Introduction to Applied Probability Modeling of Data Networks

Sidney Resnick
School of Operations Research and Information Engineering
Rhodes Hall, Cornell University
Ithaca NY 14853 USA

http://legacy.orie.cornell.edu/~sid
sir1@cornell.edu

Madrid

May 21, 2012

Introduction
Stylized facts
Broad Issues (BI's)
Approach to Modeling
Tech points
Heavy tails ⇒ LRD
References

1. Introduction

Does data network traffic behave statistically
like telephone network traffic?

Action:

- Stop assuming the two types of networks behave the same.
- Start checking.

Initially at Bellcore (now Telcordia) and later at AT&T Labs-Research, and Boston University and ... high resolution measurements (data) were collected. Usually the data consisted of counts of bits, bytes, packets etc per unit time (eg millisecond). This could then be aggregated to coarser time scales. For example

- ...
- 10 seconds
- 10 milliseconds

- ...
- 1 second
- 1 millisecond

Significant examples:

- LAN's and WAN's
  - Willinger et al. (1997)
  - Duffy et al. (1993)
  - Leland et al. (1993)
  - Willinger et al. (1995)
- WWW traffic (BU)
  - Crovella and Bestavros (1996),
  - Crovella and Bestavros (1997).

Measurements on data networks exhibit features surprising by the standards of classical queueing and telephone network models. These are called

- *invariants*

which is to networks what

- *stylized facts*

are to finance.

2. Stylized facts

1. Heavy tails abound for such things as

    - file sizes,
    - transmission rates,
    - durations (file transfers, connection lengths).

    (See Arlitt and Williamson (1996); Leland et al. (1994); Maulik et al. (2002); Resnick (2003); Resnick and Rootzén (2000); Willinger (1998); Willinger and Paxson (1998); Willinger et al. (1998).)

    Reminder: a random variable X has a heavy tail if

    $$P[X > x] \sim x^{-\alpha} L(x), \quad \alpha > 0,\ x \to \infty,$$

    and think of $L(x)$ as constant if you do not know *slow variation*.

    - Tail exhibits power law decay.
    - Limited moments:

    $$E(|X|^{\alpha+\delta}) = \infty, \quad \delta > 0.$$

    - If $1 < \alpha < 2$, no variance!

Introduction
Stylized facts
Broad Issues (BI's)
Approach to Modeling
Tech points
Heavy tails ⇒ LRD
References

2. The number of bits or packets per slot exhibits *long range dependence* across time slots (eg, Leland et al. (1993); Willinger et al. (1995). There is also a perception of *self-similarity* as the width of the time slot varies across a range of time scales exceeding a typical round trip time.

   Note: A stationary process $\{X_n\}$ possesses long range dependence if dependence between variables decays slowly as the gap between the variables increases:

$$|\mathrm{Corr}(X_0, X_h)| \le (const)h^{-\beta}, \quad 0 < \beta < 1.$$

3. Network traffic is *bursty* with rare but influential periods of very high transmission rates punctuating typical periods of modest activity.

   - Bursty is a somewhat ill defined concept associated with heavy tailed transmissions rates.
   - Introduces peak loads to the network.
   - Associated with large files transmitted over fast links.

4. Additional truism: *Traffic at a heavily loaded link per unit time (eg, 1 hour) is normally distributed provided the link is subject to a high degree of aggregation across users.*

3. Broad Issues (BI's)

BI-1: Role of statistics and applied probability:

- Statistics: Empirically identify phenomena and properties of the data so as to better understand what network data *in the wild* should look like. Less emphasis on prediction than in traditional time series analysis.
  - Examples: Identify presence of heavy tails, long range dependence, self-similarity;
  - Understand different statistical properties of various applications and protocols (ftp, http, mail, streaming audio).
- Applied probability: Build models which explain relations and explain empirically observed phenomena.
  - Example: Sizes of files stored on a server follows Pareto power law tail which causes long range dependence.

    Pardigm: Heavy tails cause long range dependence.

  - Build models of end user behavior which allow construction of simulation tools to study effect of tweaking protocols.

- Explain perceived asymptotic normality of quantity of traffic at heavily loaded hub.

BI-2: Problem of time scales: Can Applied Math, Operations Research, Applied Prob & Statistics make contributions to data network analysis and planning in *Internet time*.

- Developers have short attention spans and little patience with outsider's toys.
- Two year time horizon to write a PhD thesis and really understand something is ridiculously long time horizon for industry.
- Pessimists view: the best the mathy community can do is to cause paradigm shifts with explanations which may lag behind developments.
- Start-up mentality.
  - Take the money and run mentality.
  - "*Anyone who gets a PhD does not understand economics*."
  - Long development time for a project means other people are earning.

4. An approach to modeling: Infinite source Poisson model

Instead of passing bits, imagine pouring liquid. Suppose *sessions* characterized by

- Session initiation times are $\{\Gamma_k\}$ where
$$\{\Gamma_k\} \sim \text{ homogeneous Poisson on } (-\infty, \infty), \text{ rate } \lambda.$$
- Sequence of iid marks independent of $\{\Gamma_k\}$: Each Poisson point Γ_k receives a *mark* which characterizes input characteristics:
$$(S_k, D_k, R_k) = (\text{file}, \text{duration}, \text{rate}),$$
where
$$S_k = D_k R_k.$$
All three quantities are often empirically seen to be marginally heavy tailed:
$$\begin{aligned} P[S > x] &\sim x^{-\alpha_S} L_{\alpha_S}(x) \\ P[D > x] &\sim x^{-\alpha_D} L_{\alpha_D}(x) \\ P[R > x] &\sim x^{-\alpha_R} L_{\alpha_R}(x), \end{aligned}$$
with (usually) $1 < \alpha_S, \alpha_R, \alpha_D < 2$.

Examples of mark (S_k, D_k, R_k) structures:

- Early simple models assumed constant input rates,

$$R_k = 1$$

so that total input rate at time t is

$$M(t) = \ \# \text{ active sources at time } t.$$

- Random but constant rates R_k during the transmission interval. (Can even make rate time varying.) Possibly,
  * (S, D) jointly heavy tailed.
  * (R, S) satisfies the *conditional extreme value* (CEV) model in which

$$\frac{R - \beta(t)}{\alpha(t)}\Big| S > t$$

    has a limit distribution approximation for large t and S is heavy tailed (Das and Resnick, 2011).
  * $S \perp\!\!\!\perp R$ (Hernández-Campos et al., 2005) ?
  * $D \perp\!\!\!\perp R$ (Maulik et al., 2002)..
  * Mixture of the previous 2 cases.
  * Some asymptotic form of independence. (Danger: Hard to do mathematical modeling without actual independence.)

- Why is it difficult to decide on the the joint distribution of (S, D, R)? Statistical studies are inclusive since
  - Different types of data analyzed.
    - file sizes
    - response
    - connections
    - session of packets
  - Different amalgamation and segmentation rules. Eg, cluster packets into same *session* if
    - packets have same

      $$(\text{source ip addr, destination ip addr})$$

      and arrive within 2 seconds of each other.
    - packets have same

      $$(\text{source ip addr, destination ip addr})$$

      and same

      $$(\text{source port \#, same destination port \#})$$

      and arrive within 2 seconds of each other.
  - Different applications might have different statistical characteristics (streaming media vs http).

* Different protocols (TCP vs UDP) might have different statistical characteristics.
  - TCP, being subject to a control, tends to produce light tailed flows.
  - UDP (eg, streaming, voip) produces more heavy tailed flows.

– Different distributional assumptions lead to radically different model predictions. Not surprising.

- Fluid input models for cumulative input in time $(a, b]$:

$$A(a, b] = \text{ total work inputted in time interval}(a, b].$$

  - Process approximation to cumulative input.
    - Large time scale approximation

$$d - \lim_{T\to\infty} \frac{A(0, Tt] - b(T)}{a(T)} = X(t),$$

      where possible limits include (Kaj and Taqqu, 2008; Mikosch et al., 2002; Taqqu et al., 1997)
      - fractional Brownian motion (Gauss marginals, lrd)
      - stable Lévy motion (heavy tailed marginals, sii, ss)
      - BUT: not easy to find agreement with these approximate models and data (Guerin et al., 2003)
    - Small time scale approximations (D'Auria and Resnick, 2006, 2008) : Block time into small time slots $(k\delta, (k+1)\delta]$ and consider as $\delta \to 0$

$$\{A(k\delta, (k + 1)\delta], k \in \mathbb{N}\}.$$

      depending on the interaction of input rates and tails. Will need $\lambda = \lambda(\delta) \uparrow \infty$ (a la heavy traffic limit theorems).

- Small time scale approximation results dependent on distributional assumptions on (S, D, R). Either

  D-limit is approximately highly correlated
  normal random variables or
  D-limit is approximately highly dependent
  stable infinite variance random variables.

- Compute dependence measure across different slots. Models predict lrd for $\{A(k\delta, (k+1)\delta], k \in \mathbb{N}\}$

Introduction
Stylized facts
Broad Issues (BI's)
Approach to Modeling
Tech points
Heavy tails ⇒ LRD
References

4.1. Summary: Stylized facts and small time scale approximation

Stylized Facts	$S \perp\!\!\!\perp R$ Model	$D \perp\!\!\!\perp R$ Model
1. Heavy tails	Built in	Built in
2. $P[A(0,\delta] > x] \sim$	$x^{-(\alpha_R+\alpha_S)}$, fixed δ; $x \to \infty$	$x^{-\alpha_R}$, fixed δ; $x \to \infty$
3. LRD across slots	$\mathrm{Cov}(k) \sim c\bar{F}_S^{(0)}(k)$; fixed δ; $k \to \infty$	$\mathrm{EDM}(k) \sim c\bar{F}_D^{(0)}(k)$; fixed δ; $k \to \infty$
4. Cum traffic/slot is $N(0,1)$?	$\dfrac{(A(0,\delta] - (\text{ctering}(\delta))}{a(\delta)} \overset{d}{\approx} N(0,1)$	$\dfrac{(A(0,\delta] - (\text{ctering}(\delta))}{b(\delta)} \overset{d}{\approx} X_{\alpha_R}(\cdot)$

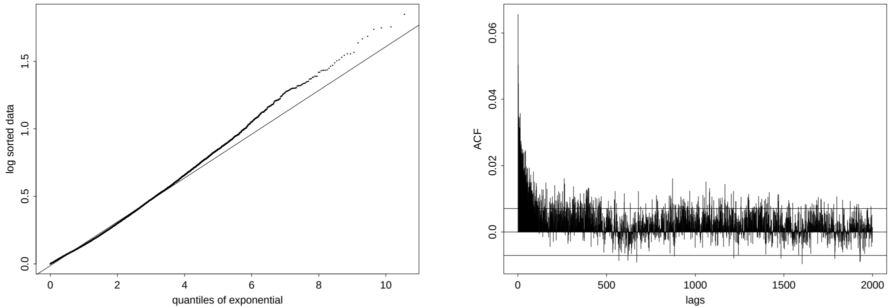
For more, ask Bernardo.

5. Some Technical Points

Tech Pt 1: Identify in the data (connections, flows, packets, sessions) a subset of points that can be modelled as **Poisson time points** and validate the choice.

- Quick & dirty (Q&D) solution: Check if interpoint distances are iid (sample acf almost 0) and exponentially distributed (qq-plots).
- Q&D Rules of Thumb:
  - Behavior of lots of humans acting independently is often well modelled by a Poisson process.
  - Starting times of machine triggered downloads cannot be modelled as Poisson process.

Example: UCB: Inter-arrival times of requests in http sessions via telephone modem (last century).



UCB data; http sessions via modem. *Left*: qqplot against exponential distribution. *Right*: autocorrelation function of interarrival times.

Introduction
Stylized facts
Broad Issues (BI's)
Approach to Modeling
Tech points
Heavy tails ⇒ LRD
References
Title Page
Page 16 of 33
Go Back
Full Screen
Close
Quit

Tech Pt 2: **Heavy tails**: A rv X has a heavy (right) tail if

$$P[X > x] \sim x^{-\alpha}, \quad x \to \infty.$$

Notes

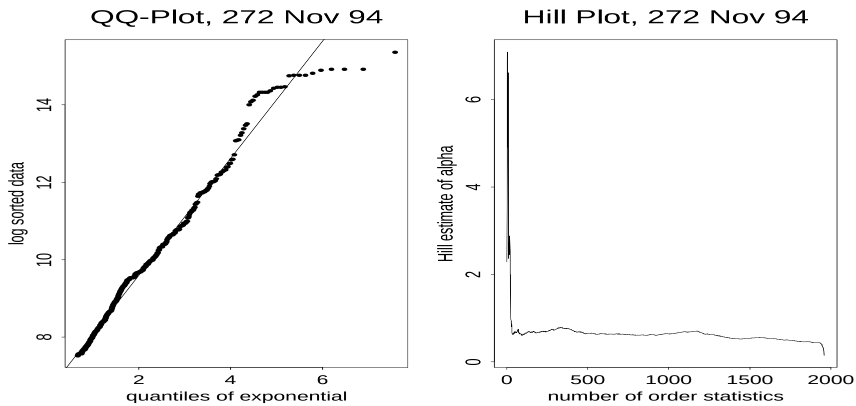
- $0 < \alpha < 1$: Very heavy: mean & variance infinite.
- $1 < \alpha < 2$: Heavy: Frequent case where mean finite but variance is infinite.
- $\alpha > 2$: Heavy with finite variance: Typical of financial data.
- For large x,

  $$P[\log X > x] \sim e^{-\alpha x}, \quad x \to \infty.$$

  So inference can be based on exponential density and thresholding techniques to account for the distribution following this law only for large x.
- For many purposes, do not need to know the whole distribution but just the tail.

Example: BU data: Influential study from mid '90's: File sizes downloaded in a web session.

Figure: Sizes of www downloads; BU experiment: QQ-plot and Hill plot.



Tech Pt 3. Checking for **independence**.

- Q&D method 1: Standard time series method checks if sample correlation function

$$\hat{\rho}(h) = \frac{\sum_{i=1}^{n-h}(X_i - \bar{X})(X_{i+h} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad h = 1, 2, \dots,$$

  is **close** to identically 0.

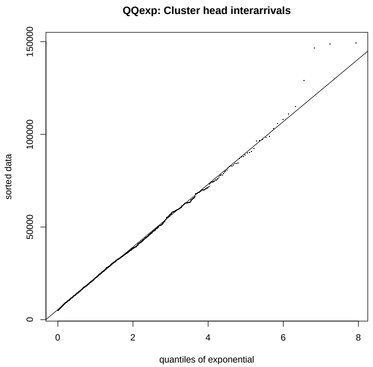
  How to put meaning to phrase **close** to 0? If

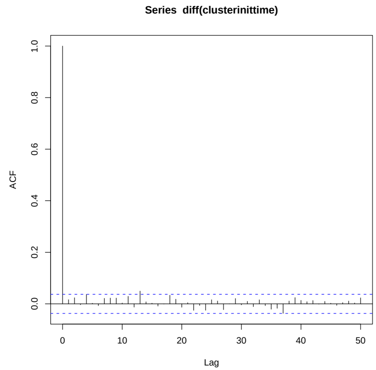
  - If finite variances, Bartlett's formula provides asymptotic normal theory.
  - If heavy tailed, Davis and Resnick formula provides asymptotic distributions for $\hat{\rho}(h)$.

- Q&D method 2: If data heavy tailed, take a function of the data (say the log) to get lighter tail and test. (But this may obscure the importance of large values.)
- Q&D method 3: Subset method. Split data into (say) 2 subsets. Plot acf of each half separately. If iid, pics should look same.

Example: UNC connection data. Data contains 173604 connection vectors ordered by the connection initiation times. Clusters are obtained by considering only the connection start times ordered as they appeared in time on the UNC link and arbitrarily using a time threshold of 5 milliseconds to separate clusters. This yields 16417 clusters.

Do the cluster heads look Poisson?



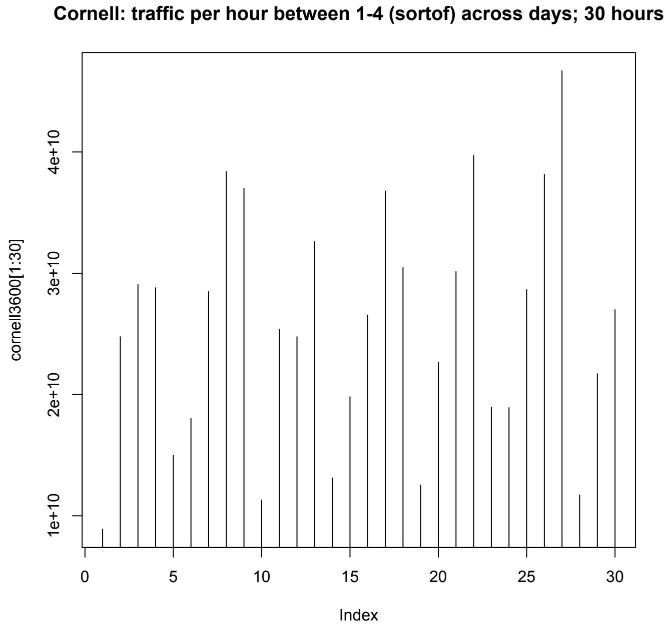


Tech Pt 4. Is the data **stationary**? Usually not and there are, for example, diurnal cycles. (There is only so much Red Bull (Jolt, Coke, ...) that a human can consume.)

- Q&D Coping Method: Take a slab of the copius data which looks stationary.
- Usually there is too much data. (!!??)
  - Unpleasant truth:

    Making the data set sufficiently large allows one to reject any hypothesis.
- Rule of thumb: With high resolution data, don't take more than 4 hours. Depending on the data, it can be a couple of minutes; eg connection data.
- Should we try to model the cycles?

But: Cornell hourly traffic between 1-4pm show buildup for the day. So have trend within day.


Cornell: traffic per hour between 1-4 (sortof) across days; 30 hours
cornell3600[1:30]
1e+10
2e+10
3e+10
4e+10
0
5
10
15
20
25
30
Index

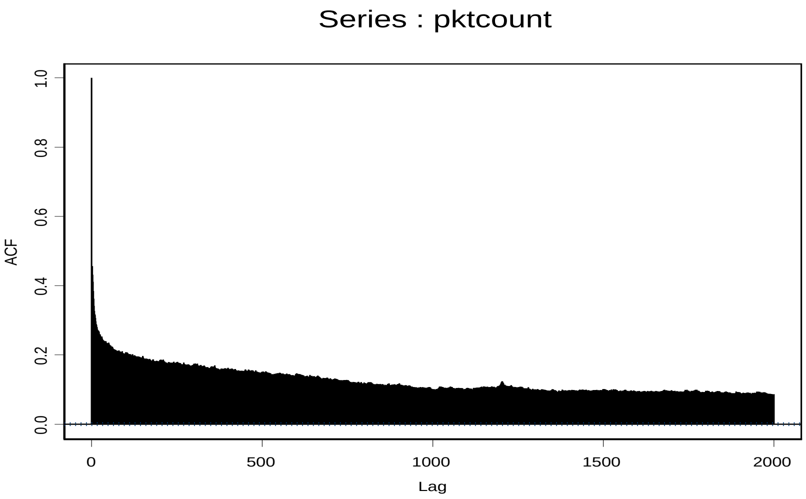
Tech Pt 5. **Long range dependence**. A stationary L_2 sequence $\{\xi_n, n \geq 1\}$ has long-range dependence if

$$\mathrm{Cov}(\xi_n, \xi_{n+h}) \sim h^{-\beta}, \quad h \to \infty$$

for $0 < \beta < 1$.

How to test? Q&D method: The sample acf should not $\to 0$ quickly as the lag increases.

EXAMPLE: CompanyX–packet counts per unit time on CompanyX's WAN (including trans-Atlantic traffic).


Series : pktcount
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Lag

6. How do heavy tails cause long range dependence?

Assume infinite source Poisson model with

$$R_k = 1 \qquad \Rightarrow \qquad S_k = D_k.$$

$$1 < \alpha = \alpha_S < 2, \quad \bar{F}_S \in RV_{-\alpha}, \quad \bar{F}_S(x) = x^{-\alpha}\ell(x).$$

Recall

$$\begin{aligned} M(t) &= \#\text{ active sources at time } t, \\ &= \text{total input rate at time } t, \\ &= \text{analogue of packet count per unit time.} \end{aligned}$$

Background and warmup:

For each fixed t, $M(t)$ is a Poisson random variable.
Why? When $1 < \alpha_S < 2$, $M(\cdot)$ has a stationary version. Assume

$$\sum_k \epsilon_{\Gamma_k} = \mathrm{PRM}(\lambda dt)$$

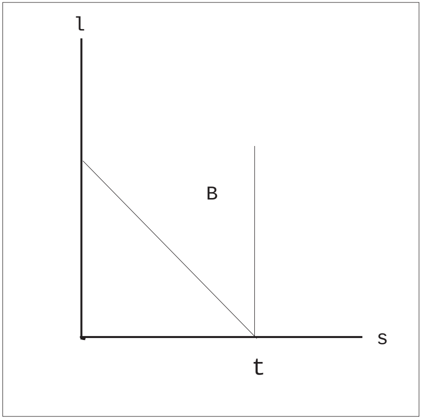
on $\mathbb{R}$. Then

$$\xi := \sum_k \epsilon_{(\Gamma_k, D_k)} = \mathrm{PRM}(\lambda dt \times F_S)$$

Introduction
Stylized facts
Broad Issues (BI's)
Approach to Modeling
Tech points
Heavy tails ⇒ LRD
References

on $\mathbb{R} \times [0, \infty)$ and

$$M(t) = \sum_k 1_{[\Gamma_k \le t < \Gamma_k + D_k]}$$
$$= \xi(\{(s,l) : s \le t < s + l\} = \xi(B)$$



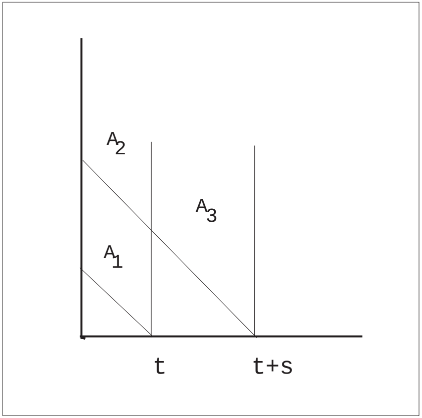
is Poisson with mean

$$E\Big(\xi(\{(s,l) : s \le t < s + l\}\Big)$$
$$= \int_{s=-\infty}^{t} \bar{F}_S(t-s)\lambda ds = \lambda\mu_S$$

The process $\{M(t), t \in \mathbb{R}\}$ is stationary with covariance function

$$\begin{aligned}&\mathrm{Cov}(M(t), M(t+s))\\ &\quad =\mathrm{Cov}\big(\xi(A_1)+\xi(A_2), \xi(A_2)+\xi(A_3)\big)\end{aligned}$$

and because $\xi(A_1) \perp\!\!\!\perp \xi(A_3)$, this is



$$\begin{aligned}&=\mathrm{Cov}\big(\xi(A_2), \xi(A_2)\big)\\ &=\mathrm{Var}\big(\xi(A_2)\big)\\ &=E\big(\xi(A_2)\big)\\ &=\int_{u=-\infty}^{t} \lambda du\, \bar{F}_S(t+s-u)\\ &=\lambda \int_{s}^{\infty} \bar{F}_S(v)dv\\ &\sim\lambda s \bar{F}_S(s)\cdot c = c' s^{-(\alpha-1)}\ell(s).\end{aligned}$$

The slow decay of the covariance as a function of the lag s characterizes *long range dependence*. □

7. References

Contents

Introduction
Stylized facts
Broad Issues (BI's)
Approach to Modeling
Tech points
Heavy tails $\Rightarrow$ LRD
References

References

M. Arlitt and C.L. Williamson. Web server workload characterization: The search for invariants (extended version). In *Proceedings of the ACM Sigmetrics Conference, Philadelphia, Pa*, 1996. Available from {mfa126,carey}@cs.usask.ca.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. In *Proceedings of the ACM SIGMETRICS '96 International Conference on Measurement and Modeling of Computer Systems*, volume 24, pages 160–169, 1996.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

B. Das and S.I. Resnick. Detecting a conditional extreme value model. *Extremes*, 14(1):29–61, 2011.

B. D'Auria and S.I. Resnick. Data network models of burstiness. *Adv. in Appl. Probab.*, 38(2):373–404, 2006.

B. D'Auria and S.I. Resnick. The influence of dependence on data network models. *Advances in Applied Probability*, 40(1):60–94, 2008.

D.E. Duffy, A.A. McIntosh, M. Rosenstein, and W. Willinger. Analyzing telecommunications traffic data from working common channel signaling subnetworks. In M.E. Tarter and M.D. Lock, editors, *Computing Science and Statistics Interface, Proceedings of the 25th Symposium on the Interface*, volume 25, pages 156– 165, San Diego, California, 1993.

C.A. Guerin, H. Nyberg, O. Perrin, S.I. Resnick, H. Rootzén, and C. Stărică. Empirical testing of the infinite source poisson data traffic model. *Stochastic Models*, 19(2):151–200, 2003.

F. Hernández-Campos, J.S. Marron, C. Park, S.I. Resnick, and K. Jaffay. Extremal dependence: Internet traffic applications. *Stochastic Models*, 21(1):1–35, 2005.

I. Kaj and M. Taqqu. Convergence to fractional brownian motion and to the telecom process: the integral representation approach. In V. Sidoravicius and M. Eulália Vares, editors, *In and Out of Equilibrium 2*, volume 60 of *Progress in Probability*. Birkhäuser Basel, 2008.

W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. Statistical analysis of high time-resolution ethernet Lan traffic measurements. In *Proceedings of the 25th Symposium on the Interface between Statistics and Computer Science*, pages 146– 155, Fairfax Station, Virginia, 1993. Interface Foundation of North America.

W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2(1):1–15, 1994. ISSN 1063-6692. doi: http://dx.doi.org/10.1109/90.282603.

K. Maulik, S.I. Resnick, and H. Rootzén. Asymptotic independence and a network traffic model. *J. Appl. Probab.*, 39(4):671–699, 2002. ISSN 0021-9002.

T. Mikosch, S.I. Resnick, H. Rootzén, and A.W. Stegeman. Is network traffic approximated by stable Lévy motion or fractional Brownian motion? *Ann. Appl. Probab.*, 12(1):23–68, 2002.

S.I. Resnick. Modeling data networks. In B. Finkenstadt and H. Rootzén, editors, *SemStat: Seminaire Europeen de Statistique, Extreme Values in Finance, Telecommunications, and the Environment*, pages 287–372. Chapman-Hall, London, 2003.

S.I. Resnick and H. Rootzén. Self-similar communication models and very heavy tails. *Ann. Appl. Probab.*, 10:753–778, 2000.

M.S. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *Computer Communications Review*, 27:5–23, 1997.

W. Willinger. Data network traffic: heavy tails are here to stay. Presentation at EXTREMES–RISK AND SAFETY, Nordic School of Public Health, Gothenberg Sweden, August 1998.

W. Willinger and V. Paxson. Where mathematics meets the Internet. *Notices of the American Mathematical Society*, 45(8):961–970, 1998.

W. Willinger, M.S. Taqqu, M. Leland, and D. Wilson. Self–similarity in high–speed packet traffic: analysis and modelling of ethernet traffic measurements. *Statist. Sci.*, 10:67–85, 1995.

W. Willinger, M.S. Taqqu, M. Leland, and D. Wilson. Self–similarity through high variability: statistical analysis of ethernet lan traffic at the source level (extended version). *IEEE/ACM Transactions on Networking*, 5(1):71–96, 1997.

W. Willinger, V. Paxson, and M.S. Taqqu. Self-similarity and heavy tails: Structural modeling of network traffic. In R.J. Adler, R.E. Feldman, and M.S. Taqqu, editors, *A Practical Guide to Heavy Tails. Statistical Techniques and Applications*, pages 27–53. Birkhäuser Boston Inc., Boston, MA, 1998.