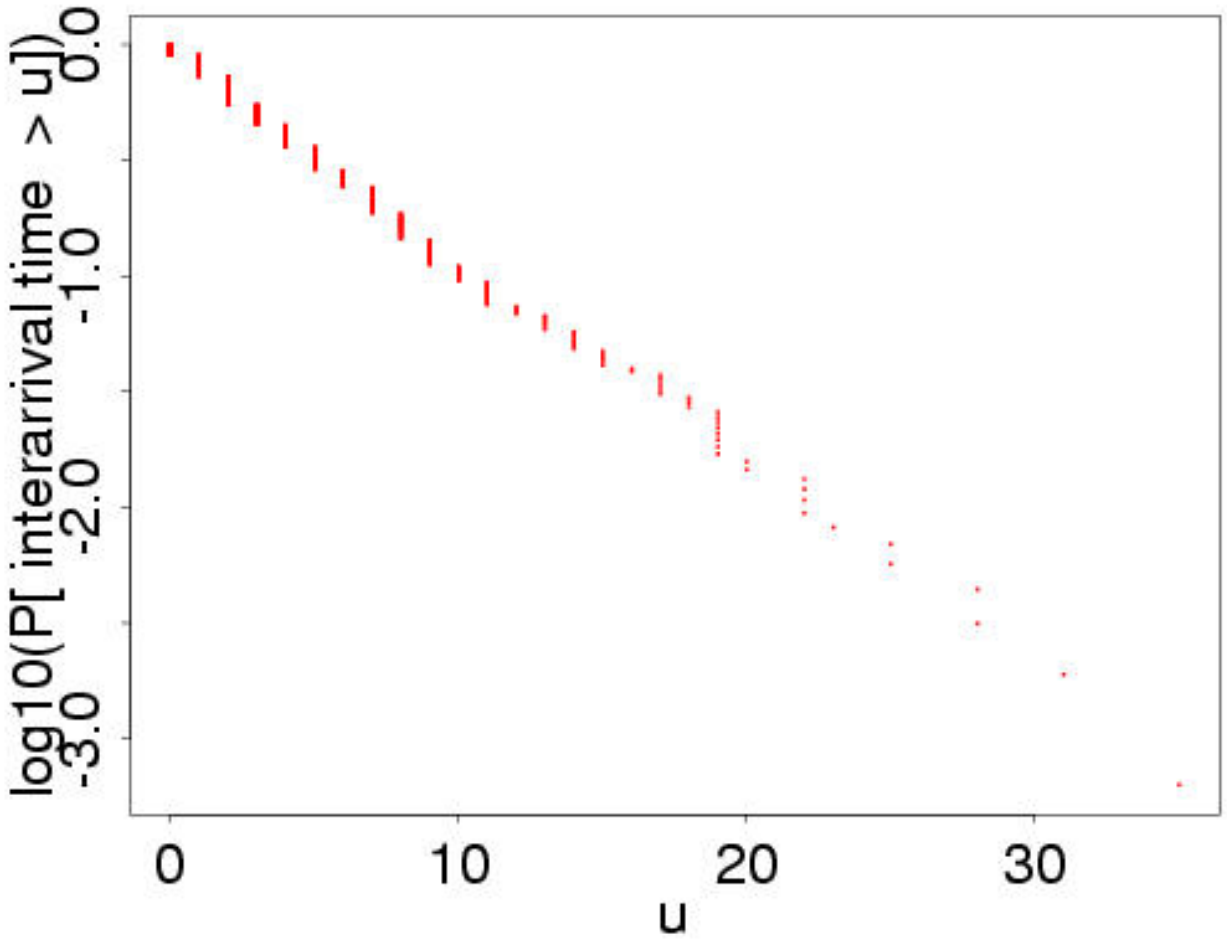
Long-range dependence (LRD) & Heavy-tail Distributions

- There are observations in the past that are correlated with current observations
- Parsimonious models available
- It changes the way we design systems

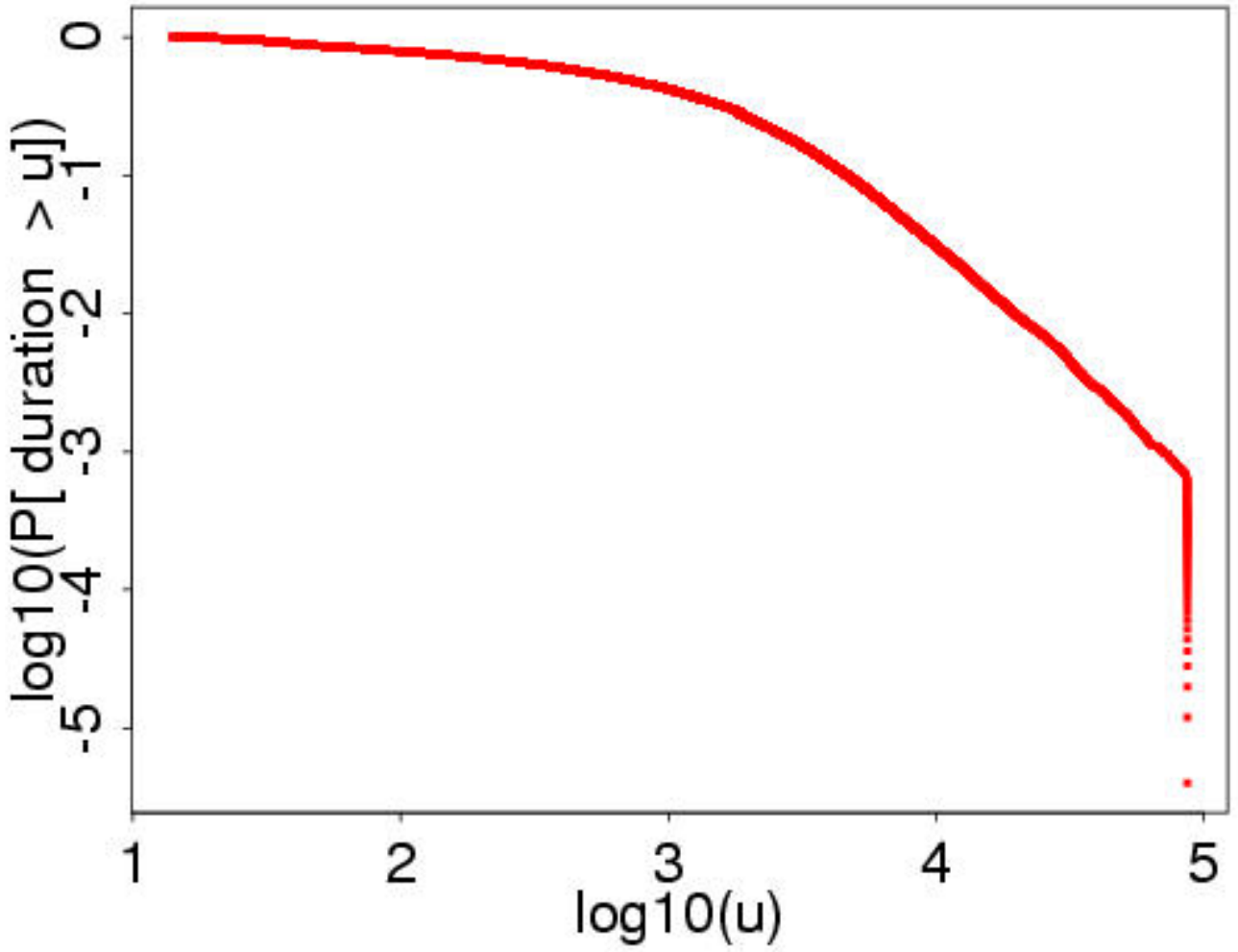
(e.g., how to deal with bursts, effect of queuing, protocol design)

Detour
Characteristics of modem calls (~1999)

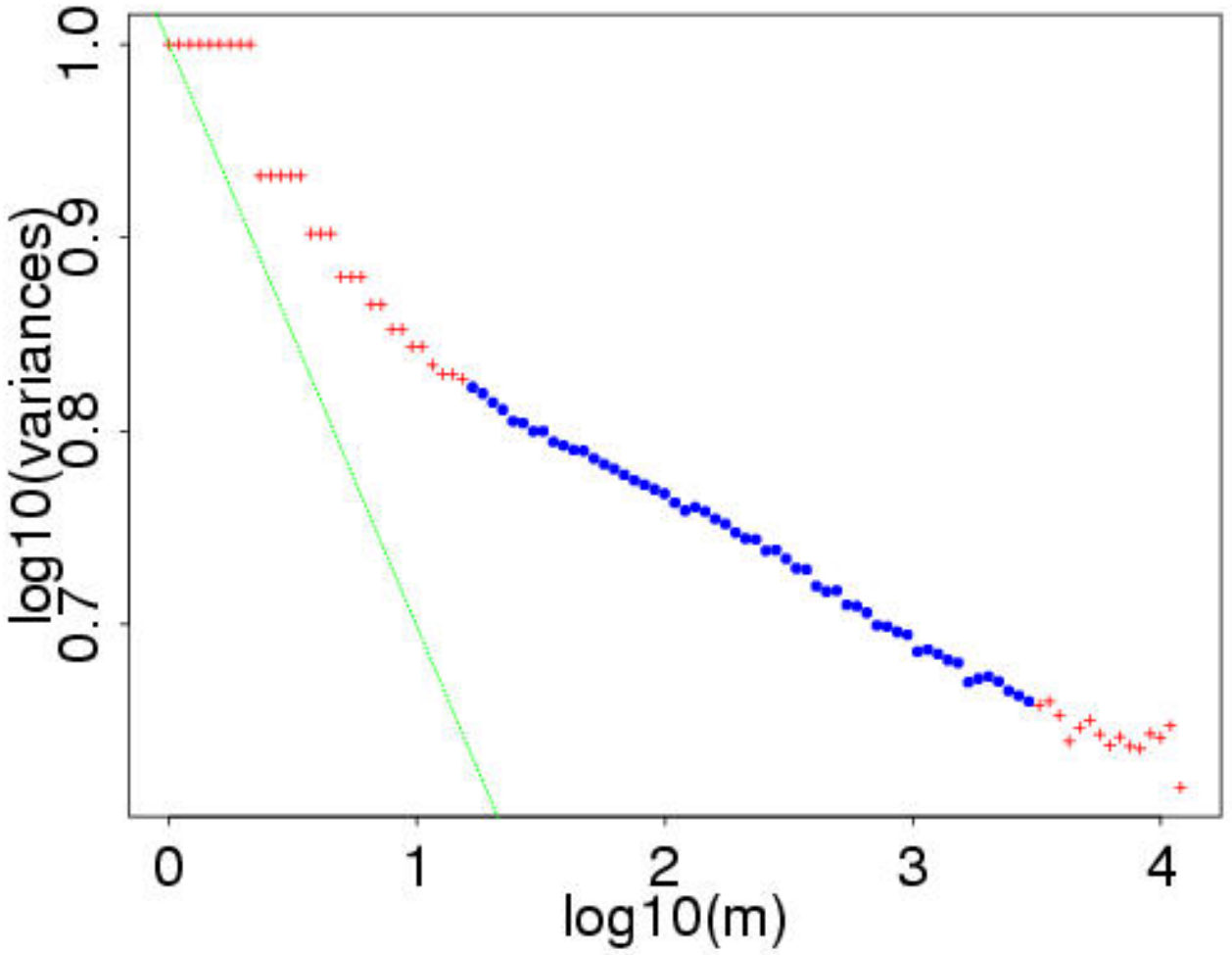
Interarrival times of modem calls



Durations of modem calls

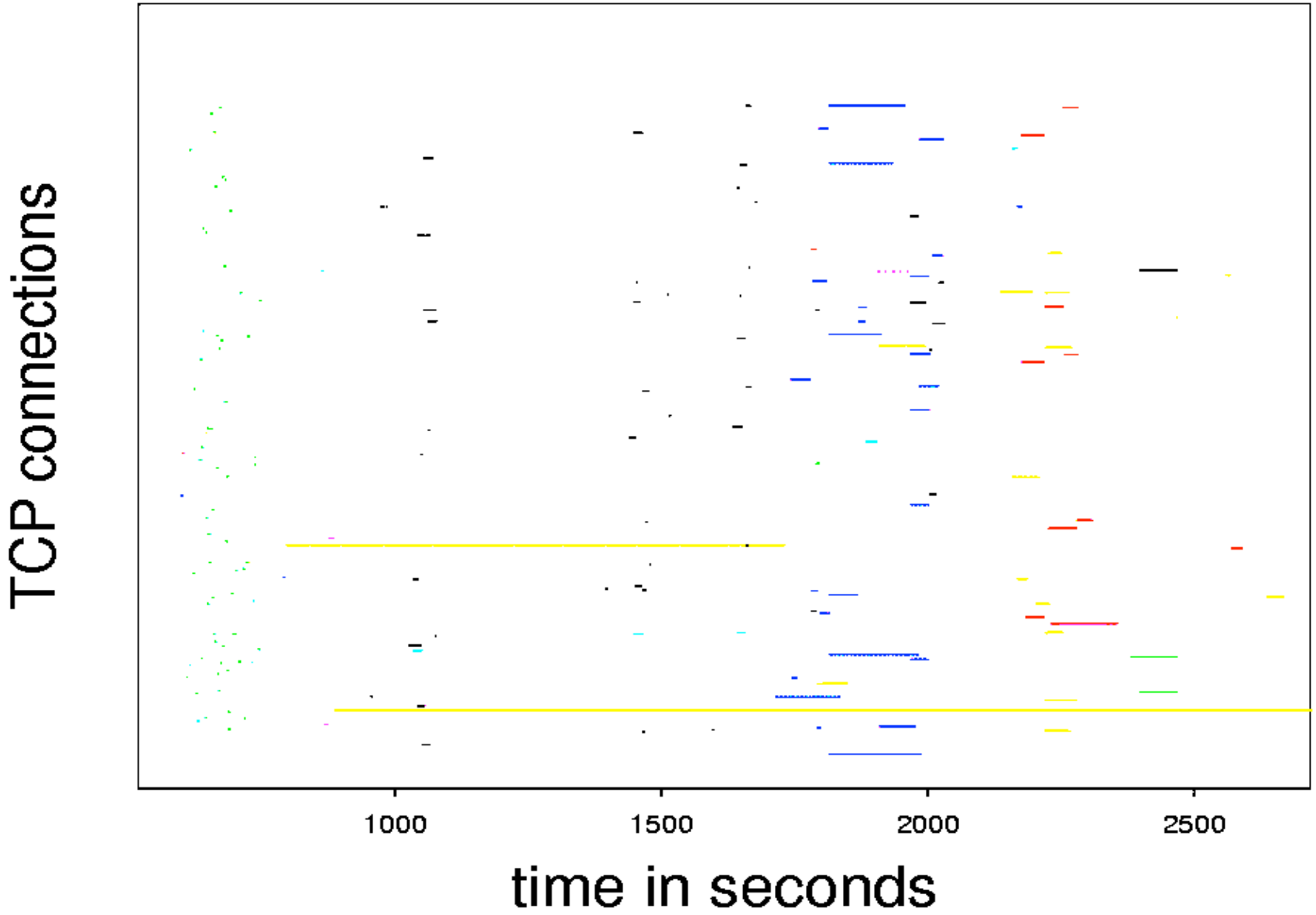


What about packets from modem calls

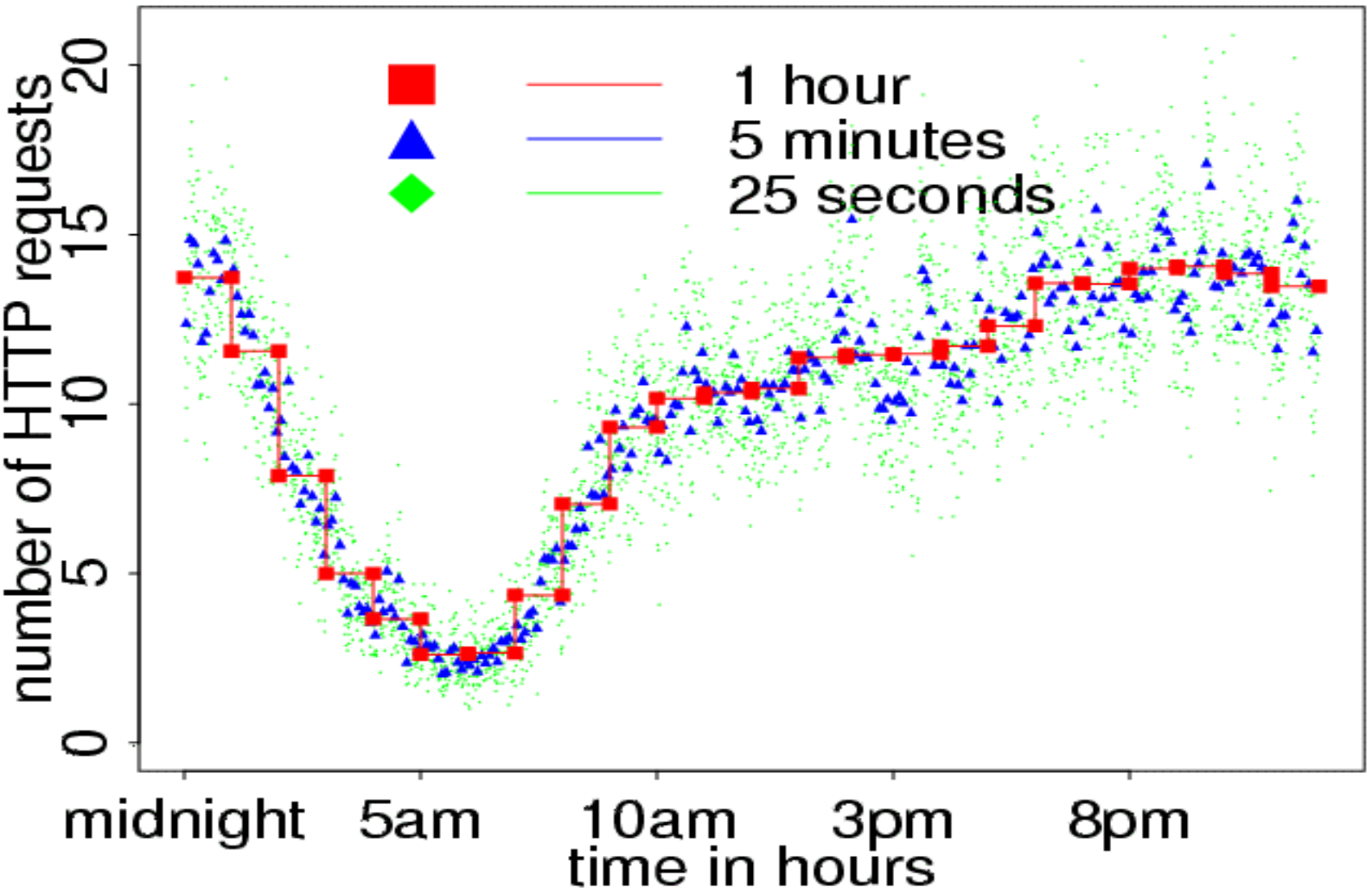


Detour
Characteristics of Web
(~2000)

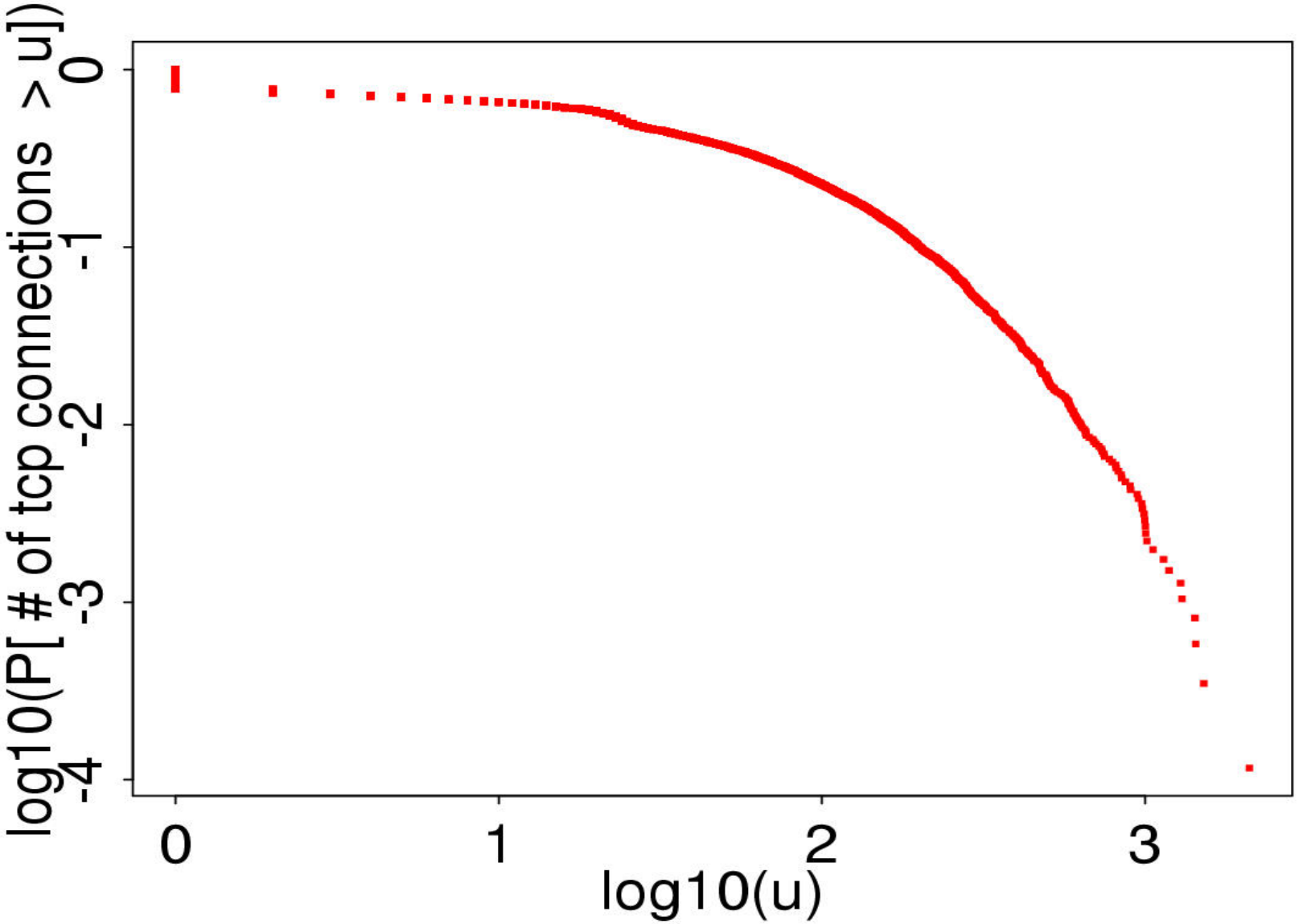
General characteristics of WWW transfers



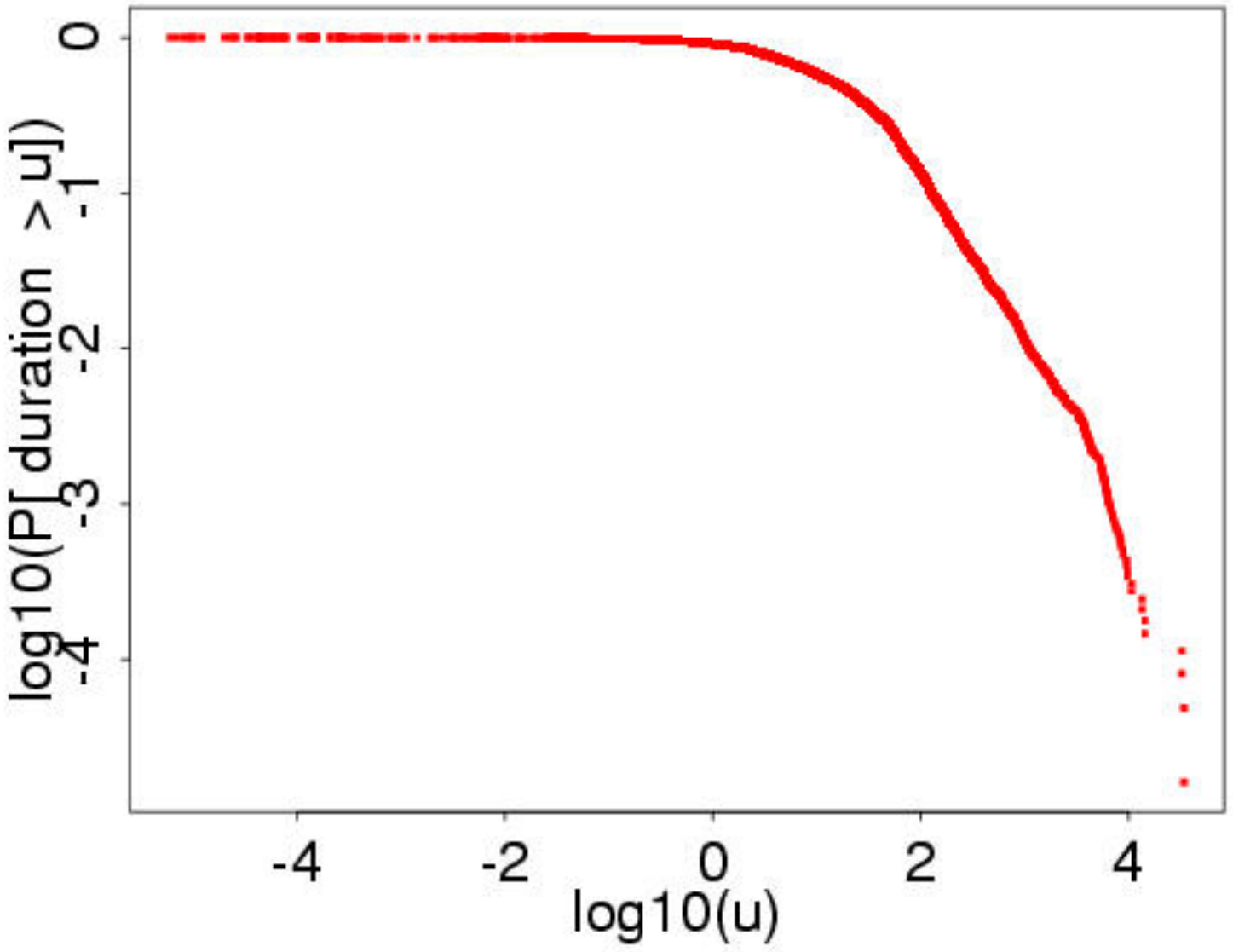
General characteristics of WWW transfers



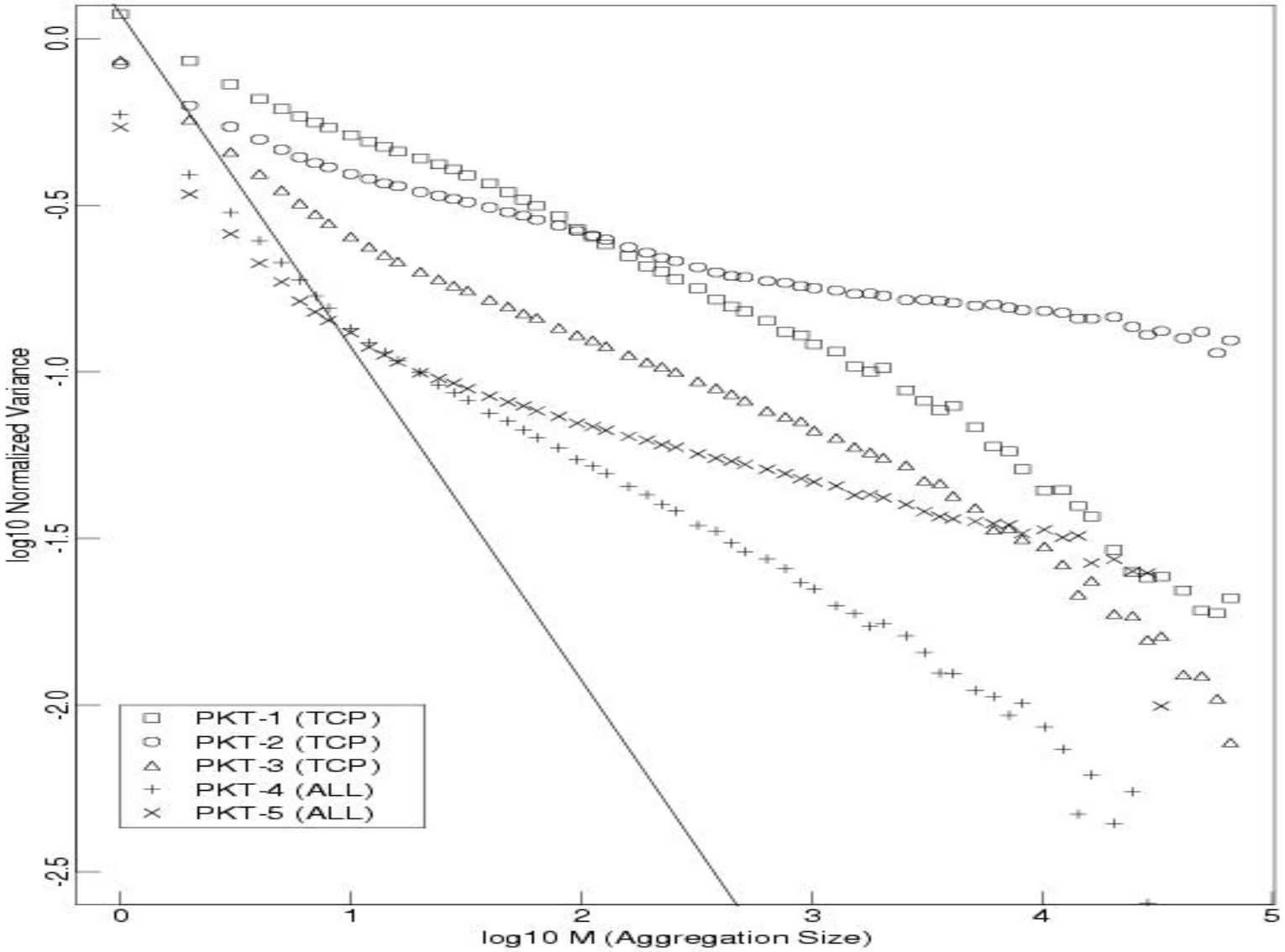
of TCP connections per session



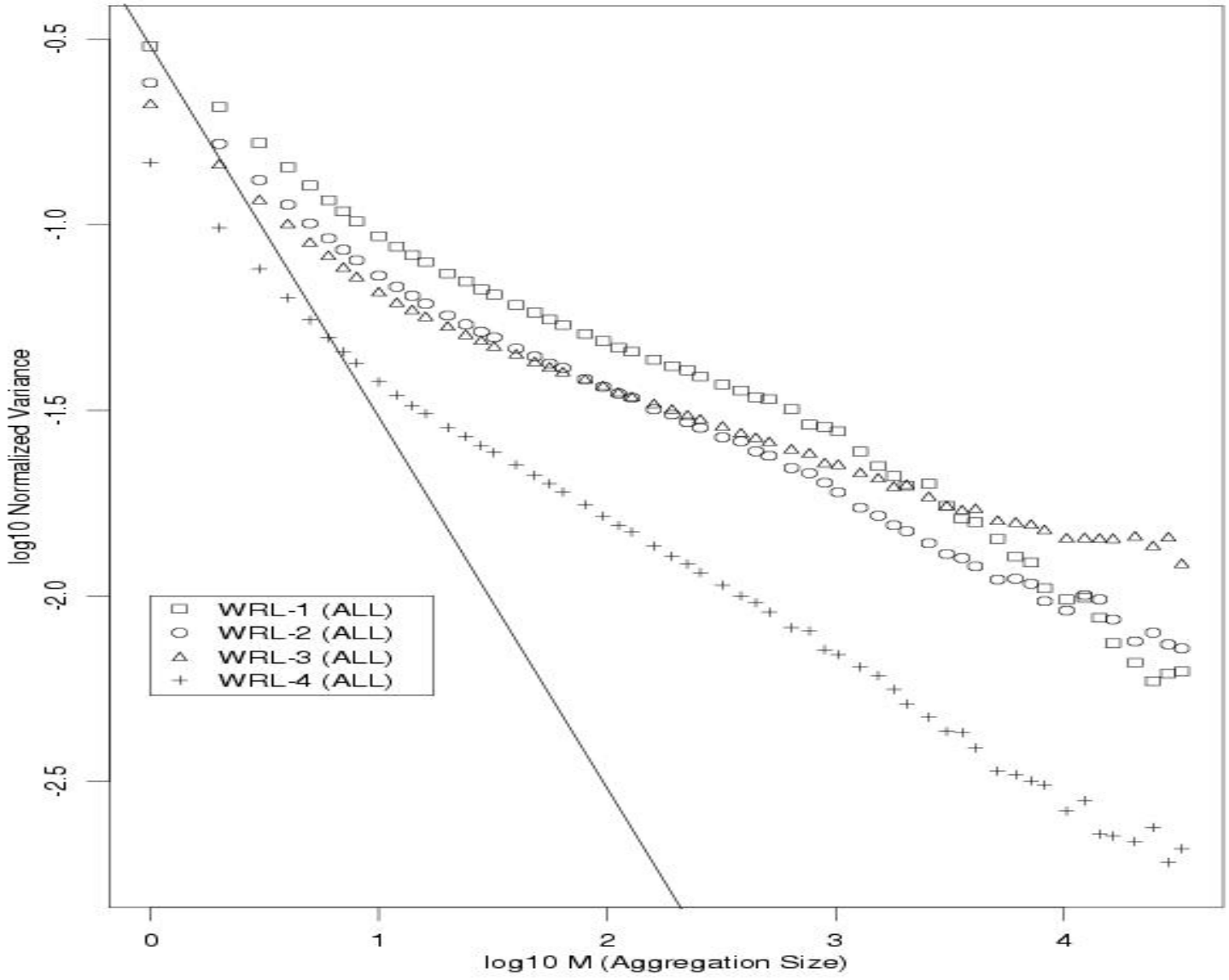
Flow durations



Self-similar?



Self-similar?



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)????

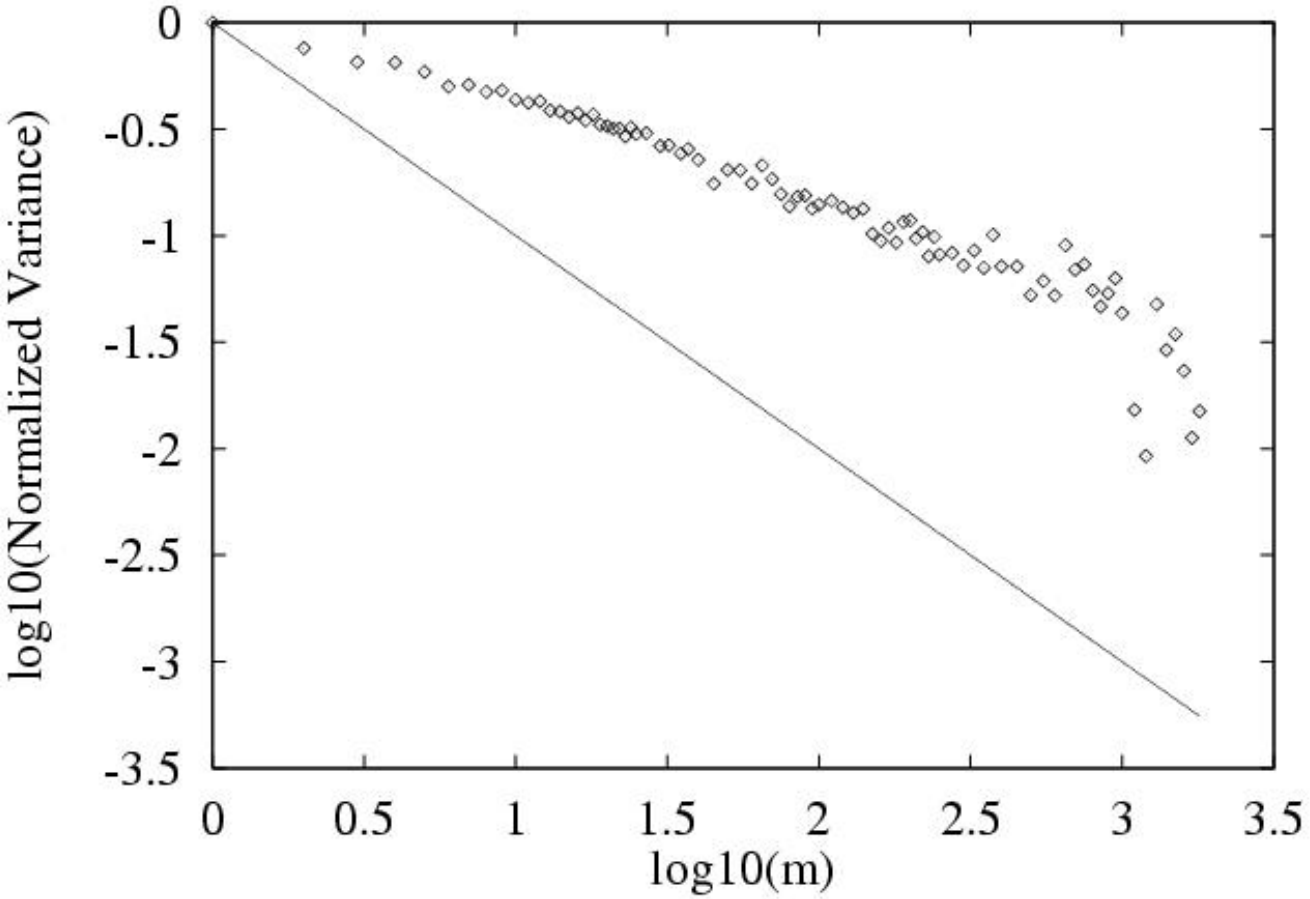
- Arrivals of User session
- # of Web requests per session
- Dist. of # of bytes, pkts, duration per request?

Web client trace analysis 1995

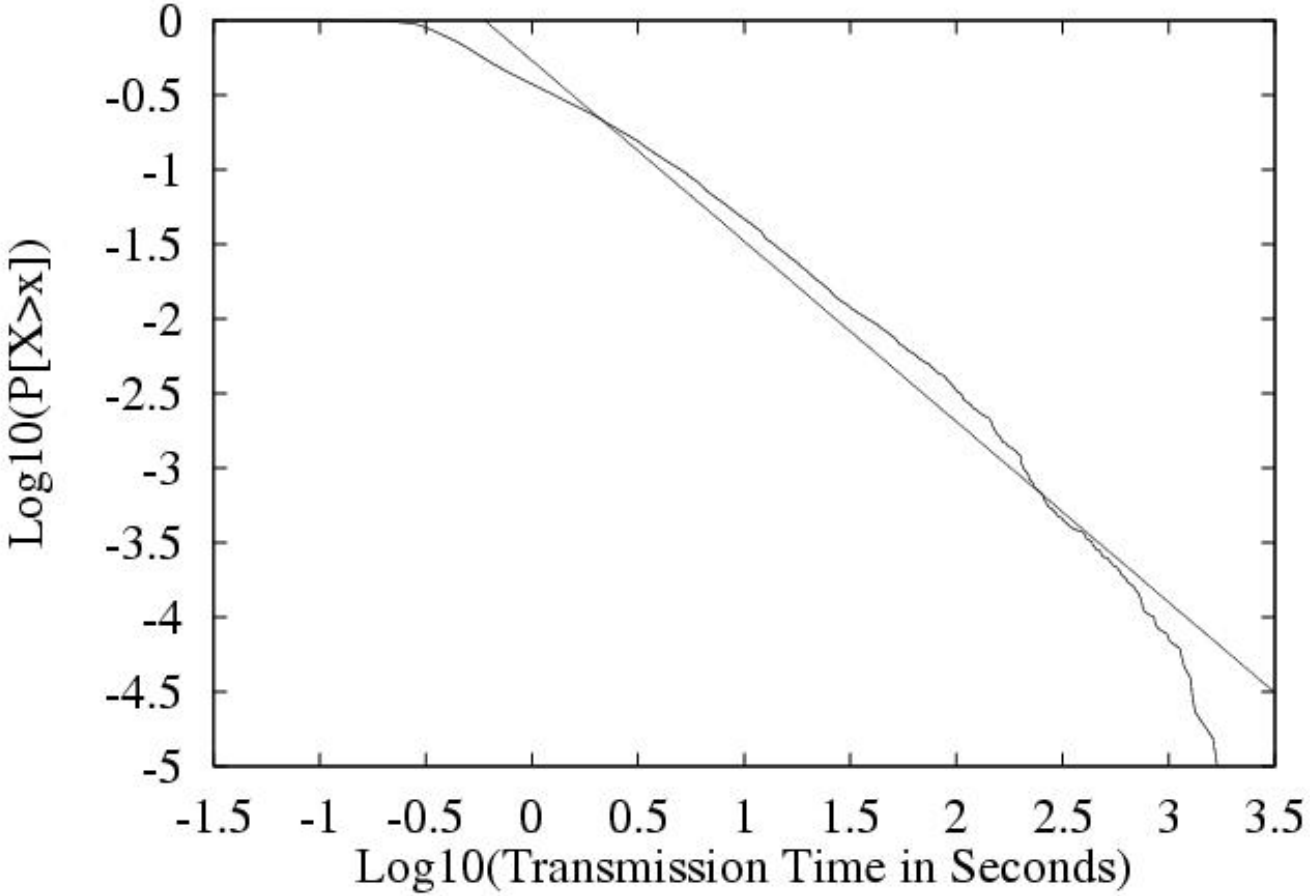
- Modified Web browser (Mosaic)
- Population: students at BU
- Duration: 21 Nov 94 to 8 May 95

Sessions	4,700
Users	591
URLs Requested	575,775
Files Transferred	130,140
Unique Files Requested	46,830
Bytes Requested	2,713 MB
Bytes Transferred	1,849 MB
Unique Bytes Requested	1,088 MB

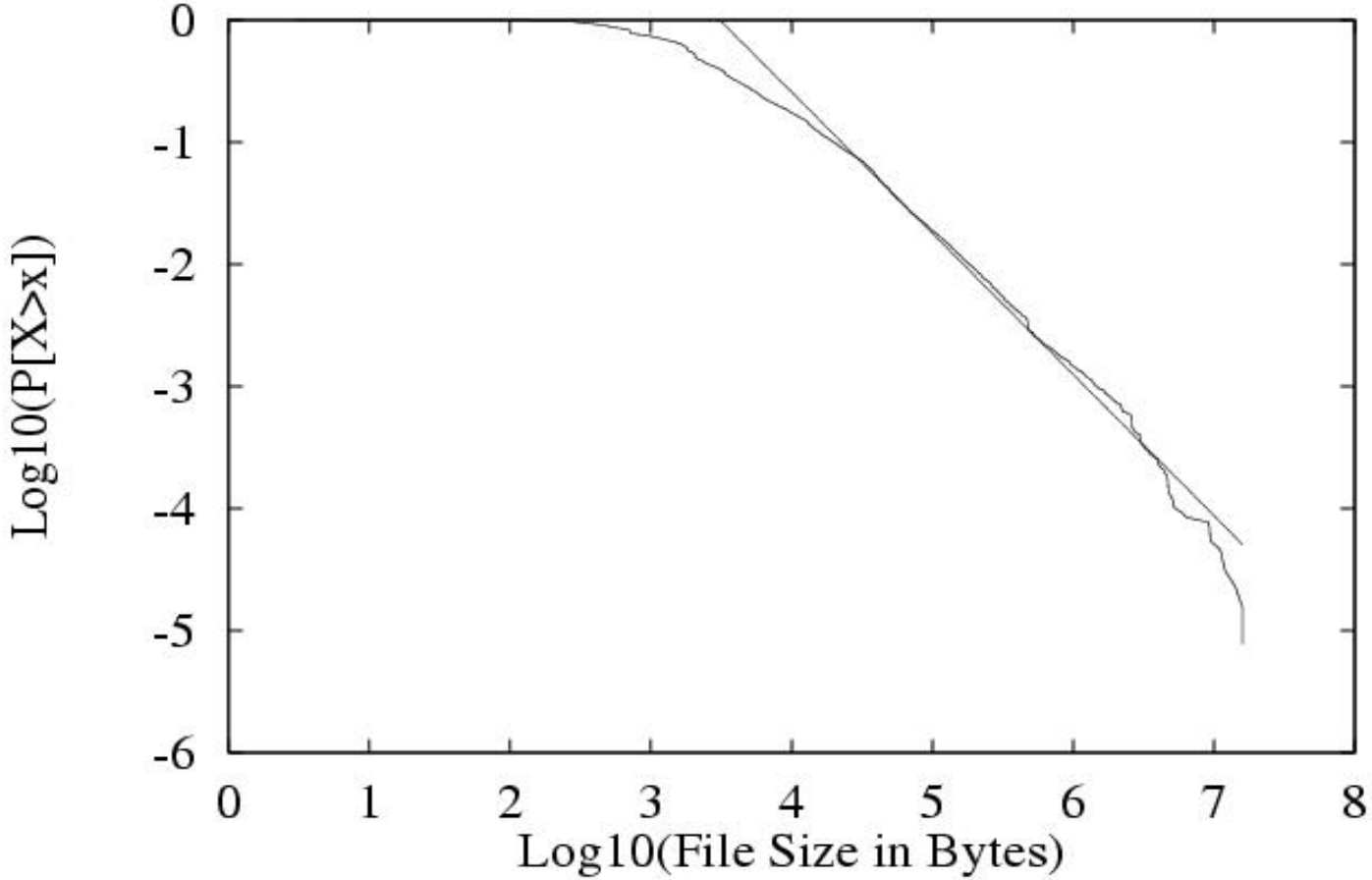
What about WEB traffic



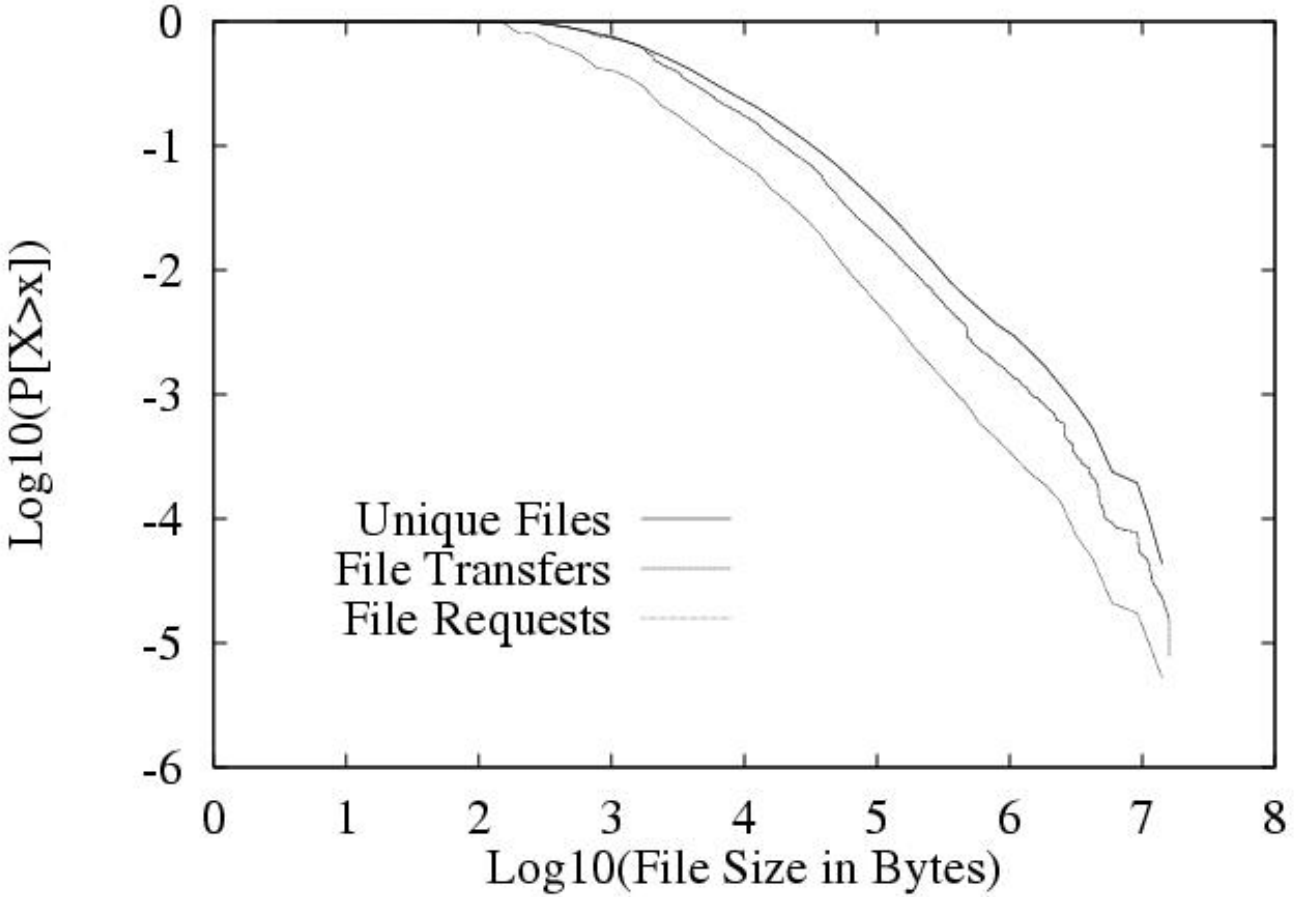
Durations of WEB transfers???



File size of WEB transfers???

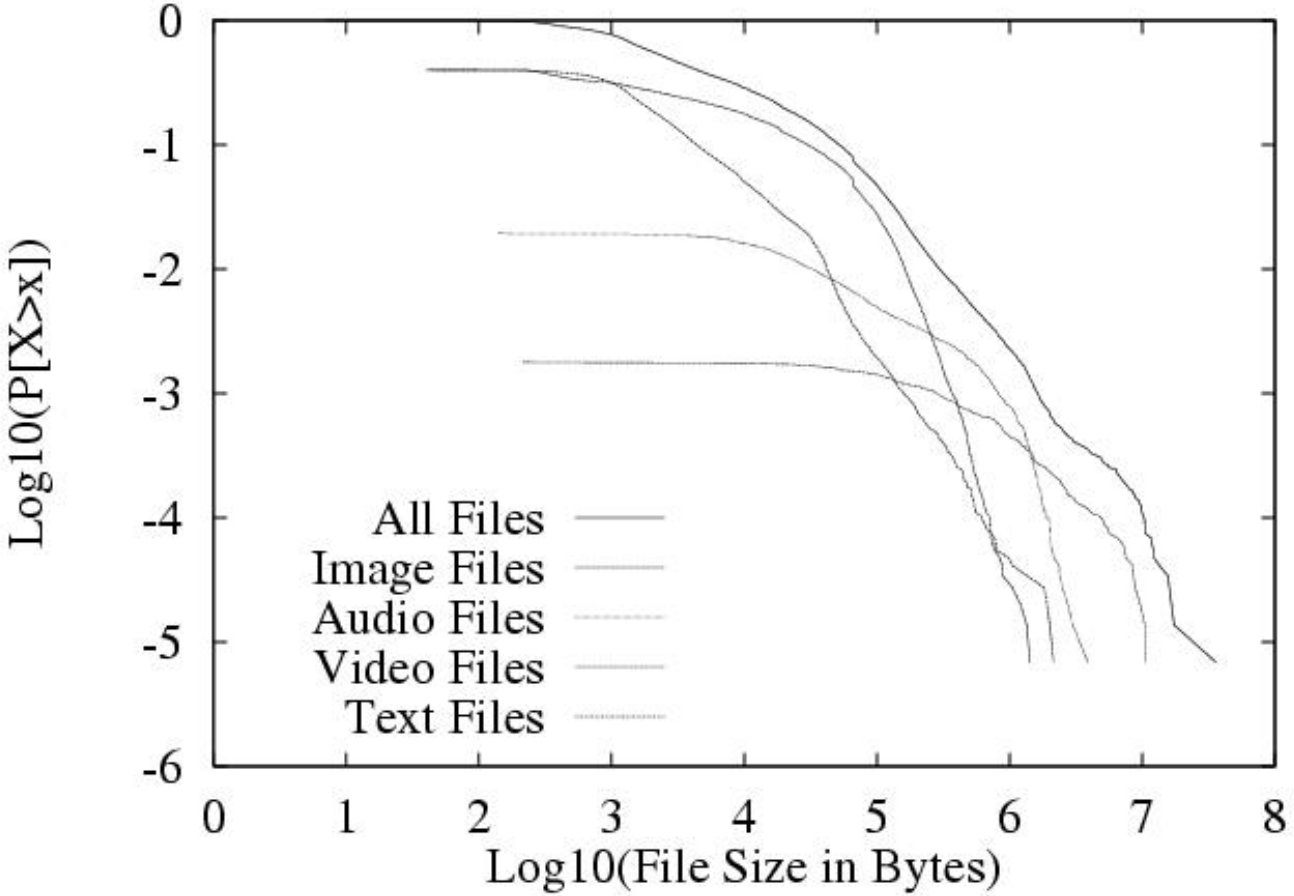


Unique files vs. files transferred?



What about the available files?

Long tail due to superposition of heterogeneous filesizes/filetypes



References

"On the Self-Similar Nature of Ethernet Traffic"
WE Leland, MS Taqqu, W Willinger, DV Wilson,
SIGCOMM 1993

"Self-similarity in World Wide Web traffic: evidence and possible causes"
Mark Crovella and Azer Bestavros
IEEE/ACM Transactions on networking 5 (6), 835-846, 1997