Modeling Telecommunications Traffic

Heavy-tails

Matthew Roughan
<matthew.roughan@adelaide.edu.au>

Discipline of Applied Mathematics
School of Mathematical Sciences
University of Adelaide

Occam's razor

Pluralitas non est ponenda sine neccesitate

William of Ockham (ca. 1285-1349)

- "Plurality should not be posited without necessity."
- alternative versions
  - “Entia non sunt multiplicanda praeter necessitatem”, or “Entities should not be multiplied beyond necessity”
  - “in vain we do by many which can be done by means of fewer”
  - “if two things are sufficient for the purpose of truth, it is superfluous to suppose another”
  - Principle of Parsimony

Quidquid latine dictum sit, alutum viditur.

Occam's razor

> I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.
>
> Enrico Fermi

- You can always get a model that fits your data better by using more parameters
- Is this really a better model?
  - Does it explain more?
  - What do the parameters mean?
  - Can it be used elsewhere, or is the model very specific to the data in question?

Principle of Parsimony

Principle of Parsimony says that models with fewer parameters are often better

- there is a tradeoff
- more parameters, better fit
- but can overfit to data, so more parameters make model less universal, and mode specific to the dataset
- more parameters make estimation harder

Heavy-tails

Heavy-tailed distributions

Heavy-tailed distributions occur in many places in Internet traffic

- in On/Off times
- in file sizes (on file systems, and web servers, and observed being transfered on networks)
- in pause times between interactions
- sometimes in marginal distribution

Salient features

- high variability
- tail event, even though low probability have a large impact on overall behaviour

Heavy-tailed distributions

Sub-exponential examples

- log-Normal: $\text{Log-}N(\mu, \sigma^2)$

$$\begin{aligned} p(x) &= \frac{1}{\sigma\sqrt{2\pi}} x^{-1} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right), \quad x > 0 \\ E[X] &= \exp(\mu + \sigma^2/2) \\ \mathrm{Var}[X] &= e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \end{aligned}$$

Obtained when log of the data follows a normal law (e.g. if there is a product of errors instead of a sum).

"Statistical Distributions", M. Evans, N. Hastings and B. Peacock, 2nd Ed., John Wiley and Sons, Inc., New York, 1993.

Heavy-tailed distributions

Sub-exponential examples

- Weibull: $\mathrm{Weibull}(b,c)$

$$
\begin{aligned}
p(x) &= \frac{cx^{c-1}}{b^c}\exp\left[-\left(\frac{x}{b}\right)^c\right], \quad x \geq 0 \\
F(x) &= 1-\exp\left[-\left(\frac{x}{b}\right)^c\right] \\
E[X] &= b\Gamma\left(\frac{c+1}{c}\right) \\
\mathrm{Var}[X] &= b^2\left[\Gamma\left(\frac{c+2}{c}\right)-\Gamma\left(\frac{c+1}{c}\right)^2\right]
\end{aligned}
$$

“Statistical Distributions”, M. Evans, N. Hastings and B. Peacock, 2nd Ed., John Wiley and Sons, Inc., New York, 1993.

Heavy-tailed distributions

Sub-exponential examples

- **Pareto** $x \geq a$

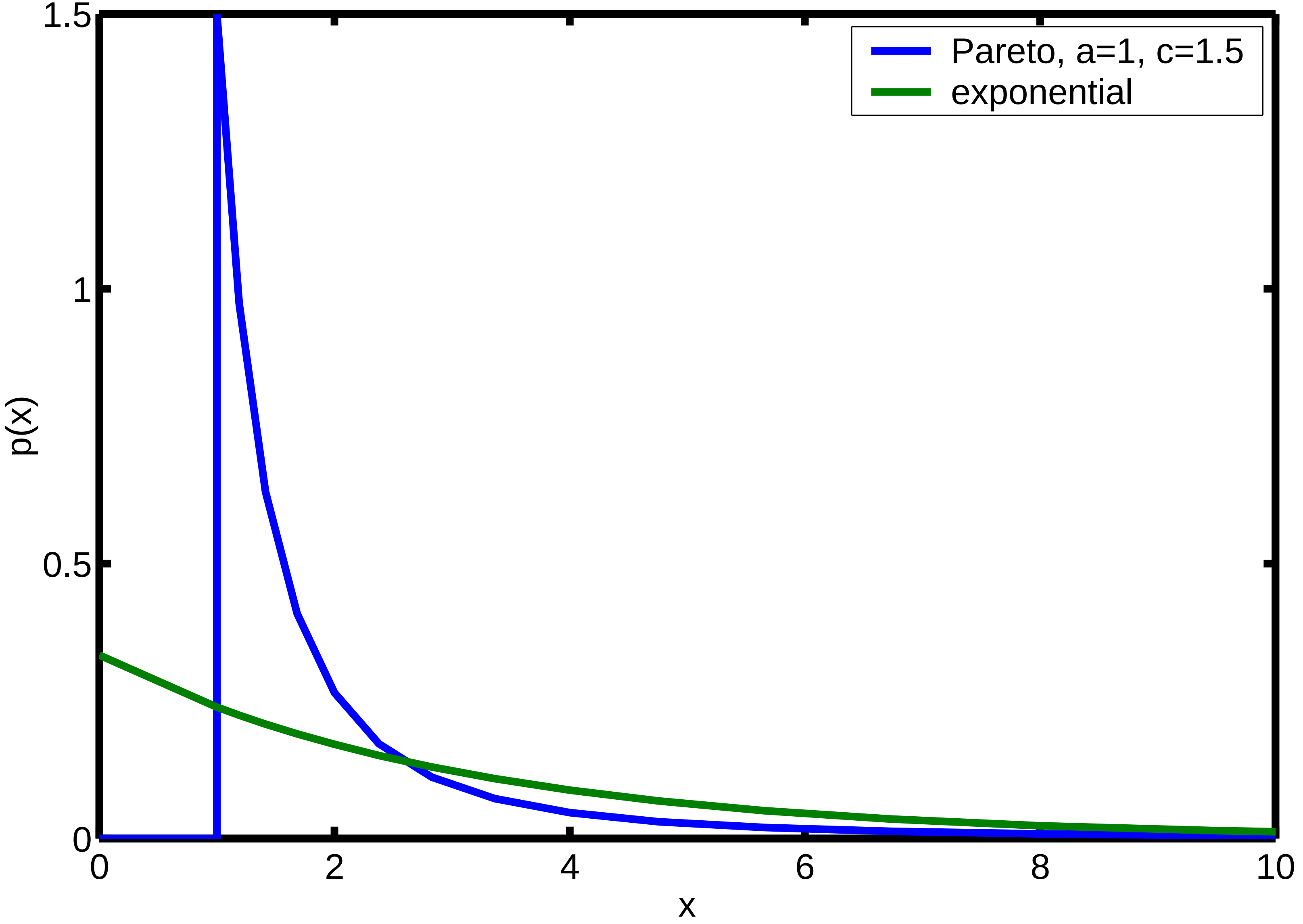
$$\begin{aligned} p(x) &= \frac{ca^c}{x^{c+1}} \\ F(x) &= 1 - \left(\frac{a}{x}\right)^c \\ E[X] &= \frac{ca}{c-1}, \quad c > 1 \\ \mathrm{Var}[X] &= \frac{ca^2}{(c-1)^2(c-2)}, \\ & \quad c > 2 \end{aligned}$$

- shifted Pareto $x \geq 0$

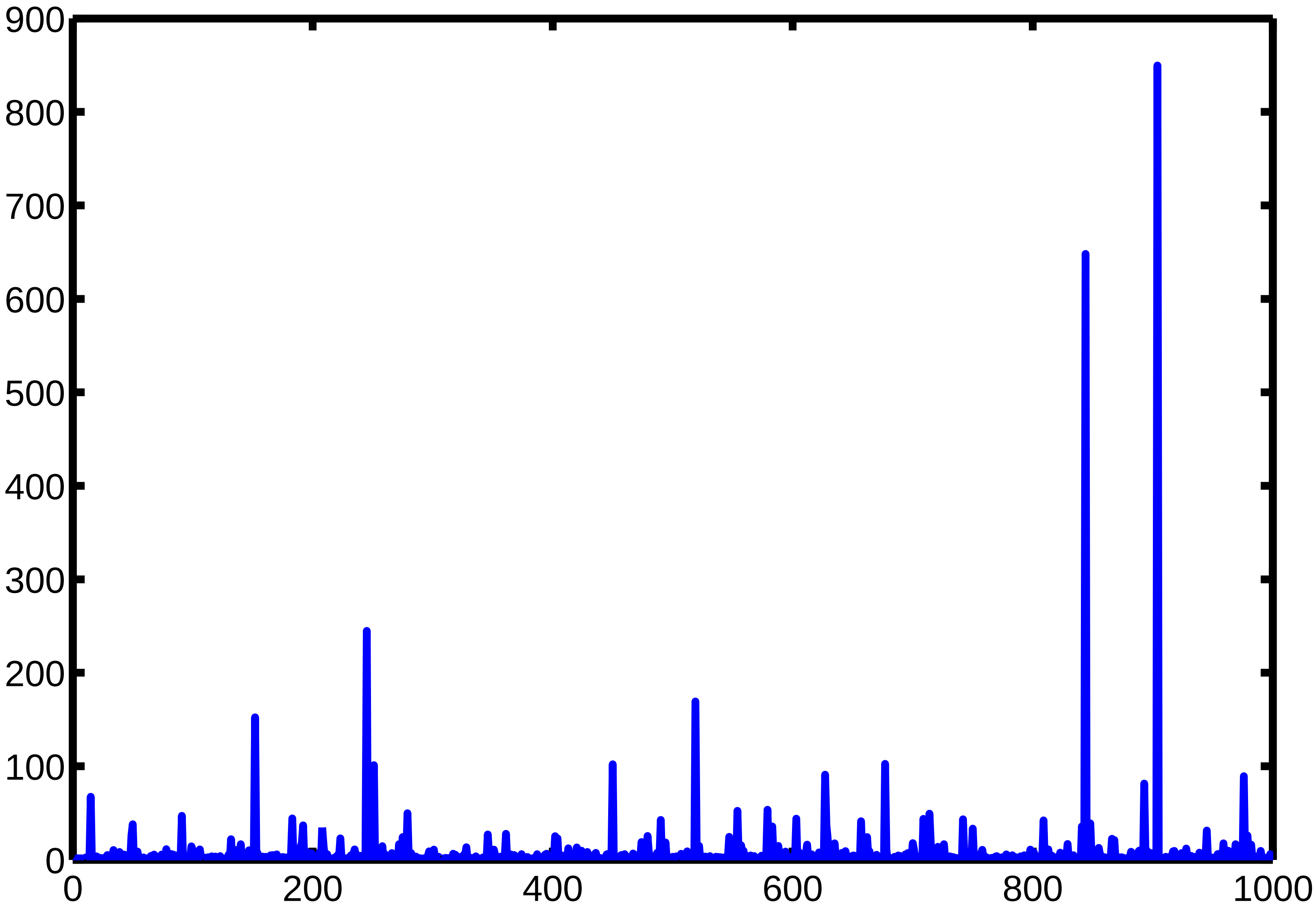
$$\begin{aligned} p(x) &= \frac{ca^c}{(x+a)^{c+1}} \\ F(x) &= F(x) = 1 - \left(\frac{a}{x+a}\right)^c \\ E[X] &= \frac{a}{c-1}, \quad c > 1 \end{aligned}$$

Otherwise known as a power-law distribution.

Pareto example



Pareto example



The CCDF

Complimentary Cumulative Distribution Function (CCDF)

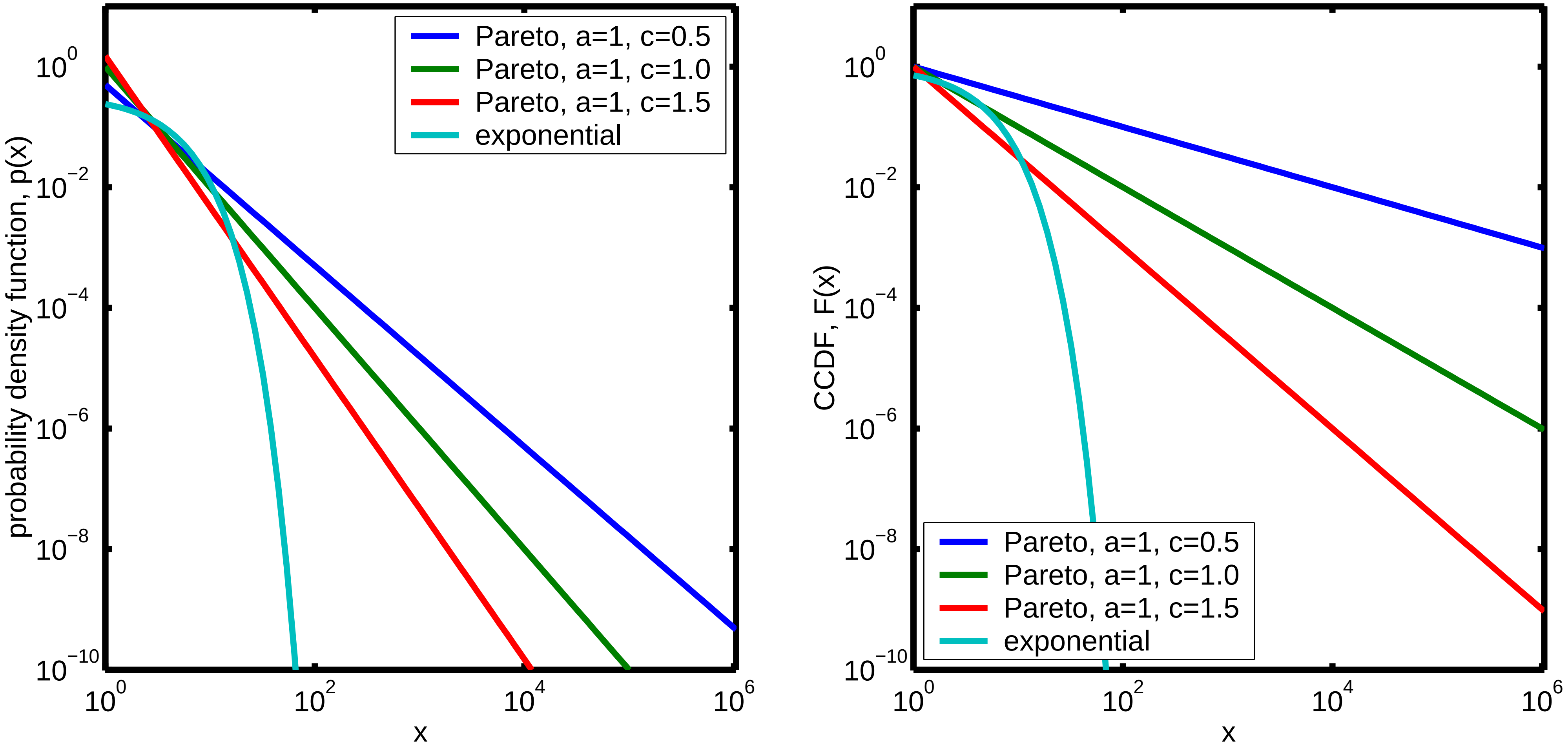
- Defined by $CCDF = F^c(x) = 1 - F(x) = p\{X > x\}$
- For a power-law distribution, the CCDF follows a power-law
- e.g. Pareto has CCDF $F^c(x) = \left(\frac{a}{x}\right)^c$
- exponent c is one larger than for the density
- e.g. Pareto has density $p(x) = \frac{ca^c}{x^{c+1}}$
- this is a more robust measurement than the density function

Log-log pictures

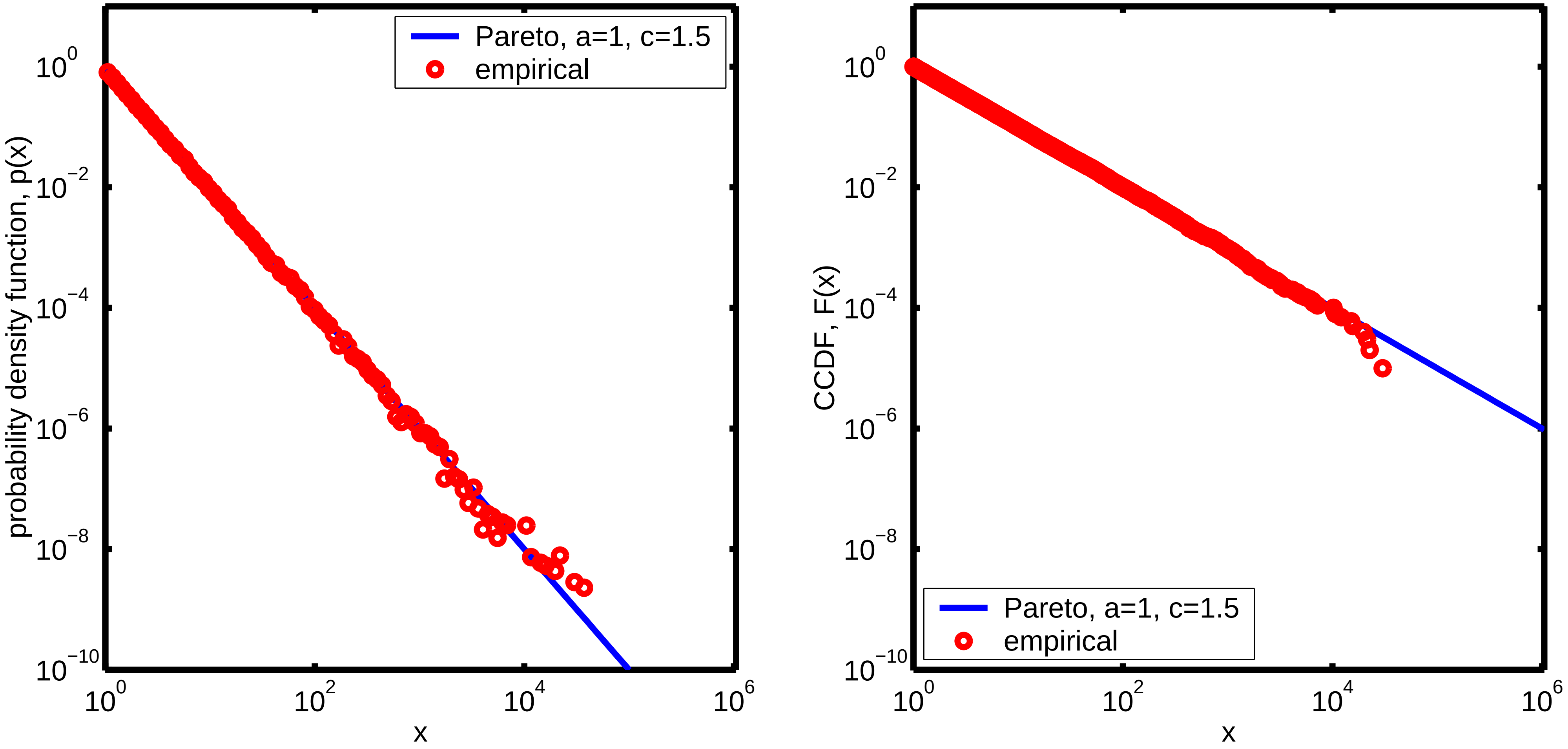
When we plot power-laws in log-log graph and they appear as straight lines.

- Plot $F^c(x) = \left(\frac{a}{x}\right)^c$ on a graph with a log-y axis
- we see $\log F^c(x) = c\,(\log a - \log x)$.
- Take $Y = \log F^c(x)$ and $X = \log x$
- We get
$$Y = k - cX$$
- So the power law appears in the graph of (X,Y) (the log-log graph) as a straight line with slope $-c$.

Pareto example



Pareto example



Pareto distributions

Properties

- **infinite variance:** if $c \leq 2$ the mean of the Pareto distribution is infinite.
- **infinite mean:** if $c \leq 1$ the mean of the Pareto distribution is infinite.
- in general, if $k = \lceil c \rceil$, then the first $k-1$ moments, and central moments of the Pareto distribution will be finite, and the kth moment (and larger moments) will be infinite.

Infinite moments

Infinite moments do not mean the value is infinite

- value is finite with probability 1
- mean is defined by an integral

$$E[X] = \int_{-\infty}^{\infty} x \, p_X(x) \, dx$$

- this integral doesn't necessary converge for all distributions $p_X(x)$
  - it may take infinite values
  - it may be undefined
- this is OK!

Truncation

- Why not truncate the distribution?
  - this would make integral converge
  - real data must surely be finite?
- truncation is problematic
  - it introduces another parameter
  - parameter is out in the tail
  - hence VERY hard to estimate
  - the distribution would have high variation anyway
- more parsimonious model is just to allow the heavy-tail
  - statistical distribution is always just a model

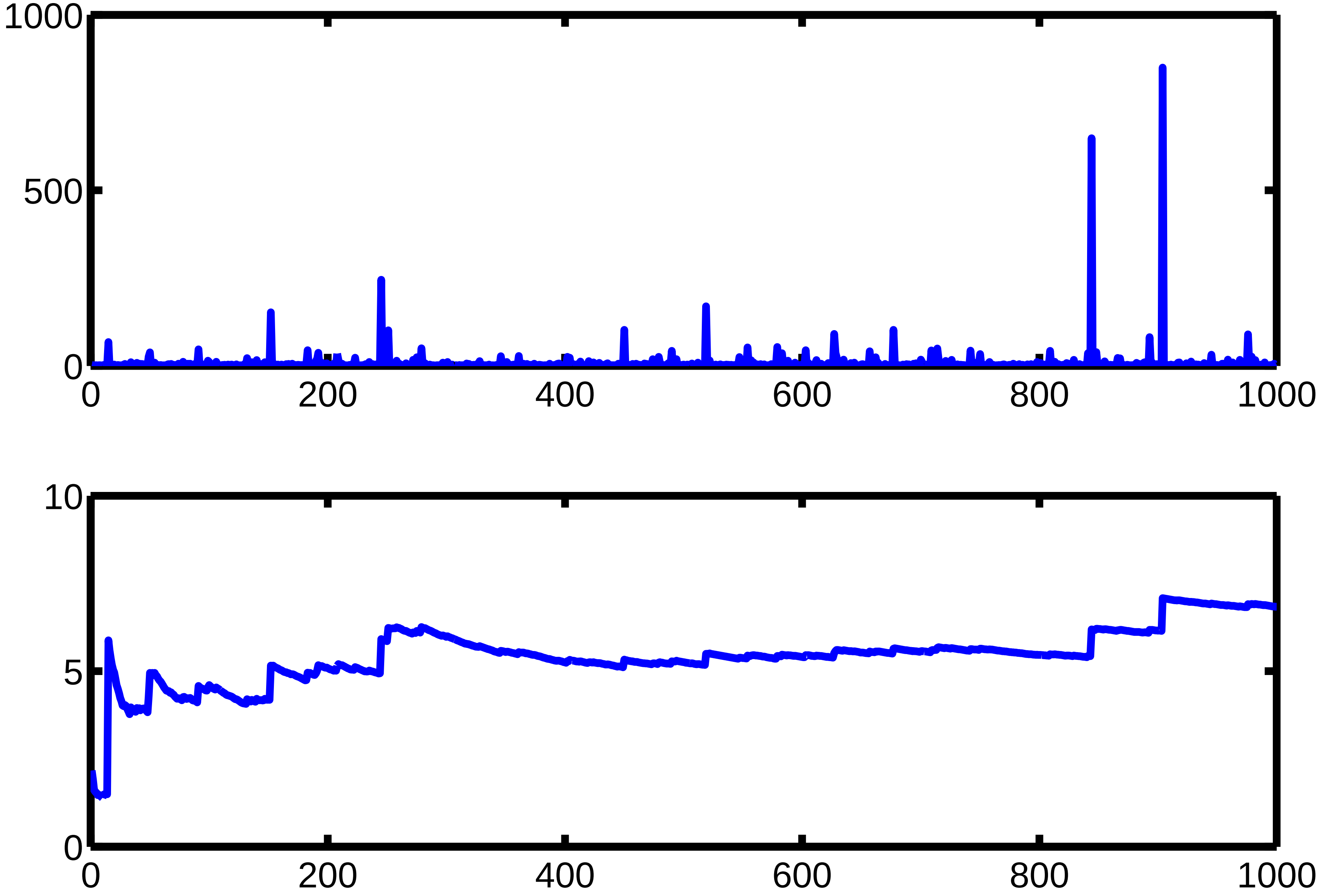
Implications of infinite moments

- estimates converge slowly or not at all
  - infinite mean distribution → sample mean doesn't converge, e.g.

    $$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \infty$$

    Hence, we cannot use this to model the data!
  - infinite variance distribution → sample variance doesn't converge
  - infinite variance distribution → sample mean converges only slowly
- poor queueing behaviour
  - we will discuss more later

Pareto example



Regular Variation

Mathematical generalization of a power-law distribution

- **Asymptotically equivalent:** two functions $g(t)$ and $h(t)$ are asymptotically equivalent at t_0, if

$$\lim_{t \to t_0} \frac{|h(t)|}{|g(t)|} = 1$$

  and this is denoted by $g(t) \overset{t_0}{\sim} h(t)$

- **Slowly varying:** a function is slowly varying at t_0 if, for all $x > 0$

$$\lim_{t \to t_0} \frac{L(xt)}{L(t)} = 1$$

- **Regularly varying:** a function is regularly varying at ∞, with exponent p, if

$$h(t) \overset{\infty}{\sim} L(t) t^p$$

Regular Variation

Examples slowly varying functions (at ∞)

- $const$, or $\log(t)$, or $e^{-b/t}$

Examples regularly varying (at ∞) distributions

- Pareto distribution $L = const$. Clearly a constant is slowly varying.
- Inverse Gamma Distribution with density function

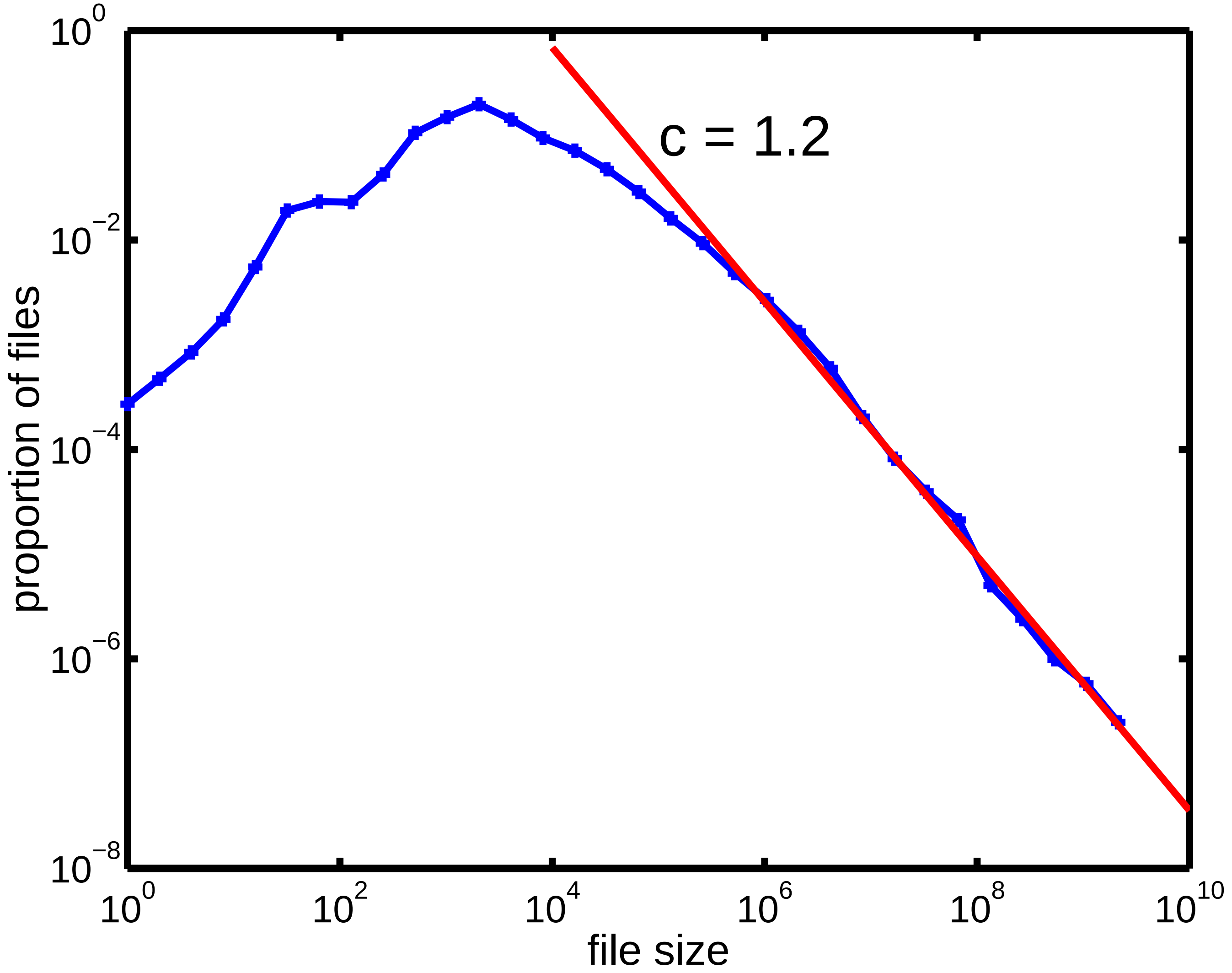
$$p(t) = \frac{b^{\alpha} e^{-b/t}}{\Gamma(\alpha)} t^{-\alpha-1}$$

where L is given by a constant times $e^{-b/x}$.

- any distribution with a power-law tail

Heavy-tails

Unix file size survey (1994)



http://www.base.com/gordoni/ufs93.html

Web data

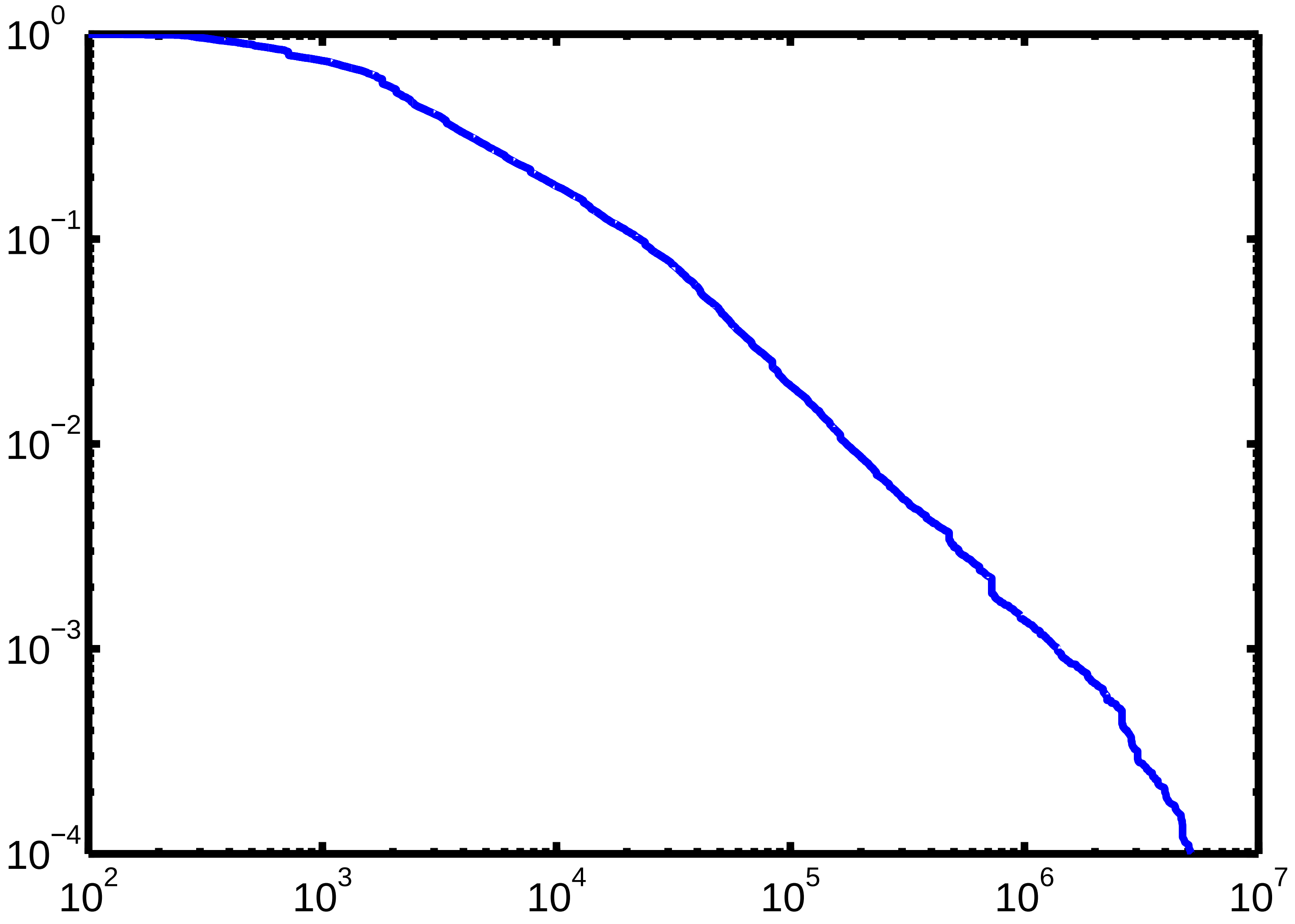
Boston Uni study, 1995 (and studies since)

- derived from instrumented client traces
- running in the Boston University Computer Science Department
- spanning the timeframe of 21 November 1994 through 8 May 1995.
- 9,633 Mosaic sessions
- 762 different users
- 1,143,839 requests for data transfer.

"Characteristics of WWW Client Traces", Carlos A. Cunha, Azer Bestavros and Mark E. Crovella, Boston University Department of Computer Science, Technical Report TR-95-010, April 1995.

Web data

Boston Uni study, 1995

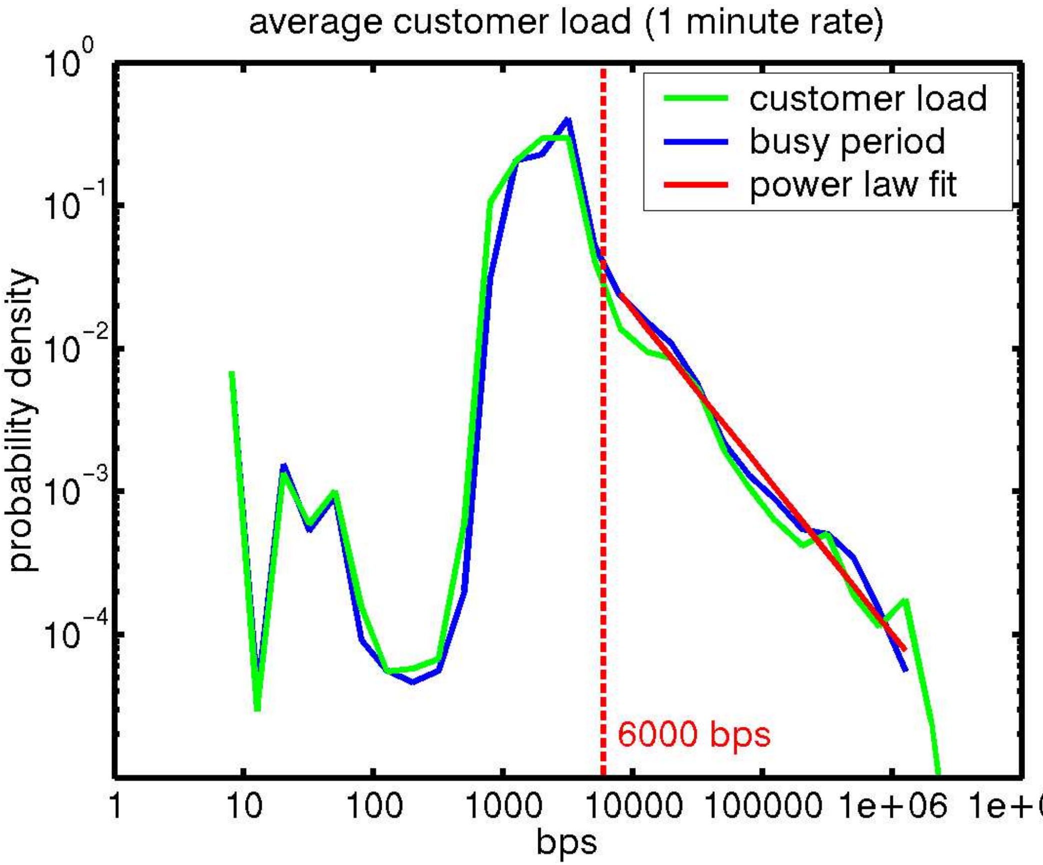


http://www.cs.bu.edu/faculty/crovella/links.html

Broadband data

AT&T Broadband performed SNMP study of their traffic at cable headends.

- one minute SNMP byte counts
- customer bit rates
- chatter below 6 kbps
- above 6 kbps, we see a power law

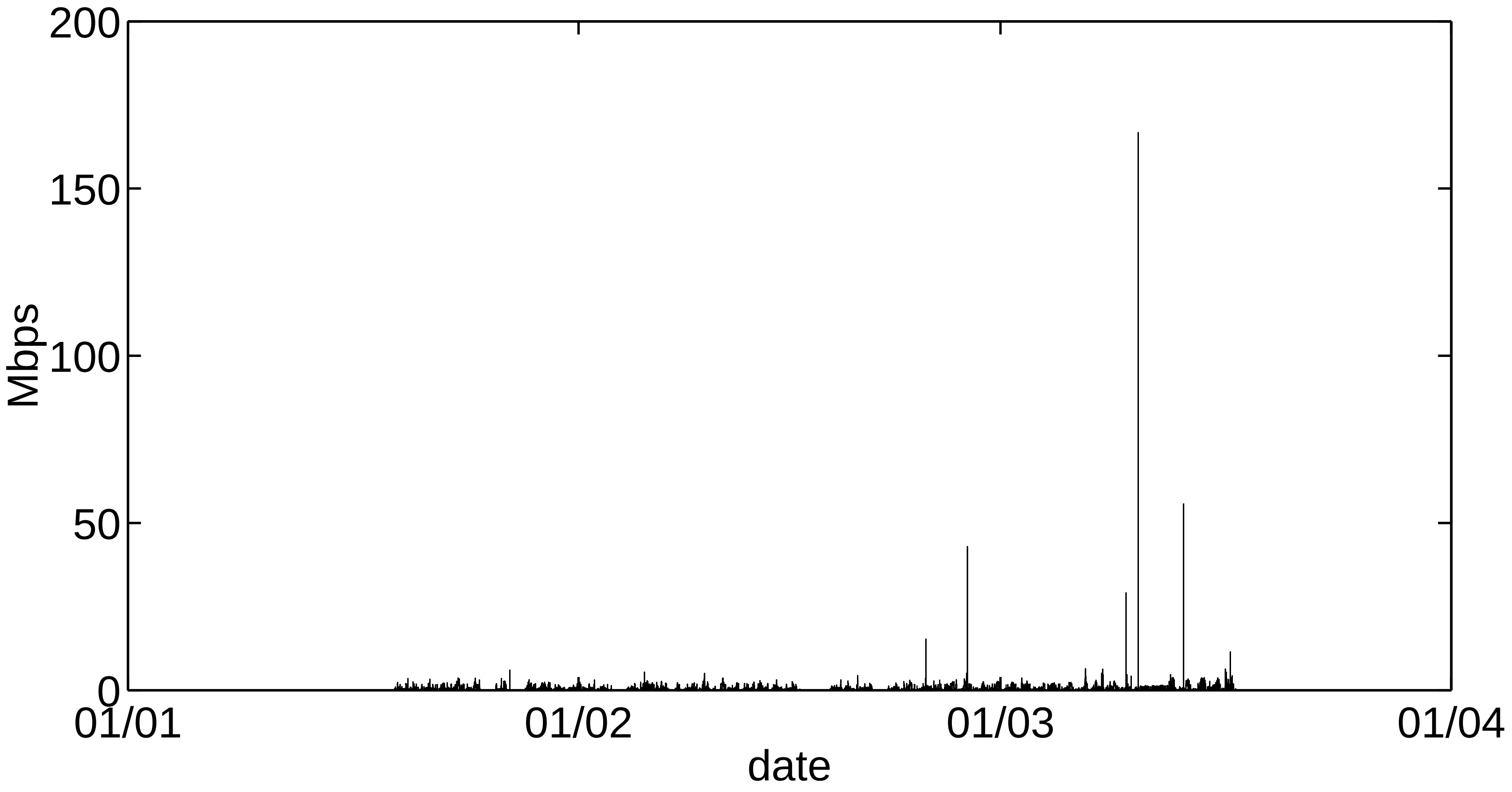


"Pragmatic Modeling of Broadband Access Traffic", Matthew Roughan and Charles R. Kalmanek, Computer Communications, vol 26/8, pp.804-816, 2003.

Broadband data

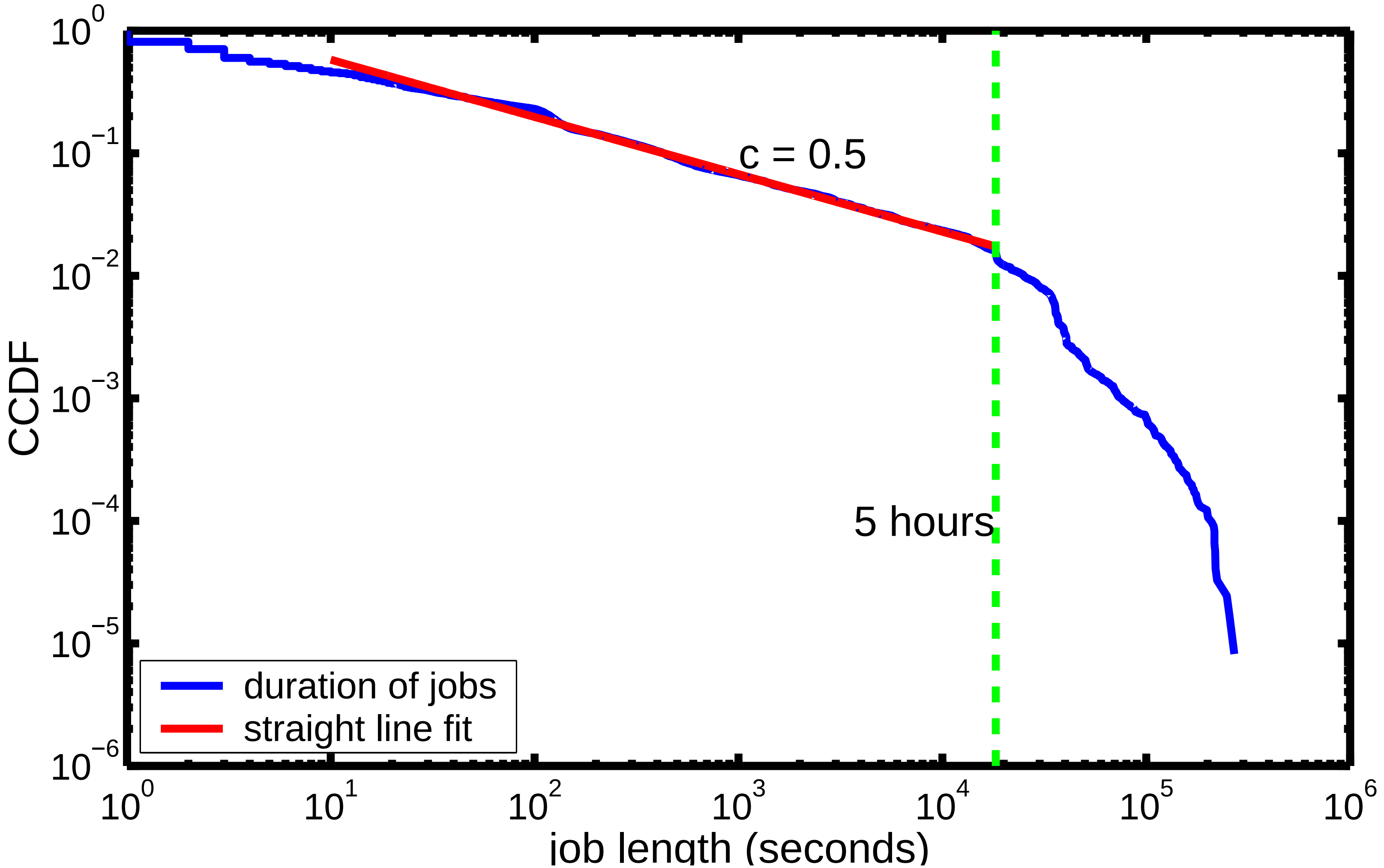
AT&T Broadband performed SNMP study.

- one minute SNMP byte counts
- total aggregate at head-end



Super computer job run lengths

CPU times of jobs on SAPACs CM5 (1999), 129625 records



Ubiquity

We seem to see these heavy-tails quite a lot.

Why?

Normal distributions

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified,if they had known of it. It reigns with serenity and in complete self-effacement,amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

Sir Francis Galton, 1889

Normal distributions

Why is the normal law so common?

- **Central Limit Theorem:** The sum of a series of random variables (satisfying certain conditions) will converge to a normal distribution.
  So Normal/Gaussian distributions tend to appear where we sum up lots of random variables, the larger the mob, the closer to Gaussianity we get.
- **Invariance:** Various operations applied to a normal distribution result in a normal distribution, e.g.
  - multiplication by a scalar
  - addition of two normal distributions

  Hence, once we get a Normal distribution, it tends to stick around.

Normal distributions

Why don't we always get a normal law?

- some processes are not well modelled by the limit of a sum of random variables
- other limits may apply
  - Poisson distribution is the limit for rare events in a large population
  - limit of the maximum of a set of random variables
- non-linear transformation applied to the random variables (which breaks the normal law).
- correlations in the data
- **infinite variance (heavy-tails)**

Generalized Central Limit Theorem

Theorem: Generalized Central Limit Theorem
Let $X_1, X_2, X_3, \ldots$ be an independent, identically distributed series of random variables. There exists a constants $a_n > 0$, and $b_n \in \mathbb{R}$ and a non-degenerate random variable Z, with

$$a_n(X_1 + X_2 + \cdots + X_n) - b_n \xrightarrow{d} Z,$$

if and only if Z is α-stable, in which case $a_n = n^{-1/\alpha}$ for some $\alpha \in (0, 2]$.

- an α-stable is the generalization of the Gaussian, to allow heavy-tails
- note the slower than CLT convergence rate

Why you expect to see heavy-tails

For the same reasons we see Gaussians

- **Generalized Central Limit Theorem:** The sum of a series of random variables stable distribution (under looser conditions than before).
- **Invariance:** Various operations applied to a normal distribution result in a normal distribution, e.g.
  - multiplication by a scalar
  - addition of two stable distributions
  - max also leads to a stable law

Actually, we might expect, given larger range of invariant behaviour for stable laws that we see them more often than Gaussians.

How can we use this here?

- use heavy-tailed distributions in renewal processes
  - On/Off process, use for On and Off times
  - renewal reward process, use for reward
- Note that these are models of a single, or fixed number of sources
- An alternative is the $M/G/\infty$ arrival process
  - traffic sources arrive as a PP
  - they stay around for a generally distributed “service time”
  - the service time would be chosen to have a heavy-tail
  - while sources are around, they generate traffic at rate r

Renewal models with heavy-tails

- most renewal theory still holds
- some special names used
  - a renewal process with heavy-tailed inter-renewal times = fractal renewal process
  - Doubly stochastic PP driven by an On/Off process with heavy-tailed On/Off times = Fractal Binomial Noise Driven Poisson Process (FBNDP)
  - A doubly stochastic PP Shot noise point process with inter-arrivals that are heavy-tailed = Fractal Shot Noise Driven Poisson Process (FSNDP)

M/G/∞ model

- the number of customers in the system is insensitive to the service time
- it just follows the simple Poisson distribution

$$p_n = \frac{e^{-\lambda}\lambda^n}{n!}$$

- hence the marginal traffic rate follows a Poisson distribution with rate r times the number of sources, as given by the above distribution.
- the service time would be chosen to have a heavy-tail
  - no impact on marginal distribution
  - results in interesting correlations in the process
- once again, this is an aggregate traffic model

Estimation of parameters

Simple methods for estimation of heavy-tailed parameters

- regression of the CCDF on a log-log graph
- MLE
- Hill estimator

MLE for Pareto parameters

Pareto density function: $p(x; a, c) = \frac{ca^c}{x^{c+1}}$. The probability of a particular set of IID samples $\{X_i\}$ from the distribution is

$$p(X_1, \ldots, X_n | a, c) = \prod_{i=1}^{n} p(X_i | a, c)$$

The Likelihood of a particular pair of parameters is defined by

$$L(a, c) = p(a, c | X_1, \ldots, X_n) = \prod_{i=1}^{n} p(a, c | X_i)$$

We could equally maximize the Log-Likelihood

$$\log L(a, c) = \sum_{i=1}^{n} \log p(a, c | X_i) = \sum_{i=1}^{n} \log c + c \log a - (c+1) \log X_i$$

MLE for Pareto parameters

We want to maximize this, so we take the derivatives WRT a and c

$$\begin{aligned}
\frac{\partial L}{\partial a} &= \sum_{i=1}^{n} c/a \\
\frac{\partial L}{\partial c} &= \sum_{i=1}^{n} \frac{1}{c} + \log a - \log X_i \\
&= \frac{n}{c} - \sum_{i=1}^{n} \log X_i - \log a
\end{aligned}$$

The first is only zero for $c = 0$, so we can't really use it, instead we choose $\hat{a} = \min_i X_i$.

MLE for Pareto parameters

Assume we have the estimate $\hat{a}$ for a, setting the derivative $\frac{\partial L}{\partial c}$ to zero results in

$$\begin{aligned} \frac{n}{c} - \sum_{i=1}^{n} \log X_i - \log a &= 0 \\ \frac{n}{c} &= \sum_{i=1}^{n} \log(X_i/a) \\ \frac{1}{c} &= \frac{1}{n} \sum_{i=1}^{n} \log(X_i/a) \\ c &= \left[\frac{1}{n} \sum_{i=1}^{n} \log(X_i/a) \right]^{-1} \end{aligned}$$

MLE for Pareto parameters

Given a Pareto distribution we form the Maximum Likelihood Estimator (MLE) of the parameters by

$$\begin{aligned} \hat{a} &= \min X_i \\ \hat{c} &= \left(\frac{1}{n} \sum_{i=1}^{n} \log(X_i/\hat{a}) \right)^{-1} \end{aligned}$$

Using this is complicated by the fact that the body of the distribution may not follow a power-law, so we have to only apply to the tail.

- how do you choose the tail?
- Hill estimator allows you to visually see how much of the tail you need to get a good estimate.

Hill estimator

How it works: $F^c(x) \sim \left(\frac{a}{x}\right)^c$

We can estimate $F^c(x)$ using the order statistics $X_{(i)}$ of the data, e.g. $\hat{F}^c(i/n) = X_{(i)}$. Assume the body of the distribution is not a power-law, but the tail, past the ith order stat $X_{(i)}$ does follow a power-law, then for $x > X_{(i)}$

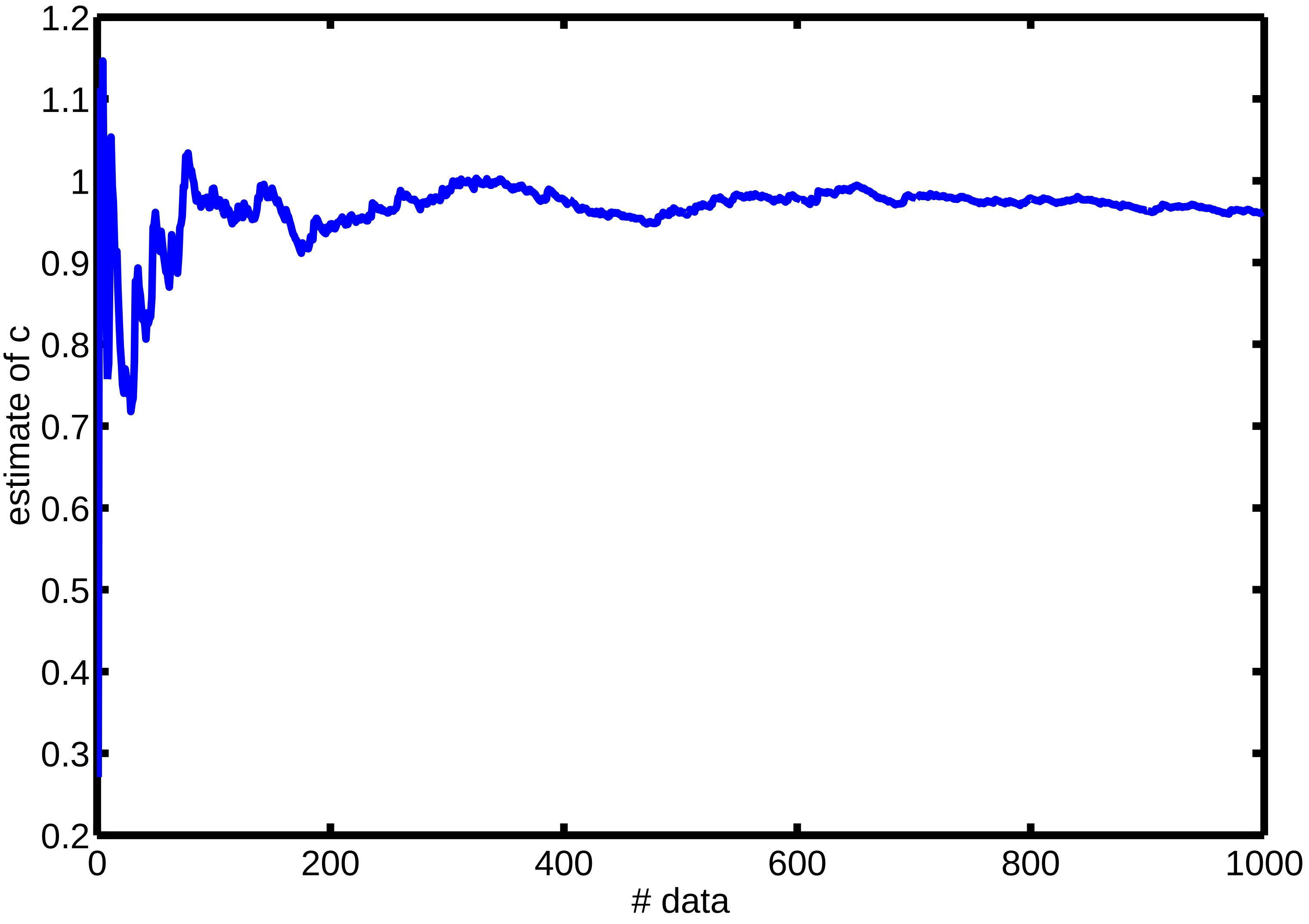
$$\hat{F}^c(x) \propto \left(\frac{X_{(i)}}{x}\right)^c$$

So we use the MLE estimator on the tail data, e.g. we use the data $X_{(i+1)}, \ldots, X_{(n)}$, where $\hat{a} = X_{(i)}$. That is

$$\hat{c}_i = \left(\frac{1}{n-i} \sum_{k=i+1}^{n} \log(X_{(k)}/X_{(i)})\right)^{-1}$$

Hill estimator

A simple estimator of power-law tail parameter c



Hill estimator: matlab code

Matlab version

```
function [mle_est, hill_est] = ...
             hill_estimate(input_data, do_plot)

n = length(input_data);
k = (1:n);
order_stats = fliplr(sort(input_data’));
t = log(order_stats);
mle_est = (sum(t)/n).ˆ(-1); % mle for Pareto dist.

hill_est = cumsum(t(1:n-1))./k(1:n-1) - t(k(1:n-1)+1);

if (do_plot)
  plot(k(1:n-1), hill_est.ˆ(-1));
end
```