# Modeling Data Network Sessions

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

APS Stockholm

http://people.orie.cornell.edu/∼sid
sir1@cornell.edu

July 7, 2011

**Work with:** Luis Lopez-Oliveros (Cornell), B. Das (ETH), J. Heffernan, Bernardo D'Auria (Madrid), T. Mikosch, H. Rootzen, A. Stegeman

# 1. Introduction: Network Input Models Based on Sessions.

*Session:* higher order construct typically obtained by amalgamating either

- packets (see Part II)
- connections
- flows or reponses (one way connections)
- groups of connections.

according to some *chosen* – and undoubtedly not unique– definition.

Search for (or assume)

- Poisson structure
- Distributional descriptors of sessions.

Hope (or assume)

- Sessions are independent.

## Which hat are you wearing?

- Mathematical modeler hat (Part I):
  - Superimpose sessions,

$$A(0,t] = \text{total end-user input from sessions in } (0,t]$$

    and construct limit theorems to explain observed features of "traffic".

    * Large time scale distributional process limits for

$$A(0,Tt] \qquad \text{as } T \to \infty.$$

    * Small time scale distributional limits for

$$\{A((k-1)\delta, k\delta], k \geq 0\}, \qquad \text{as } \delta \downarrow 0,$$

    and understanding of dependence structure.

- Empirical modeler hat (Part II)
  - Analyze Cornell netflow data.
  - In contrast to netflows, for packet level modeling, no obvious definition of session.
    - Must amalgamate packets into sessions according to some rule. Usually the rule is a threshold rule, sometimes time based.
    - For example: *successively arriving packets are in the same session if arrival times are within 2 seconds of each other.*
  - See what properties are empirically present?
  - Heavy tail analysis to understand statistical characteristics.
  - Search for Poisson time points in the mess.

# 2. Part I: Modeling Approach

Sessions characterized by

- Initiation times $\{\Gamma_k\}$ where often assume for tractability

$$\{\Gamma_k\} \sim \text{ Poisson, rate } \lambda.$$

- The mark or session descriptor of $\Gamma_k$:

$$(S_k, D_k, R_k) = (\textbf{S}\text{ize}, \textbf{D}\text{uration}, \textbf{R}\text{ate }),$$

  where $R$ is average rate in session:

$$R_k = S_k / D_k.$$

- Approach: Consider a family of models indexed by $T \to \infty$ or $\delta \to 0$; move through the family by taking a limit on $T$ or $\delta$ and seek a limit model providing a useful approximation. Consider either:

- Large time scale approximations; model family indexed by $T$ and $T \to \infty$:

$$d - \lim_{T\to\infty} \frac{A(0,Tt] - b_T(t)}{a_T(t)}$$

  and will permit $\lambda = \lambda(T) \to \infty$.
- Small time scale approximations: $A(k\delta,(k+1)\delta]$, the work inputted in $(k\delta,(k+1)\delta]$.

  Approximation as $\delta \to 0$? Will need $\lambda = \lambda(\delta) \uparrow \infty$ (a la heavy traffic limit theorems).

- Compute dependence measures.
- Results dependent on distributional assumptions on $(S, D, R)$.

## 2.1. Difficulties

1. What is the joint distribution of $(S, D, R)$? Statistical studies are inconclusive since

   - Different types of data analyzed.

     – file sizes

     – responses

     – flows

     – connections

     – packet headers

   - Sessions constructed using different amalgamation rules: eg,

     – amalgamate packets using 2s threshold rule when

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

     same?

     – Or amalgamate using 2s threshold rule when

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

     same.

- Segmenting sessions according to various rules leads to different statistical characteristics.
  - Segment by peak rate.
  - Segment by application (HTTP, SMTP, FTP, VOIP, streaming, … ).
  - Segment by protocol (TCP, UDP, … ).
- Different types of data and collection methods may be heavily or lightly influenced by censoring. Some sessions start before the collection window and some end after. Take into account or ignore?

2. Distributional possibilities (remember $S = DR$ so any two determine the third):

- $(S, D)$ 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.
- $S \perp\!\!\!\perp R$ (Hernández-Campos et al., 2005) ?
- $D \perp\!\!\!\perp R$ ?
- Mixture of the previous 2 cases?
- Some asymptotic form of independence?
  - Possibly no pair of $(S, D, R)$ is truly independent in practice.
  - BUT: For mathematical analysis of model, asymptotic independence rather than independence increases the cost in complexity?

3. Oversimplifications:

   (a) Connection initiations–as opposed to sessions–may be better modeled as point process with clusters where a primary connection ("get NY Times") triggers subsidiary connections ("get ads, get pics, get text");

   - cluster structure may depend on definitions of session.
   - the statistical structure of clusters is not obvious.

   (b) Transmission rate is not constant over the connection interval.

   (c) Typically, these models only of end-user behavior and do not model the influence of congestion and queueing; but applicable to studying things like protocol design. (BUT: certain types of data reflect effect of protocols.)

   (d) Empirical studies typically ignore issue of censoring; this issue will increase in importance with the ubiquity of streaming.

## 2.2. Assessing dependence structure of $(S, D, R)$.

Consider

$$\begin{aligned} S &= \text{file size, total info in session, \# bytes} \\ D &= \text{duration of transmission or session,} \\ R &= \text{throughput, } \textit{average} \text{ rate } = S/D. \end{aligned}$$

All three, are usually (alas–not always) empirically seen to be heavy tailed:

$$\begin{aligned} P[S > x] &= x^{-\alpha_S}\ell_S(x) \\ P[D > x] &= x^{-\alpha_D}\ell_D(x) \\ P[R > x] &= x^{-\alpha_R}\ell_R(x). \end{aligned}$$

Note:

1. Segmenting data in various ways (eg by peak rate) may preclude certain segments from having $R$ heavy tailed.
2. For some data, the effect of censoring is noticeable since there are multiple values of duration equal to the length of collection window.

# 3. Small time scale results

## 3.1. $S \perp\!\!\!\perp R$: Assumptions

Assume:

- Sessions begin: $\{\Gamma_k\}$, homogeneous Poisson rate $\lambda$.
- For the $k$-th session, independently attach iid marks $(S_k, D_k, R_k)$; $S, R$ independent, heavy tailed; $S = DR$,

$$S \sim F_S(x) \quad R \sim F_R(x).$$

- $1 < \alpha_S, \alpha_D, \alpha_R < 2$; finite means, infinite 2nd moments.
- Distribution tail of $D$ given by

$$\bar{F}_D(l) \sim \mathbb{E}\left(\frac{1}{R}\right)^{\alpha_S} \bar{F}_S(l),$$

provided assume (Breiman (1965))

$$\mathbb{E}\left[\frac{1}{R}\right]^{\alpha_S+\eta} < \infty,$$

for some $\eta > 0$. (Other assumptions possible!)

- Time slots $(k\delta, (k+1)\delta], k = 0, \pm 1, \pm 2, \dots\}$.

- Limiting procedure shrinks the observation windows ($\delta \to 0$). To get limit, increase the arrival rate $\lambda = \lambda(\delta)$ of sessions.
- Heavy traffic limit theorem philosophy; move through a family of models indexed by $\delta$ as $\delta \downarrow 0$. Choice of $\lambda$:

$$\lambda(\delta) = \frac{1}{\delta \bar{F}_R(\delta^{-1})}.$$

- Since $1 < \alpha_R < 2$, this choice of $\lambda$ guarantees as $\delta \to 0$,

$$\lambda(\delta) = \frac{1}{\delta^{\alpha_R+1}\ell_R(\delta^{-1})} \to \infty \quad \text{and} \quad \delta\lambda(\delta) = \frac{1}{\delta^{\alpha_R}\ell_R(\delta^{-1})} \to \infty.$$

- Seek limit behavior of

$$\mathbf{A}(\delta) := \{A(k\delta, (k+1)\delta], -\infty < k < \infty\}$$

where

$$A(k\delta, (k+1)\delta] = \text{ work inputted in time } (k\delta, (k+1)\delta],$$

as

  – $\delta \to 0$, OR
  – $\delta$ is fixed and we study $\mathrm{Cov}(A(0,\delta], A(k\delta,(k+1)\delta])$ as $k \to \infty$ to seek LRD.

## 3.2. $D \perp\!\!\!\perp R$: Model Assumptions:

- Same choice of Poisson rate $\lambda$ as before:

$$\lambda(\delta) = \frac{1}{\delta \bar{F}_R(\delta^{-1})}.$$

- The file size distribution $F_S$ is determined by $F_D$ and $F_R$.
- Tails of traffic sum is heavier than when $S \perp\!\!\!\perp R$. Now for fixed $\delta$, as $x \to \infty$,

$$P\big[A(\delta) > x\big] \sim c\bar{F}_R(x).$$

- Oops! Infinite 2nd moment for $A(0, \delta]$, $A(k\delta, (k+1)\delta]$
  - Cannot use correlations for assessing decay of dependence as $k \to \infty$.
  - Used EDM to measure LRD.

## 3.3. Summary: Compare model outcomes from two assumptions

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

# 4. Large time scale modeling

Kaj and Taqqu (2008), Konstantopoulos and Lin (1998), Mikosch et al. (2002), Taqqu et al. (1997)

Sessions characterized by

- Initiation times

$$\{\Gamma_k\} \sim \text{ homogeneous Poisson on } (-\infty, \infty), \text{ rate } \lambda.$$

- Marks for Poisson points $\{(D_k, R_k)\}$. Assume $\{R_k\} \perp\!\!\!\perp \{D_k\}$ and $R_k$ iid with finite 2nd moment and $\{(D_k, R_k)\} \perp\!\!\!\perp \{\Gamma_k\}$ and session length distribution has heavy tail:

$$\mathbb{P}(D_k > x) = \bar{F}_D(x) = x^{-\alpha}\ell(x), \quad x > 0,\ 1 < \alpha < 2\,, \qquad (1)$$

  where $\ell$ is a slowly varying function. Then $\{(\Gamma_k, D_k, R_k)\}$ are points of a 3-dimensional Poisson process.

  Simplest case: Assume

$$R_k \equiv 1 \text{ so } S_k = D_k \text{ and } 1 < \alpha_S = \alpha < 2.$$

  Assumed to make tutorial more comprehensible.

- Number of sessions in progress at $t$

$$M(t) = \sum_{k=1}^{\infty} 1_{[\Gamma_k \le t < \Gamma_k + D_k]}$$

- Cumulative input in $[0,t]$ if $R \equiv 1$,

$$A(t) = \int_0^t M(s)ds.$$

- $\alpha \in (1,2)$ means the variance of $D_k$ is infinite and its mean $\mu_D$ is finite.

- Quantile function

$$b(t) = \left(\frac{1}{1-F_D}\right)^{\leftarrow}(t) =: \inf\{x : \frac{1}{1-F_D(x)} \geq t\}, \quad t > 0, \quad (2)$$

which is regularly varying with index $1/\alpha$.

- Scale time in cumulative input by $T \to \infty$ and if necessary allow $\lambda = \lambda(T) \to \infty$ (at what rate?) and consider for the $T$th model:

$$A_T(t) = A(Tt).$$

- Family of models indexed by $T$. As we move through the family by letting $T \to \infty$, is there an informative limit?

## Growth conditions

- $\lambda = \lambda(T)$: the parameter governing the initiation of sessions in the $T$th model;
- Suppose $\lambda = \lambda(T)$ is non-decreasing function of $T$.
- Asymptotic behavior of $A_T(\cdot)$ is very different depending on whether
  - *slow*,
  - *moderate* or
  - *fast growth*

  holds for $\lambda(T)$.

**Lemma 1 (Slow/moderate/fast growth.)** *Suppose for simplicity* $R_k \equiv 1$ *and*

- $\bar{F}_D \in RV_{-\alpha}$, $1 < \alpha < 2$; *the quantile function of* $F_D$ *is*

$$b(t) = \left(\frac{1}{1 - F_D}\right)^{\leftarrow}(t).$$

- *Session initiations in the* $T$*th model form a stationary PRM on* $\mathbb{R}$, *rate* $\lambda = \lambda(T)$.

*Then* $M_T(t)$, *the number of active sources at time* $t$ *in the* $T$*th model is a stationary process on* $\mathbb{R}$ *and the following are equivalent and define slow growth:*

1.
$$\lim_{T\to\infty} \frac{b(\lambda T)}{T} = 0,$$

2.
$$\lim_{T\to\infty} \lambda T\, \bar{F}_D(T) = 0,$$

3.
$$\lim_{T\to\infty} \mathrm{Cov}\,\left(M_T(0), M_T(T)\right) = 0.$$

*Remark: If* $\lambda(T)$ *is constant or bounded, then always have slow growth.*

*Fast growth*: *replace* 0 *by* $\infty$.
*Moderate growth*: *replace* 0 *by* $0 < c < \infty$.

## Very different approximations

Kaj and Taqqu (2008), Konstantopoulos and Lin (1998), Mikosch et al. (2002), Taqqu et al. (1997)

**Theorem 1** *If slow growth holds, then the process* $(A(Tt), t \geq 0)$ *describing the total cumulative input in* $[0, Tt]$, $t \geq 0$, *satisfies the limit relation*

$$X^{(T)}(\cdot) := \frac{A(T\cdot) - T\lambda\mu_D(\cdot)}{b(\lambda T)} \stackrel{fidi}{\rightarrow} X_\alpha(\cdot), \qquad (3)$$

*where*

- $X_\alpha(\cdot)$ *is* $\alpha$*-stable Lévy motion;*
- $\stackrel{fidi}{\rightarrow}$ *denotes convergence of the finite dimensional distributions.*

**Theorem 2** *If fast growth holds, then the process $(A(Tt), t \geq 0)$ describing the total accumulated input in $[0, Tt]$, $t \geq 0$, satisfies the limit relation*

$$\frac{A(T\cdot) - \lambda\mu_D T(\cdot)}{[\lambda T^3 \bar{F}_D(T)\sigma^2]^{1/2}} \stackrel{d}{\to} B_H(\cdot)\,,$$

*where*

- $\stackrel{d}{\to}$ *denotes weak convergence in* $(\mathbb{D}[0,\infty), J_1)$,
- $B_H$ *is standard fractional Brownian motion,*
- $H = (3-\alpha)/2$
- $\sigma^2$ *is a constant.*

See Gaigalas and Kaj (2003), Kaj (1999, 2002), Kaj and Taqqu (2008), Konstantopoulos and Lin (1998), Mikosch and Stegeman (1999), Mikosch et al. (2002), Resnick (2003), Resnick and van den Berg (2000).

# 5. Part II: Empirical–Statistical Modeling

## 5.1. Does traffic ever look stable?

- Question: Do you ever see "traffic" that looks like stable Lévy motion?
- Answer: No. (Guerin et al., 2003)
- Network invariant:

  > "Traffic at a heavily loaded link which is sufficiently aggregated across users looks Gaussian."

- I have never met a networking person who believes traffic could be stable and at networking conferences have gotten several hostile reactions to the suggestion that it might be.
- Theory says traffic could look stable so how does one explain the lack of sightings?
- What happens to different streams, some of which could be Gaussian and some stable, when they are combined?

# 6. Cornell Netflow Data

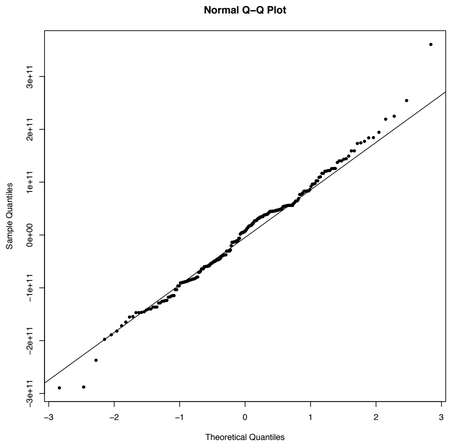
The traditional Cisco definition (simplified) of a network flow:

> a uni-directional sequence of packets with the same source and destination IP address, source and destination port, protocol (TCP, UDP, ...).

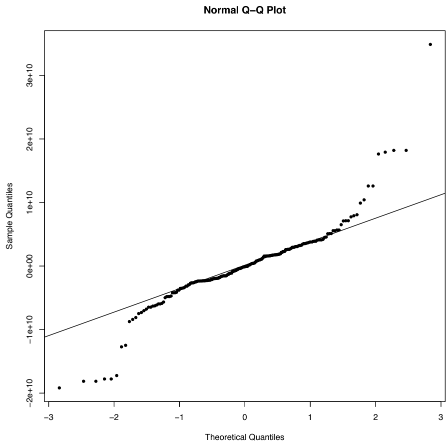
- So each flow generates a datum: (source ip, destination ip, source port, destination port, protocol (TCP, UDP, ...), ..., #bytes in flow, start time, end time).
- 55 days, 4 hours per day, 1-5pm, time rounded to seconds by collector.
- We separated flows into TCP flows and UDP flows.
- TCP=web, mail, ftp, ssh, some p2p, ... —source send rate restricted by send/ack system and if ack doesn't come back in a time threshold, send rate halved.
- UDP= streaming media, some file sharing, voip; no control; sender just sends. Such flows may be long.
- In Cornell data, TCP $\approx 68\%$ of bytes but $> 80\%$ of flows.

## Analysis of data: Segment by TCP and UDP.

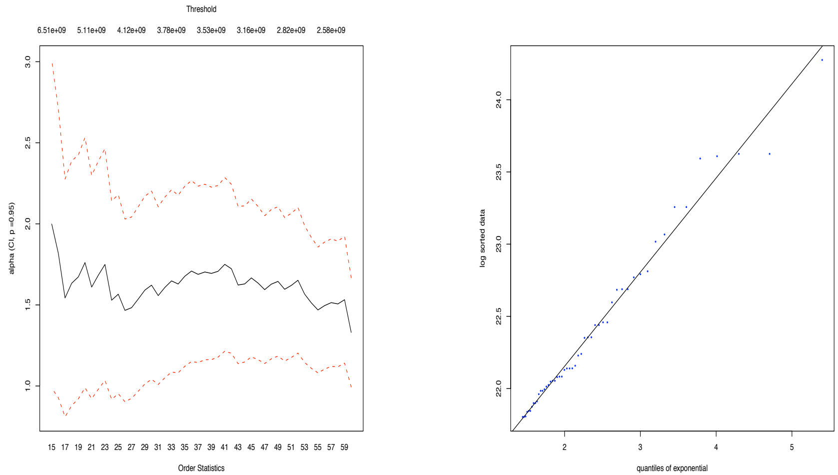
- $A_k^{(TCP)}$ and $A_k^{(UDP)}$ = the total number of TCP and UDP bytes in the $k$th hour, $k = 1, \ldots, 220$.
- Detrend and remove daily seasonality. (With or without this massage the conclusions are the same.)
- $A_k^{(TCP)}$, $k = 1, \ldots, 220$ looks Gaussian.
- $A_k^{(UDP)}$, $k = 1, \ldots, 220$ looks heavy tailed.
- Combining the flows creates traffic that looks Gaussian.



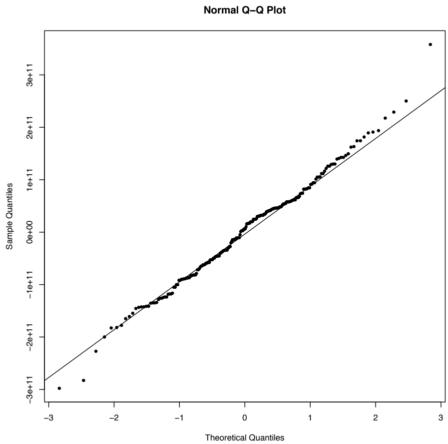

Normal QQ plot of TCP hourly data; Anderson-Darling test p-value: 0.1369. Reminder: If $p$-value $> 0.05$, do not reject hypothesis of normality at level 0.05.




Normal QQ plot of UDP hourly data; Anderson-Darling test p-value: $9.18 \times 10^{-16}$.




Hourly UDP: Left: Hill plot; right QQ plot of log data with line through the biggest 55 points.




$A^{(TCP)} + A^{(UDP)}$: Anderson-Darling test p-value: 0.2117.

## The Lopez-Oliveros & Resnick Explanation:

Suppose there are $p$ independent streams. Define

- For $j$th stream, session arrival rate $\lambda^{(j)} = \lambda^{(j)}(T)$, $j = 1, \ldots, p$;
- Cumulative arrival rate $\lambda = \sum_{j=1}^{p} \lambda^{(j)}$,
- Each stream has regularly varying tail probabilities for session durations with index $\alpha_D^{(j)}$, $j = 1, \ldots, p$; set

$$\alpha_D := \bigwedge_{j=1}^{n} \alpha_D^{(j)}.$$

- Assume

$$\liminf_{T\to\infty} \bigvee_{j:\alpha_D^{(j)}=\alpha_D} \lambda^{(j)}/\lambda > 0,$$

so that the proportion of traffic with heaviest tailed duration is greater than positive constant.

- Mixture: $F_D := \sum_{j=1}^{p}(\lambda^{(j)}/\lambda)F_D^{(j)}$ with $\bar{F}_D(x) \in RV_{-\alpha_D}$ for each fixed $T$.
- $A^{(j)}(t) :=$ cumulative input of the $j$th stream in $[0, t]$.
- $A(t) :=$ cumulative input in $[0, t]$ aggregated over all $p$ streams.
- Each of the $p$ streams satisfies one of the 3 conditions–fast, moderate, or slow.

## How to get FBM limit and Gaussian Traffic:

**Scenario Fast**: There is at least one stream whose arrival rate satisfies fast growth; implies the aggregated stream's arrival rate also satisfies fast-growth:

$$\lambda T \bar{F}_D(T) \to \infty, \quad T \to \infty.$$

**Conclusion**: The cumulative traffic aggregated over the $p$ streams looks like FBM:

$$\frac{A(Tt) - c_T(t)}{a(T)} \Rightarrow B_H(t), \quad T \to \infty.$$

# So?

In practice port 80 carries HTTP web plus other traffic;

- Dominant application; on typical server $> 80\%$ of connections.
- This dominant input rate $\lambda_{\text{http}}$ leads to *fast growth*.

## Other cases.

**Scenario Slow:** All streams satisfy slow growth, hence

$$\lambda T \bar{F}_L(T) \to 0.$$

**Conclusion**: The cumulative traffic aggregated over the $p$ streams looks like stable Lévy motion:

$$\frac{A(T\cdot) - \lambda \mu_D T\cdot}{a_{\text{Slow}}(T)} \Rightarrow X_\alpha(\cdot), \qquad \text{Scenario Slow.}$$

# 7. Packet Level Modeling: Segmenting Sessions by Peak Rate

Consider packet level modeling from data consisting of *packet headers*. This is quite different from netflow data.

When a file is sent through the Internet:

1. The file is divided into *packets*.
2. Headers: Our data consists of (anonymized) packet *headers*. Packets numbered and labeled with
   - Source and destination IP addresses
   - Source and destination port numbers
   - Packet size
   - Internet protocol (TCP?, UDP?, ...)
   - Time stamp of arrival at sensor.
3. Stored and forwarded through routers.
4. Reassembled into the original file upon delivery.

No information when connections are closed or when flows end. How to construct sessions? Choose a rule such as (Sarvotham et al., 2005):

> A session is a cluster of packets with same source and destination network addresses, such that the delay between any two succesive packets in the cluster is less than a threshold t(= 2s).

## Modest goals:

- Use packet header data to model Internet transmissions.
- Amalgamate packets into higher order entities, or sessions, according to a sensible rule. Such amalgamation should simplify modeling and allow fluid or continuous type applied prob models?
  - Amalgamation rule far from unique.
  - Do you have enough manpower & computer power to explore alternatives?
- Since the Internet behaves partly as a result of human stimulation, hope somewhere in this mess of data there lurk Poisson points. Identify such points. Hope the chosen amalgamation rule is conducive to identification of Poisson points.

- Teach a computer to mimic Internet sessions and hence end user Internet behavior.

## 7.1. Sessions

[Sarvotham et al. (2005)] For this study:

> A *session* is a cluster of packets with same source and destination network addresses, such that the delay between any two succesive packets in the cluster is less than a threshold $t(=2s)$.

Other definitions possible.

## 7.2. Session descriptors:

For each session, compute the following descriptors:

- $S$ : Number of bytes transmitted (size).
- $D$ : Duration of the session.
- $R = S/D$ : Average transfer rate.
- $\Gamma$: Starting time.
- For studying burstiness, some measure of peak rate.

## 7.3. Sample public domain data set.

http://wand.cs.waikato.ac.nz/wits/

- TCP traffic (www, email, FTP)
- Traffic sent to a University of Auckland server on December 8, 1999, between 3 and 4 pm.
- Raw data: 1,177,497 packet headers.
- Harvest working data set of the form $\{(S_i, D_i, R_i, \Gamma_i) : 1 \le i \le 44,136\}$.

Originally used by Sarvotham et al. (2005) to study of sources of burstiness: Burstiness is important in order to understand congestion because of the sudden peak loads it introduces to the network; qos concerns.

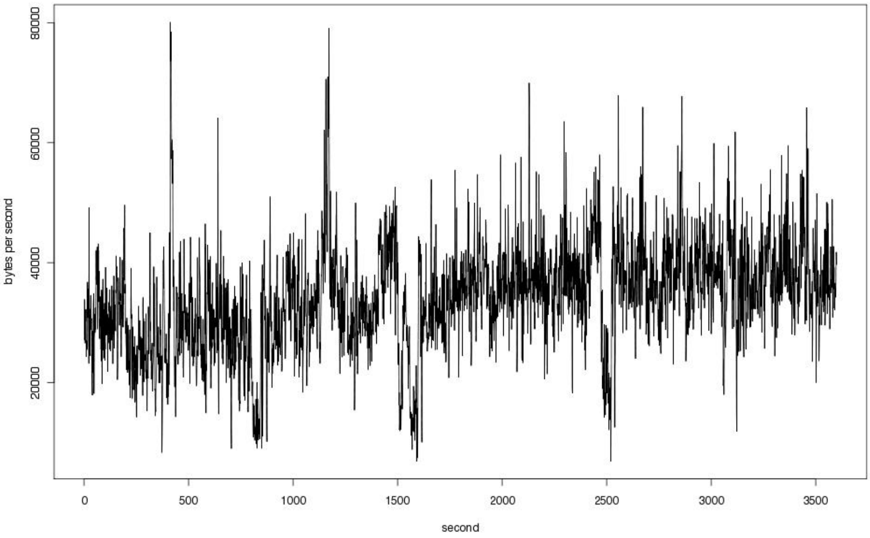



Figure 1: Bytes per time (seconds) process.

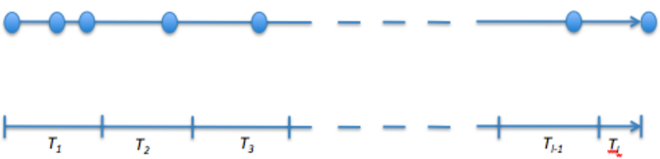
## 7.4. The alpha-beta split

Sarvotham et al. (2005)

**Definition 1 ($\delta$-maximum input)** Pick $\delta > 0$. Divide each session into consecutive intervals of length $\delta$ (being sloppy at end). If a session is thus divided into $l$ subintervals, let

$$B_i = \#\text{ bytes transmitted over the } i\text{th subinterval}, \ i = 1, \dots, l.$$

The $\delta$-maximum input of a session is defined as $M_\delta = \bigvee_{i=1}^{l} B_i$.




**Definition 2 (Alpha-beta split)** Choose a high threshold $u$. A session with a $\delta$-maximum input $M_\delta$ is called

- *alpha*, if $M_\delta \geq u$,
- *beta*, if $M_\delta < u$.

## Empirical features

Sarvotham et al. (2005) found:

- Alpha-sessions are the major source of burstiness.
- In alpha-sessions: $R \perp\!\!\!\perp S$ (sort of).
- In beta-sessions: $R \perp\!\!\!\perp D$ (sort of).
- Split usually produces huge beta-group ($\approx$ tens of thousands) vs. tiny alpha-group ($< 100$).
- Implies segmented sessions have different distribution in each segment.

*sort of* = as measured by correlation.

Does further segmentation of the beta-group produce meaningful information?

Goals:

- Better description of dependence structure of $(S, D, R)$ within segment
- Simulation model.

## 7.5. Finer segmentation

Split the data into $m$ groups of approximately equal size according to the empirical quantiles of the burstiness predictor or covariate; need new definition of peak rate (that actually computes peak rate/session).

Will use $m = 10$ and speak of the decile groups. Split into decile groups.

Features:

- Rather than a beta-group, we have 9 groups each with the peak rate covariate in a given decile range.
- Claim the alpha-beta split masks further structure and it is informative to take into account the explicit level of the covariate.

## 7.6. Peak rate

**Definition 3** For a session with $n$ packets:

$$B_i' = \# \text{ bytes in } i\text{th packet},$$
$$T_i' = \text{interarrival time between } i\text{th and } (i+1)\text{st packet},$$

$i = 1, \ldots, n-1$. For $k = 2, \ldots, n$, the *peak rate of order $k$* is

$$P^{(k)} = \bigvee_{i=1}^{n-k+1} \frac{\sum_{j=i}^{i+k-1} B_j'}{\sum_{j=i}^{i+k-2} T_j'}.$$

The $P^{(k)}$ is the maximum transfer rate using only $k$ consecutive packets.

The peak rate is defined as

$$P^{\vee} = \bigvee_{k=2}^{n} P^{(k)}.$$

[Makes sense empirically but would be difficult to work with analytically.]


CORNELL
Intro
Math modeling
Small time scale results
Large time scale modeling
Part II: Empirical
Cornell Data
Sessions
Distn Structure
Dependence
Finale
Title Page
Go Back
Full Screen
Close
Quit

## Outline

- Divide the 44,136 sessions into 10 groups according to the deciles of $P^{\vee}$.
- Study the marginals of $(S, D, R)$ in the 10 decile groups. (Heavy tails?)
- Study dependence structure of $(S, D)$ using EVT across the decile groups.
- For our definition of peak rate, $P^{\vee}$, within a decile group, data sessions are initiated according to a homogeneous Poisson process.
  - Not true for other peak rate definitions of Sarvotham et al. (2005).]

# 8. Structure of $(S, D, R)$.

## 8.1. Heavy tails

**Definition 4 (Heavy tails)** Call $Y$ has *heavy tailed* if its cdf $F$ satisfies

$$\bar{F}(y) = y^{-1/\gamma}\ell(y),$$

where $\ell$ is slowly varying and $\gamma > 0$.

Quickie summary:

- In each decile group, $(S, D)$ jointly heavy tailed w/o asymptotic independence.

  Jointly heavy tailed means: $\exists$ scaling functions $b_S(t), b_D(t) \to \infty$ and a limit measure $\nu$ s.t.

  $$tP[\Big(\frac{S}{b_S(t)}, \frac{D}{b_D(t)}\Big) \in A] \to \nu(A)$$

  for sets bounded away from $(0, 0)$.

  w/o asymptotic independence means

  $$\nu((0, \infty]^2) > 0.$$

- $R$ is only heavy tailed for the highest decile group;

- $R$ does *not* appear to be even in a domain of attraction for any of the 9 lower decile groups.

## 8.2. Estimation of $\gamma$'s

**Definition 5 (Hill estimator)** Let $\{X_1, \dots, X_n\}$ be iid (or stationary + mixing) with order statistics

$$X_{1:n} \leq \cdots \leq X_{n:n}.$$

The *Hill estimator* of $\gamma > 0$ is

$$\hat{\gamma}_{k,n} = \frac{1}{k} \sum_{i=n-k+1}^{n} \log \frac{X_{i:n}}{X_{n-k:n}}. \tag{4}$$

**Theorem 3 (Consistency of Hill)** *If the distribution is heavy tailed +additional second order condition, as* $k \to \infty, n \to \infty, k/n \to 0$*:*

$$\sqrt{k}(\hat{\gamma}_{k,n} - \gamma) \xrightarrow{d} N(0, \gamma^2). \tag{5}$$

Equivalent to peaks over threshold method and MLE.

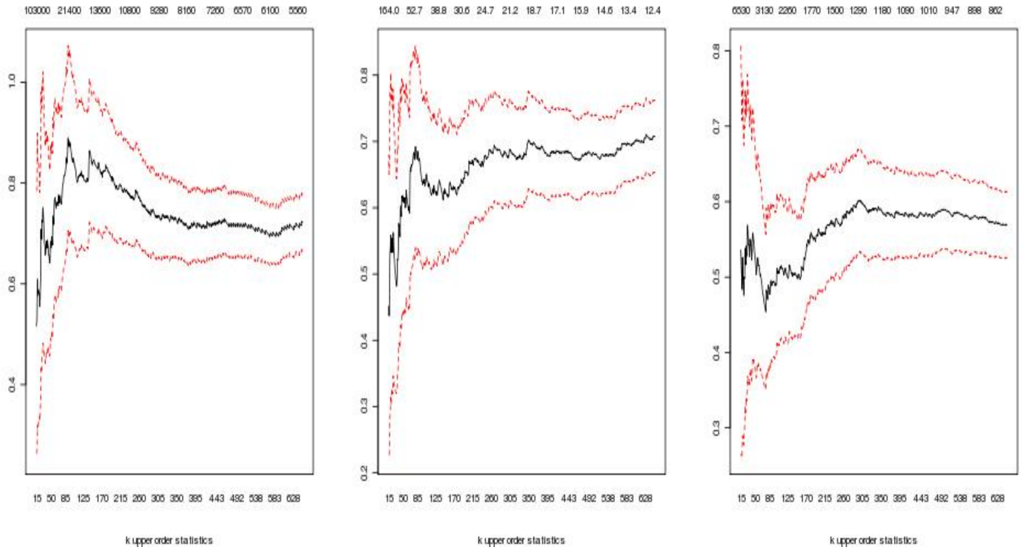



Figure 2: Size, duration and rate in the 10th decile group

| decile | $\gamma_S$ | s.e. | $\gamma_D$ | s.e. | $\gamma_R$ | s.e. |
|---|---|---|---|---|---|---|
| 1 | 0.56 | 0.056 | 0.60 | 0.028 | | |
| 2 | 0.55 | 0.061 | 0.47 | 0.023 | | |
| 3 | 0.62 | 0.044 | 0.63 | 0.034 | | |
| 4 | 0.62 | 0.036 | 0.62 | 0.029 | | |
| 5 | 0.61 | 0.035 | 0.55 | 0.029 | | |
| 6 | 0.69 | 0.040 | 0.55 | 0.028 | | |
| 7 | 0.88 | 0.042 | 0.73 | 0.037 | | |
| 8 | 0.77 | 0.045 | 0.71 | 0.033 | | |
| 9 | 0.70 | 0.037 | 0.69 | 0.032 | | |
| 10 | 0.73 | 0.034 | 0.68 | 0.032 | 0.58 | 0.027 |

Table 1: Summary of Hill estimates with asymptotic standard errors for the shape parameter of $S$, $D$ and $R$.

Conclude: It appears that marginal distributions vary by decile.

# 9. Dependence structure of $(S, D)$

- Dependence structure varies by decile. Seen already in simple scatter plots.
- Assess dependence by computing *angular measures* which give favored directions for big values of $(S, D)$.
  - Standardize the pairs to have the same tails by *ranks method*.

    *Standardize* means putting coordinates $(S, D)$ on same scale by non-linear monotone transformation so that $b_S(t), b_D(t)$ replaced by $t, t$.
  - Threshold the resulting pairs and keep only those data pairs outside a large circle.
  - Convert to polar coordinates.
  - Make density plot of $\theta$-coordinate. The corresponding measure is $\mathbb{S}(d\theta)$, called the *angular* or *spectral* measure. So

$$\mathbb{S}(d\boldsymbol{\theta}) = \text{prob standardized point far away from (0,0)} \\ \text{is in direction } d\boldsymbol{\theta}.$$

## Simple scatter plot.

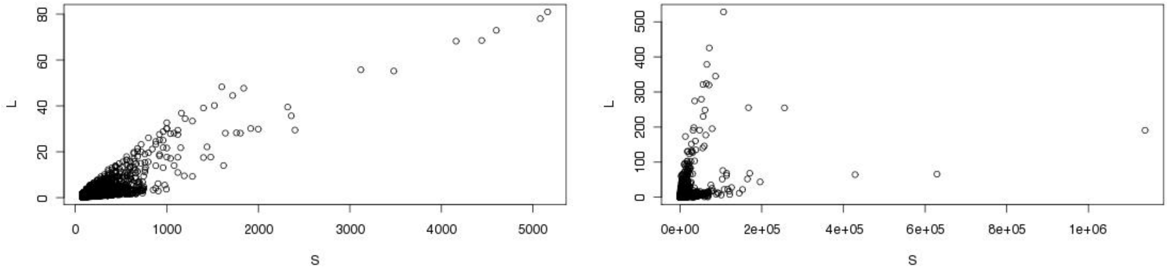



Figure 3: $S$ vs. $D$ for (left) 1st $P^{\vee}$ decile and (right) 6th $P^{\vee}$ decile.

## Alpha-beta-like split; angular measures $(S, D)$

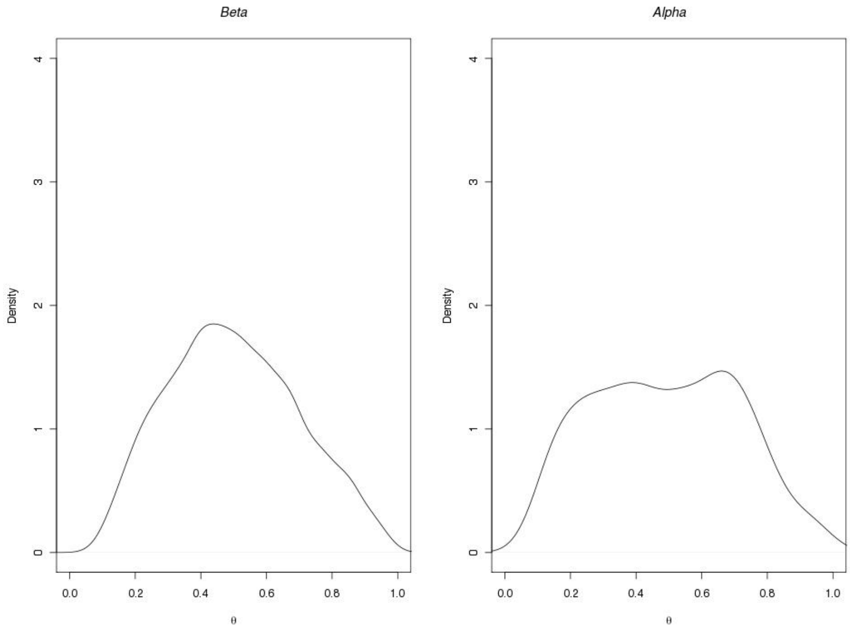



Figure 4: Non-parametric estimates of the spectral density of $\mathbb{S}$ for (left) a beta aggregate of 9 deciles and (right) an alpha group.

## Finer split: plots reasonably symmetric, unimodal

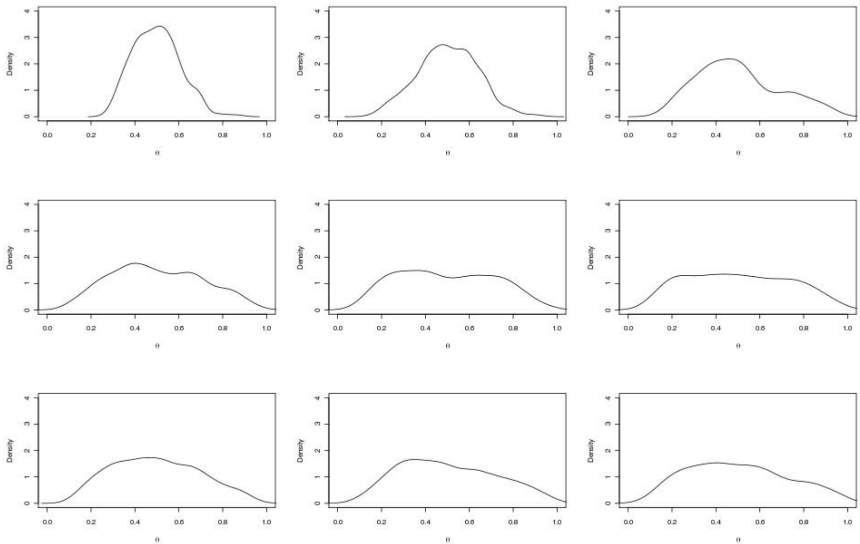



Figure 5: Non-parametric estimates of the spectral density of $\mathbb{S}$ from left to right and top to bottom: 1st to 9th decile groups.

## Comments

- Seek to relate the explicit level of $P^\vee$ with the dependence structure of $(S, D)$.

  Seek global model: Hope the spectral measure $\mathbb{S}$ can be approximated by some $\mathbb{S}_\psi$ from a parametric family of densities $h_\psi(\theta)$ where a (generalized linear) model links $g(\psi) \sim$ decile group. So

  $$\text{session} \longrightarrow \text{peak rate } P^\vee \xrightarrow{glm} \psi \longrightarrow h_\psi(\theta) \longrightarrow \text{simulated standardized } (S, D) \longrightarrow \text{simulated } (S, D).$$

- Using QQ plots and sample acf's can check

  within decile groups, session initiation times look Poisson.

  This is not true across the whole data set–only when the data is segmented by decile group; also not true with other definitions of peak rate.

## 9.1. Global model: Toward a parametric model for the spectral density $\mathbb{S}$

Logistic model:

$$h_\psi(t) = \frac{1}{2}\left(\frac{1}{\psi} - 1\right) t^{-1-1/\psi}(1-t)^{-1-1/\psi}[t^{-1/\psi} + (1-t)^{-1/\psi}]^{\psi-2}, \quad (6)$$
$$= h(t), \quad 0 \le t \le 1,$$

with $\psi \in (0,1)$.

Features:

- Symmetric.
- For $\psi < 0.5$ : $h$ is unimodal and as $\psi \to 0$ we obtain perfect dependence.
- For $\psi > 0.5$ : $h$ is bimodal and as $\psi \to 1$ we obtain asymptotic independence.

This allows us to quantify the effect of $P^\vee$ on the dependence between $S$ and $D$.

## Parametric vs non-parametric density estimates.

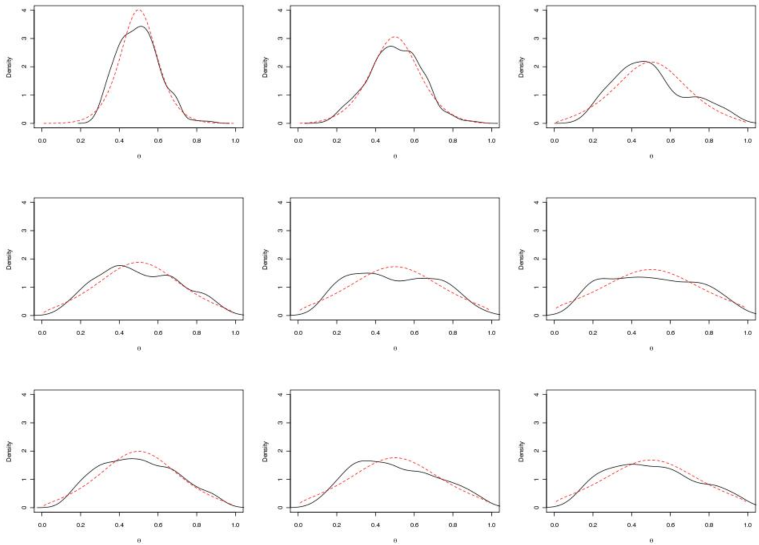


Figure 6: Parametric estimates of the spectral density $\mathbb{S}$ superimposed to non-parametric counterparts, from left to right and top to bottom: 1st to 9th decile groups.

## Dependence of $(S, D)$ as a function of $P^\vee$

Fit a global trend logistic model where

$$g^{-1}(\psi) = \beta_0 + \beta_1 \log(P^\vee).$$

After some experimenting choose link function $g$

$$g(x) = \frac{1/2}{1 + e^{-x}}.$$

## Sketch of simulation of heterogeneous traffic:

1. In the existing data set, each session has an associated $P^\vee$. Form the EDF. Get a bootstrap sample of $P^\vee$ from this EDF and divide into $m = 10$ samples according to the quantiles.
2. For each group, simulate the starting times of the sessions via homogeneous Poisson process.
3. For each $P_j^\vee$, compute the corresponding value of $\psi_j$ from the GLM and use it to simulate an angle $\Theta_j$ from the logistic distribution.
4. Simulate the radius component $N_j$; use Pareto for the heavy tail.
5. Transform to Cartesian coordinates and then invert using fitted marginal distributions to get back to the original scale where $(S_j, D_j)$ do not have same tails.Compute $R_j = S_j/D_j$.

## What about $R$?

- Except for highest decile, $R$ is not in a domain of attraction and not heavy tailed.
- We have evidence that

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

  for $\alpha(t)>0$ and $G$ a pm.

  Allows application of an emerging theory of conditional extreme values (CEV).
- Evidence that $R|D$ cannot be modeled. How come?

## Credit:

López-Oliveros and Resnick (2011) + ideas of Jan Heffernan.

# 10. Final thoughts

- Consistency issues for $(S, D, R)$ where $S/D = R$. Why are the models
  - $(S, D)$ jointly heavy tailed and
  - $R|S > t$ has a limiting type

  consistent.

  Need theory to match empirical observation.
- Conditional on (deciles of) peak rate, sessions arrive according to a homogeneous Poisson. Use this theoretically? Overall traffic is a mixture?
- Sessions are better behaved after conditioning on peak rate. But peak rate is difficult to analyze based on packet level models.
- Segmentation of sessions by application: HTTP, mail, streaming, ftp, ssh, . . . .
  - Difficult to map application on ports so difficult to identify applications. Many applications use port 80 which is supposedly reserved for HTTP.

– As opposed to 1999 data set used for peak rate study, a 2011 data set from CAIDA shows signifcant censoring in 1 hour collection interval.
– Port 25 $\leftrightarrow$ mail easiest to study; eg arrivals of sessions form Poisson process in between Poisson interruptions.

## References


L. Breiman. On some limit theorems similar to the arc-sin law. *Theory Probab. Appl.*, 10:323–331, 1965.

R. Gaigalas and I. Kaj. Convergence of scaled renewal processes and a packet arrival model. *Bernoulli*, 9(4):671–703, 2003. ISSN 1350-7265.

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. Convergence of scaled renewal processes o fractional brownian motion. Preprint: Department of Mathematics, Uppsala University, Box 480, S-751 06 Uppsala, Sweden, 1999.

I. Kaj. *Stochastic Modeling in Broadband Communications Systems*. SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. ISBN 0-89871-519-9.

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.

T. Konstantopoulos and S.J. Lin. Macroscopic models for long-range dependent network traffic. *Queueing Systems Theory Appl.*, 28(1-3): 215–243, 1998. ISSN 0257-0130.

L. López-Oliveros and S. Resnick. Extremal dependence analysis of network sessions. *Extremes*, 14(1):1–28, 2011. ISSN 1386-1999. URL http://dx.doi.org/10.1007/s10687-009-0096-4.

T. Mikosch and A. Stegeman. The interplay between heavy tails and rates in self-similar network traffic. Technical Report. University of Groningen, Department of Mathematics, Available via www.cs.rug.nl/˜eke/iwi/preprints/, 1999.

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 E. van den Berg. Weak convergence of high-speed network traffic models. *J. Appl. Probab.*, 37(2):575–597, 2000.

S. Sarvotham, R. Riedi, and R. Baraniuk. Network and user driven on-off source model for network traffic. *Computer Networks*, 48:335–350, 2005. Special Issue on *Long-range Dependent Traffic*.

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.