The Fundamentals of Heavy Tails
Properties, Emergence, & Identification

Jayakrishnan Nair, Adam Wierman, Bert Zwart

“The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking [...] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world.”

– Clay Shirky, as quoted in Newsweek & the Guardian
Why am I doing a tutorial on heavy tails?

Because we’re writing a book on the topic...

Why are we writing a book on the topic?

Because heavy-tailed phenomena are everywhere!
Why am I doing a tutorial on heavy tails?
Because we’re writing a book on the topic...

Why are we writing a book on the topic?
Because heavy-tailed phenomena are everywhere!
**BUT, they are extremely misunderstood.**

“The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking […] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world.”

– Clay Shirky, as quoted in Newsweek & the Guardian
Heavy-tailed phenomena are treated as something **Mysterious, Surprising, & Controversial**

"The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking [...] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world."

– Clay Shirky, as quoted in Newsweek & the Guardian
Heavy-tailed phenomena are treated as something **Mysterious, Surprising, & Controversial**

Our intuition is flawed because intro probability classes focus on light-tailed distributions

Simple, appealing statistical approaches have BIG problems
Heavy-tailed phenomena are treated as something **Mysterious, Surprising, & Controversial**

**On Power-Law Relationships of the Internet Topology**

Michalis Faloutsos  
U.C. Riverside  
Dept. of Comp. Science  
michalis@cs.ucr.edu

Petros Faloutsos  
U. of Toronto  
Dept. of Comp. Science  
pfal@cs.toronto.edu

Christos Faloutsos *  
Carnegie Mellon Univ.  
Dept. of Comp. Science  
christos@cs.cmu.edu

1999 Sigcomm paper – 4500+ citations!

**On the Bias of Traceroute Sampling**

Dimitri Achlioptas  
Microsoft Research  
Microsoft Corporation  
Redmond, WA 98052  
oplas@microsoft.com

David Kempe  
Department of Computer Science  
University of Southern California  
Los Angeles, CA 90089  
dkempe@usc.edu

Aaron Clauset  
Department of Computer Science  
University of New Mexico  
Albuquerque, NM 87131  
aron.cs.unm.edu

Christopher Moore  
Department of Computer Science  
University of New Mexico  
Albuquerque, NM 87131  
mboore@cs.unm.edu

2005, STOC

**Understanding Internet Topology: Principles, Models, and Validation**

David Alderson, *Member, IEEE*, Lun Li, *Student Member, IEEE*, Walter Willinger, *Fellow, IEEE*, and John C. Doyle, *Member, IEEE*

2005, ToN

Similar stories in electricity nets, citation nets, ...
Heavy-tailed phenomena are treated as something mysterious, surprising, & controversial.

1. Properties
2. Emergence
3. Identification
What is a heavy-tailed distribution?

A distribution with a “tail” that is “heavier” than an Exponential
What is a heavy-tailed distribution?

A distribution with a “tail” that is “heavier” than an Exponential.
What is a heavy-tailed distribution?

A distribution with a “tail” that is “heavier” than an Exponential

**Canonical Example: The Pareto Distribution a.k.a. the “power-law” distribution**

\[
\Pr(X > x) = F(x) = \left( \frac{x_{\min}}{x} \right)^\alpha \quad \text{for} \ x \geq x_{\min}
\]

**density:** \[
f(x) = \frac{\alpha x_{\min}^\alpha}{x^{\alpha+1}}
\]
What is a heavy-tailed distribution?

A distribution with a “tail” that is “heavier” than an Exponential

**Canonical Example:** The Pareto Distribution a.k.a. the “power-law” distribution

Many other examples: LogNormal, Weibull, Zipf, Cauchy, Student’s t, Frechet, ...

\[
\Pr(X > x) = e^{-\mu x}
\]

\[
X: \log X \sim Normal
\]

\[
\bar{F}(x) = e^{-(x/\lambda)k}
\]
What is a heavy-tailed distribution?

A distribution with a “tail” that is “heavier” than an Exponential

![Exponential distribution graph]

**Exponential:** $e^{-\mu x}$

**Canonical Example:** The Pareto Distribution a.k.a. the “power-law” distribution

**Many other examples:** LogNormal, Weibull, Zipf, Cauchy, Student’s t, Frechet, ...

**Many subclasses:** Regularly varying, Subexponential, Long-tailed, Fat-tailed, ...
Heavy-tailed distributions have many strange & beautiful properties

- The “Pareto principle”: 80% of the wealth owned by 20% of the population, etc.
- Infinite variance or even infinite mean
- Events that are much larger than the mean happen “frequently”

These are driven by 3 “defining” properties:

1) Scale invariance
2) The “catastrophe principle”
3) The residual life ”blows up”
Scale invariance
Scale invariance

$F$ is scale invariant if there exists an $x_0$ and a $g$ such that

$$F(\lambda x) = g(\lambda) F(x)$$

for all $\lambda, x$ such that $\lambda x \geq x_0$.

“change of scale”
Scale invariance

$F$ is scale invariant if there exists an $x_0$ and a $g$ such that $F/(\lambda x) = g(\lambda)F(x)$ for all $\lambda, x$ such that $\lambda x \geq x_0$.

**Theorem:** A distribution is scale invariant if and only if it is Pareto.

**Example:** Pareto distributions

$$F/(\lambda x) = \left(\frac{x_{\text{min}}}{\lambda x}\right)^{\alpha} = F(x) \left(\frac{1}{\lambda}\right)^{\alpha}$$
Scale invariance

$F$ is scale invariant if there exists an $x_0$ and a $g$ such that
\[ F(\lambda x) = g(\lambda)F(x) \]
for all $\lambda, x$ such that $\lambda x \geq x_0$.

Asymptotic scale invariance

$F$ is asymptotically scale invariant if there exists a continuous, finite $g$ such that
\[ \lim_{x \to \infty} \frac{F(\lambda x)}{F(x)} = g(\lambda) \]
for all $\lambda$. 
**Example:** Regularly varying distributions

\[ F \text{ is regularly varying if } F(x) = x^{-\rho} L(x), \text{ where } L(x) \text{ is slowly varying,} \]

i.e., \( \lim_{x \to \infty} \frac{L(xy)}{L(x)} = 1 \text{ for all } y > 0. \)

**Theorem:** A distribution is asymptotically scale invariant iff it is regularly varying.

**Asymptotic scale invariance**

\( F \) is asymptotically scale invariant if there exists a continuous, finite \( g \) such that

\[ \lim_{x \to \infty} \frac{F(\lambda x)}{F(x)} = g(\lambda) \text{ for all } \lambda. \]
Example: Regularly varying distributions

$F$ is regularly varying if $\bar{F}(x) = x^{-\rho} L(x)$, where $L(x)$ is slowly varying, i.e., $\lim_{x \to \infty} \frac{L(xy)}{L(x)} = 1$ for all $y > 0$.

Regularly varying distributions are extremely useful. They basically behave like Pareto distributions with respect to the tail:

→ “Karamata” theorems
→ “Tauberian” theorems
Heavy-tailed distributions have many strange & beautiful properties

- The “Pareto principle”: 80% of the wealth owned by 20% of the population, etc.
- Infinite variance or even infinite mean
- Events that are much larger than the mean happen “frequently”

These are driven by 3 “defining” properties

1) Scale invariance
2) The “catastrophe principle”
3) The residual life “blows up”
A thought experiment

During lecture I polled my 50 students about their heights and
the number of twitter followers they have...

The sum of the heights was ~300 feet.
The sum of the number of twitter followers was 1,025,000

What led to these large values?
A thought experiment

During lecture I polled my 50 students about their heights and the number of twitter followers they have...

The sum of the heights was ~300 feet.
The sum of the number of twitter followers was 1,025,000

A bunch of people were probably just over 6’ tall
(Maybe the basketball teams were in the class.)

“Conspiracy principle”

One person was probably a twitter celebrity and had ~1 million followers.

“Catastrophe principle”
Example
Consider $X_1 + X_2$ i.i.d Weibull.
Given $X_1 + X_2 = d$, what is the marginal density of $X_1$?

- Light-tailed Weibull
- Exponential
- Heavy-tailed Weibull

"Conspiracy principle"
"Catastrophe principle"
“Catastrophe principle”

\[
\Pr(\max(X_1, \ldots, X_n) > t) \sim \Pr(X_1 + \ldots + X_n > t)
\]

\[
\Rightarrow \Pr(\max(X_1, \ldots, X_n) > t | X_1 + \ldots + X_n > t) \to 1
\]

“Conspiracy principle”

\[
\Pr(\max(X_1, \ldots, X_n) > t) = o(\Pr(X_1 + \ldots + X_n > t))
\]
"Catastrophe principle"

\[
\Pr(\max(X_1, \ldots, X_n) > t) \sim \Pr(X_1 + \cdots + X_n > t)
\]

\[
\Rightarrow \Pr(\max(X_1, \ldots, X_n) > t | X_1 + \cdots + X_n > t) \to 1
\]

Extremely useful for random walks, queues, etc.

"Principle of a single big jump"
Subexponential distributions

$F$ is subexponential if for i.i.d. $X_i$, $Pr(X_1 + \cdots + X_n > t) \sim nPr(X_1 > t)$

"Catastrophe principle"

$Pr(\max(X_1, \ldots, X_n) > t) \sim Pr(X_1 + \cdots + X_n > t)$
$
\Rightarrow Pr(\max(X_1, \ldots, X_n) > t | X_1 + \cdots + X_n > t) \to 1$
Subexponential distributions

$F$ is subexponential if for i.i.d. $X_i$, $\Pr(X_1 + \cdots + X_n > t) \sim n \Pr(X_1 > t)$
Heavy-tailed distributions have many strange & beautiful properties

- The “Pareto principle”: 80% of the wealth owned by 20% of the population, etc.
- Infinite variance or even infinite mean
- Events that are much larger than the mean happen “frequently”

These are driven by 3 “defining” properties

1) Scale invariance
2) The “catastrophe principle”
3) The residual life “blows up”
A thought experiment

What happens to the expected remaining waiting time as we wait
...for a table at a restaurant?
...for a bus?
...for the response to an email?

Residual life

The remaining wait drops as you wait

If you don’t get it quickly, you never will...
The distribution of residual life

The distribution of remaining waiting time given you have already waited $x$ time is $R_x(t) = \frac{F(x+t)}{F(x)}$.

Examples:

Exponential: $R_x(t) = \frac{e^{-\mu(x+t)}}{e^{-\mu x}} = e^{-\mu t}$ → “memoryless”

Pareto: $R_x(t) = \left(\frac{x_{\text{min}}}{x+t}\right)^\alpha = \left(1 + \frac{t}{x}\right)^{-\alpha}$ → Increasing in $x$
The distribution of residual life

The distribution of remaining waiting time given you have already waited $x$ time is $R_x(t) = \frac{F(x+t)}{F(x)}$.

**Mean residual life**

$$m(x) = E[X - x | X > x] = \int R_x(t) dt$$

**Hazard rate**

$$q(x) = \frac{f(x)}{F(x)} = R'_x(0)$$

Heavy-tailed distributions “tend” to have decreasing hazard rates & increasing mean residual lives

Light-tailed distributions “tend” to have increasing hazard rates & decreasing mean residual lives
What happens to the expected remaining waiting time as we wait
...for a table at a restaurant?
...for a bus?
...for the response to an email?

**BUT:** not all heavy-tailed distributions have DHR / IMRL
some light-tailed distributions are DHR / IMRL

Heavy-tailed distributions “tend” to have decreasing hazard rates & increasing mean residual lives
Light-tailed distributions “tend” to have increasing hazard rates & decreasing mean residual lives
Long-tailed distributions
$F$ is long-tailed if $\lim_{x \to \infty} \frac{\bar{F}(x+t)}{\bar{F}(x)} = 1$ for all $t$

**BUT**: not all heavy-tailed distributions have DHR / IMRL
some light-tailed distributions are DHR / IMRL

*Heavy-tailed distributions* “tend” to have decreasing hazard rates & increasing mean residual lives
*Light-tailed distributions* “tend” to have increasing hazard rates & decreasing mean residual lives
$F$ is long-tailed if $\lim_{x \to \infty} \bar{R}_x(t) = \lim_{x \to \infty} \frac{\bar{F}(x+t)}{\bar{F}(x)} = 1$ for all $t$
Long-tailed distributions

- Regularly Varying
  - Pareto
- Subexponential
  - Weibull
- LogNormal

Heavy-tailed → Difficult to work with in general
Long-tailed distributions

- Pareto
- Subexponential
- Regularly Varying
- Weibull
- LogNormal

Asymptotically scale invariant $\rightarrow$ Useful analytic properties

Heavy-tailed $\rightarrow$ Difficult to work with in general

Useful properties and theoretical insights are often associated with long-tailed distributions, which are characterized by heavy tails in their probability density functions.

However, working with these distributions can be challenging due to their pathological behavior, making them difficult to handle in practical applications.
Long-tailed distributions

Regularly Varying

Subexponential

Pareto

Weibull

LogNormal

Catastrophe principle

Useful for studying random walks

Heavy-tailed

Difficult to work with in general
Residual life “blows up” → Useful for studying extremes

Heavy-tailed → Difficult to work with in general

Long-tailed distributions

Pareto

Weibull

Regularly Varying

Subexponential

LogNormal
Heavy-tailed phenomena are treated as something mysterious, surprising, & controversial.

1. Properties

2. Emergence

3. Identification
We’ve all been taught that the Normal is “normal”...because of the Central Limit Theorem

But the Central Limit Theorem we’re taught is not complete!
A quick review

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

**Law of Large Numbers (LLN):**

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to E[X_i] \text{ a.s. when } E[X_i] < \infty$$

$$\sum_{i=1}^{n} X_i = nE[X_i] + o(n)$$
A quick review

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

**Central Limit Theorem (CLT):**

$$\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} X_i - nE[X_i] \right) \to Z \sim \text{Normal}(0, \sigma^2)$$

when $\text{Var}[X_i] = \sigma^2 < \infty$.

$$\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$$
A quick review

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT):

$$\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} X_i - nE[X_i] \right) \rightarrow Z \sim \text{Normal}(0, \sigma^2)$$

when $\text{Var}[X_i] = \sigma^2 < \infty$.

Two key assumptions

$$\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$$
A quick review

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT):  
\[
\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} X_i - nE[X_i] \right) \rightarrow Z \sim \text{Normal}(0, \sigma^2)
\]

when $\text{Var}[X_i] = \sigma^2 < \infty$.

$$
\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})
$$

What if $\text{Var}[X_i] = \infty$?
A quick review

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT):

$$\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} X_i - nE[X_i] \right) \to Z \sim \text{Normal}(0, \sigma^2)$$

when $\text{Var}[X_i] = \sigma^2 < \infty$.

$$\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$$

What if $\text{Var}[X_i] = \infty$?
A quick review

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT):

$$\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} X_i - nE[X_i] \right) \to Z \sim Normal(0, \sigma^2)$$
when $Var[X_i] = \sigma^2 < \infty$.

What if $Var[X_i] = \infty$?

The Generalized Central Limit Theorem (GCLT):

$$\frac{1}{a_n} \left( \sum_{i=1}^{n} X_i - b_n \right) \to Z \left\{ \begin{array}{ll}
Normal(0, \sigma^2) \\
\text{Regularly varying } \alpha \in (0,2) \end{array} \right.$$

$$\sum_{i=1}^{n} X_i = nE[X_i] + n^{1/\alpha} Z + o(n^{1/\alpha})$$
A quick review

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT):

$$\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} X_i - nE[X_i] \right) \rightarrow Z \sim Normal(0, \sigma^2)$$

when $\text{Var}[X_i] = \sigma^2 < \infty$.

What if $\text{Var}[X_i] = \infty$?

The Generalized Central Limit Theorem (GCLT):

$$\frac{1}{a_n} \left( \sum_{i=1}^{n} X_i - b_n \right) \rightarrow Z \begin{cases} Normal(0, \sigma^2) \\ \text{Regularly varying } \alpha \in (0,2) \end{cases}$$

Finite variance $\rightarrow$ Light-tailed (Normal)

Infinite variance $\rightarrow$ Heavy-tailed (power law)
Returning to our original question...

Consider i.i.d. $X_i$. How does $\sum_{i=1}^{n} X_i$ grow?

Either the Normal distribution OR a power-law distribution can emerge!
Returning to our original question...

Consider i.i.d. \( X_i \). How does \( \sum_{i=1}^{n} X_i \) grow?

Either the Normal distribution OR a power-law distribution can emerge!

...but this isn’t the only question one can ask about \( \sum_{i=1}^{n} X_i \).

What is the distribution of the "ruin" time?

The ruin time is always heavy-tailed!
What is the distribution of the "ruin" time?

The ruin time is always heavy-tailed!

Consider a symmetric 1-D random walk

The distribution of ruin time satisfies $\Pr(T > x) \sim \frac{\sqrt{2/\pi}}{\sqrt{x}}$
We’ve all been taught that the Normal is “normal”
...because of the Central Limit Theorem

Heavy-tails are more “normal” than the Normal!

1. Additive Processes
2. Multiplicative Processes
3. Extremal Processes
A simple multiplicative process

\[ P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n, \] where \( Y_i \) are i.i.d. and positive

Ex: incomes, populations, fragmentation, twitter popularity...

“Rich get richer”
Multiplicative processes almost always lead to heavy tails

An example:

\[ \begin{align*}
Y_1, Y_2 &\sim \text{Exponential}(\mu) \\
\Pr(Y_1 \cdot Y_2 > x) &\geq \Pr(Y_1 > \sqrt{x})^2 \\
&= e^{-2\mu\sqrt{x}} \\
\Rightarrow \quad Y_1 \cdot Y_2 &\quad \text{is heavy-tailed!}
\end{align*} \]
Multiplicative processes almost always lead to heavy tails

\[ P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n \]

\[ \log P_n = \log Y_1 + \log Y_2 + \cdots + \log Y_n \]

Central Limit Theorem

\[ \log P_n = n E[X_i] + \sqrt{n}Z + o(\sqrt{n}), \text{ where } Z \sim \text{Normal}(0, \sigma^2) \]

when Var\([X_i]\) = \(\sigma^2 < \infty\).

\[ \left(\frac{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}{\mu}\right)^{1/\sqrt{n}} \rightarrow H \sim \text{LogNormal}(0, \sigma^2) \]

where \(\mu = e^{E[\log Y_i]}\)

and Var\([\log Y_i]\) = \(\sigma^2 < \infty\).
Multiplicative central limit theorem

\[
\left( \frac{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}{\mu} \right)^{1/\sqrt{n}} \to H \sim LogNormal(0, \sigma^2)
\]

where \( \mu = e^{E[\log Y_i]} \)
and \( \text{Var}[\log Y_i] = \sigma^2 < \infty \).
Multiplicative central limit theorem

\[
\left( \frac{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}{\mu} \right)^{1/\sqrt{n}} \rightarrow H \sim \text{LogNormal}(0, \sigma^2)
\]

where \( \mu = e^{\mathbb{E}[\log Y_i]} \) and \( \text{Var}[\log Y_i] = \sigma^2 < \infty. \)

Satisfied by all distributions with finite mean and many with infinite mean.
A simple multiplicative process

\[ P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n, \text{ where } Y_i \text{ are i.i.d. and positive} \]

Ex: incomes, populations, fragmentation, twitter popularity...

"Rich get richer"

LogNormals emerge

Heavy-tails
A simple multiplicative process

\[ P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n, \text{ where } Y_i \text{ are i.i.d. and positive} \]

Ex: incomes, populations, fragmentation, twitter popularity...

Multiplicative process with a lower barrier

\[ P_n = \min(P_{n-1}Y_n, \epsilon) \]

Distributions that are approximately power-law emerge

Multiplicative process with noise

\[ P_n = P_{n-1}Y_n + Q_n \]
A simple multiplicative process

\[ P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n, \text{ where } Y_i \text{ are i.i.d. and positive} \]

Ex: incomes, populations, fragmentation, twitter popularity...

Multiplicative process with a lower barrier

\[ P_n = \min(P_{n-1} Y_n, \epsilon) \]

Under minor technical conditions, \( P_n \to F \) such that

\[ \lim_{x \to \infty} \frac{\log F(x)}{\log x} = s^* \]

where \( s^* = \sup(s \geq 0 | E[Y_1^s] \leq 1) \)

"Nearly" regularly varying
We’ve all been taught that the Normal is “normal” ...because of the Central Limit Theorem

Heavy-tails are more “normal” than the Normal!

1. Additive Processes
2. Multiplicative Processes
3. Extremal Processes
A simple extremal process

\[ M_n = \max(X_1, X_2, \ldots, X_n) \]

Ex: engineering for floods, earthquakes, etc. Progression of world records

"Extreme value theory"
\[ M_n = \max(X_1, X_2, \ldots, X_n) \]

**How does \( M_n \) scale?**

\[ \frac{M_n - b_n}{a_n} \]

**A simple example**

\( X_i \sim \text{Exponential}(\mu) \)

\[ \Pr(\max(X_1, \ldots, X_n) > a_n t + b_n) = F(a_n t + b_n)^n \]

\[ = (1 - e^{-a_n t - b_n})^n \]

\[ = (1 - e^{-t - \log n})^n \]

\[ \rightarrow e^{-e^{-t}}: \text{Gumbel distribution} \]
$M_n = \max(X_1, X_2, ..., X_n)$

**How does $M_n$ scale?**

$\frac{M_n - b_n}{a_n}$

"Extremal Central Limit Theorem"

$\frac{M_n - b_n}{a_n} \to Z \begin{cases} 
\text{Frechet} & \text{Heavy-tailed} \\
\text{Weibull} & \text{Heavy or light-tailed} \\
\text{Gumbel} & \text{Light-tailed}
\end{cases}$
How does $M_n$ scale?

$M_n = \max(X_1, X_2, \ldots, X_n)$

"Extremal Central Limit Theorem"

$$\frac{M_n - b_n}{a_n} \rightarrow Z \left\{ \begin{array}{ll}
\text{Frechet} & \text{iff } X_i \text{ are regularly varying} \\
\text{Weibull} & \text{e.g. when } X_i \text{ are Uniform} \\
\text{Gumbel} & \text{e.g. when } X_i \text{ are LogNormal}
\end{array} \right.$$
A simple extremal process

\[ M_n = \max(X_1, X_2, \ldots, X_n) \]

Ex: engineering for floods, earthquakes, etc. Progression of world records

Either heavy-tailed or light-tailed distributions can emerge as \( n \to \infty \)

...but this isn’t the only question one can ask about \( M_n \).

What is the distribution of the time until a new “record” is set?

The time until a record is always heavy-tailed!
The time until a record is always heavy-tailed!

$T_k$: Time between $k$ & $k + 1^{st}$ record

$$\Pr(T_k > n) \sim \frac{2^{k-1}}{n}$$

What is the distribution of the time until a new “record” is set?

The time until a record is always heavy-tailed!
We’ve all been taught that the Normal is “normal”...
because of the Central Limit Theorem

Heavy-tails are more “normal” than the Normal!

1. Additive Processes
2. Multiplicative Processes
3. Extremal Processes
Heavy-tailed phenomena are treated as something mysterious, surprising, and controversial.

1. Properties

2. Emergence

3. Identification
Heavy-tailed phenomena are treated as something **Mysteries, Surprising, & Controversial**

1999 Sigcomm paper – 4500+ citations!

**On Power-Law Relationships of the Internet Topology**

Michalis Faloutsos  
U.C. Riverside  
Dept. of Comp. Science  
michalis@cs.ucr.edu

Petros Faloutsos  
U. of Toronto  
Dept. of Comp. Science  
pfal@cs.toronto.edu

Christos Faloutsos *  
Carnegie Mellon Univ.  
Dept. of Comp. Science  
christos@cs.cmu.edu

BUT...

Similar stories in electricity nets, citation nets, ...

2005, STOC

On the Bias of Traceroute Sampling

or, Power-law Degree Distributions in Regular Graphs

Dimitris Achilleos  
Microsoft Research  
Redmond, WA 98052  
oplas@microsoft.com

David Kempe  
Department of Computer Science  
University of Southern California  
Los Angeles, CA 90089  
dkempe@usc.edu

Aaron Clauset  
Department of Computer Science  
University of New Mexico  
Albuquerque, NM 87131  
aaron@cs.unm.edu

Christopher Moore  
Department of Computer Science  
University of New Mexico  
Albuquerque, NM 87131  
mbores@cs.unm.edu

Understanding Internet Topology: Principles, Models, and Validation

David Alderson, Member, IEEE, Lun Li, Student Member, IEEE, Walter Willinger, Fellow, IEEE, and  
John C. Doyle, Member, IEEE

2005, ToN
A “typical” approach for identifying of heavy tails: Linear Regression

Heavy-tailed or light-tailed?

“frequency plot”
A “typical” approach for identifying heavy tails: Linear Regression

Heavy-tailed or light-tailed?

log-linear scale

Log(frequency)

value
A “typical” approach for identifying of heavy tails: **Linear Regression**

Heavy-tailed or light-tailed?

**log-linear scale**

\[ f(x) = Ce^{-\mu x} \quad \Rightarrow \quad \log f(x) = \log C - \mu x \]
A “typical” approach for identifying of heavy tails: **Linear Regression**

Heavy-tailed or light-tailed?

- **Heavy-tailed**
  - Log-log scale:
    - \( f(x) = Cx^{-\alpha-1} \rightarrow \log f(x) = \log C - (\alpha + 1)\log x \)
    - **Power-law tail**

- **Light-tailed**
  - Log-linear scale:
    - \( f(x) = Ce^{-\mu x} \rightarrow \log f(x) = \log C - \mu x \)
    - **Exponential tail**

Linear ⇒ Power-law tail

Linear ⇒ Exponential tail
$f(x) = Cx^{-\alpha - 1}$ \implies \log f(x) = \log C - (\alpha + 1) \log x$

Regression $\Rightarrow$ Estimate of tail index ($\alpha$)

Log-log scale

Linear $\Rightarrow$ Power-law tail
Is it really linear?
Is the estimate of $\alpha$ accurate?

$$f(x) = Cx^{-\alpha - 1} \rightarrow \log f(x) = \log C - (\alpha + 1) \log x$$

Log-log scale

Linear $\Rightarrow$ Power-law tail

Regression $\Rightarrow$

Estimate of tail index ($\alpha$)
Pr(\(X > x\)) = \(\bar{F}(x) = C' x^\alpha\)

\(f(x) = C x^{-\alpha -1}\)

\(\log f(x) = \log C - (\alpha + 1) \log x\)

\(\hat{\alpha} = 1.9\)

\(\hat{\alpha} = 1.4\)

True \(\alpha = 2.0\)
This simple change is extremely important...

Pr(\(X > x\))

"rank plot"

log-log scale

"frequency plot"
This simple change is extremely important...

The data is from an Exponential!
This mistake has happened A LOT!

Electricity grid degree distribution

\[ \alpha = 3 \] (from Science)

\[ \Pr(X > x) \]

“rank plot”

log-linear scale

“frequency plot”

log-log scale
This mistake has happened A LOT!

WWW degree distribution

Pr($X > x$)

α = 1.7

“rank plot”

log-log scale

log-log scale

α = 1.1

(from Science)

“frequency plot”
This simple change is extremely important...

But, this is still an error-prone approach

Pr(X > x)

“rank plot”

Regression ⇒
Estimate of tail index (α)

Log-log scale

Linear ⇒
Power-law tail

...other distributions can be nearly linear too

Lognormal

Weibull
This simple change is extremely important...
But, this is still an error-prone approach.

Regression $\Rightarrow$ Estimate of tail index ($\alpha$)

...assumptions of regression are not met
...tail is much noisier than the body

Log-log scale

Pr($X > x$)

“rank plot”

Linear $\Rightarrow$ Power-law tail

...other distributions can be nearly linear too
A completely different approach: Maximum Likelihood Estimation (MLE)

What is the $\alpha$ for which the data is most “likely”?

$$L(x; \alpha) = \prod_{i=1}^{n} \frac{\alpha x_{\text{min}}^\alpha}{x_i^{\alpha+1}}$$

$$\log L(x; \alpha) = \sum_{i=1}^{n} \log(\alpha x_{\text{min}}^\alpha) - \log x_i^\alpha$$

Maximizing gives $\hat{\alpha}_{MLE} = \frac{n}{\sum_{i=1}^{n} \log(x_i/x_{\text{min}})}$

This has many nice properties:
- $\hat{\alpha}_{MLE}$ is the minimal variance, unbiased estimator.
- $\hat{\alpha}_{MLE}$ is asymptotically efficient.
A completely different approach: **Maximum Likelihood Estimation (MLE)**

**Weighted Least Squares Regression (WLS)**

asymptotically for large data sets, when weights are chosen as $w_i = 1 / (\log x_i - \log x_0)$.

$$\hat{a}_{WLS} = \frac{- \sum_{i=1}^{n} \log(\hat{r}_i/n)}{\sum_{i=1}^{n} \log(x_i/x_0)}$$

$$\sim \frac{n}{\sum_{i=1}^{n} \log(x_i/x_0)}$$

$$= \hat{a}_{MLE}$$
A completely different approach: Maximum Likelihood Estimation (MLE)

Weighted Least Squares Regression (WLS)

asymptotically for large data sets, when weights are chosen as $w_i = 1/(\log x_i - \log x_0)$.

“Listen to your body”
A quick summary of where we are:

Suppose data comes from a power-law (Pareto) distribution \( F(x) = \left( \frac{x_0}{x} \right) ^ \alpha \).

Then, we can identify this visually with a log-log plot, and we can estimate \( \alpha \) using either MLE or WLS.
Suppose data comes from a power-law (Pareto) distribution $F(x) = \left(\frac{x_0}{x}\right)^\alpha$.

Then, we can identify this visually with a log-log plot, and we can estimate $\alpha$ using either MLE or WLS.

What if the data is not exactly a power-law?
What if only the tail is power-law?

But, where does the tail start?
Can we just use MLE/WLS on the “tail”?

Impossible to answer...
An example

Suppose we have a mixture of power laws:

$$\bar{F}(x) = q\bar{F}_1(x) + (1 - q)\bar{F}_2(x)$$

$\alpha_1 < \alpha_2$

We want $\hat{\alpha}_{MLE} \rightarrow \alpha_1$ as $n \rightarrow \infty$.

...but, suppose we use $x_{\text{min}}$ as our cutoff:

$$\frac{1}{\hat{\alpha}_{MLE}} \rightarrow \frac{q\bar{F}_1(x_{\text{min}})}{\alpha_1 \bar{F}(x_{\text{min}})} + \frac{(1 - q)\bar{F}_2(x_{\text{min}})}{\alpha_2 \bar{F}(x_{\text{min}})} \neq \alpha_1$$
Identifying power-law distributions
“Listen to your body”

v.s.

Identifying power-law tails
“Let the tail do the talking”

MLE/WLS

v.s.

Extreme value theory
Returning to our example

Suppose we have a mixture of power laws:

$$\bar{F}(x) = q\bar{F}_1(x) + (1 - q)\bar{F}_2(x)$$

\[\alpha_1 < \alpha_2\]

We want $\hat{\alpha}_{MLE} \to \alpha_1$ as $n \to \infty$.

...but, suppose we use $x_{\text{min}}$ as our cutoff:

$$\frac{1}{\hat{\alpha}_{MLE}} = \frac{q\bar{F}_1(x_{\text{min}})}{\alpha_1\bar{F}(x_{\text{min}})} + \frac{(1 - q)\bar{F}_2(x_{\text{min}})}{\alpha_2\bar{F}(x_{\text{min}})}$$

The bias disappears as $x_{\text{min}} \to \infty$!
The idea: **Improve robustness by throwing away nearly all the data!**

\[ x_{\text{min}} \rightarrow x_{\text{min}}(n), \quad \text{where } x_{\text{min}}(n) \rightarrow \infty \text{ as } n \rightarrow \infty. \]

\[
\begin{align*}
+ \text{ Larger } x_{\text{min}}(n) & \Rightarrow \text{ Small bias} \\
- \text{ Larger } x_{\text{min}}(n) & \Rightarrow \text{ Larger variance}
\end{align*}
\]
The idea: Improve robustness by throwing away nearly all the data!

\[ x_{\text{min}} \rightarrow x_{\text{min}}(n), \text{ where } x_{\text{min}}(n) \rightarrow \infty \text{ as } n \rightarrow \infty. \]

The Hill Estimator

\[ \hat{\alpha}(k, n) = \frac{1}{k} \sum_{i=1}^{k} \log \left( \frac{x(i)}{x(k)} \right) \]

where \( x(k) \) is the \( k \)th largest data point.

Looks almost like the MLE, but uses order \( k \)th order statistic.
The idea: **Improve robustness by throwing away nearly all the data!**

\[ x_{\text{min}} \rightarrow x_{\text{min}}(n), \text{ where } x_{\text{min}}(n) \rightarrow \infty \text{ as } n \rightarrow \infty. \]

**The Hill Estimator**

\[ \hat{\alpha}(k, n) = \frac{1}{k} \sum_{i=1}^{k} \log \left( \frac{x(i)}{x(k)} \right) \]

where \( x(k) \) is the \( k \)th largest data point.

Looks almost like the MLE, but uses order \( k \)th order statistic

\[ \hat{\alpha}(k, n) \rightarrow \alpha \text{ as } n \rightarrow \infty \text{ if } \frac{k(n)}{n} \rightarrow 0 \text{ & } k(n) \rightarrow \infty \]

...how do we choose \( k \)?

\[ \text{throw away nearly all the data, but keep enough data for consistency} \]
The idea: **Improve robustness by throwing away nearly all the data!**

$x_{\text{min}} \rightarrow x_{\text{min}}(n)$, where $x_{\text{min}}(n) \rightarrow \infty$ as $n \rightarrow \infty$.

The Hill Estimator

$\hat{\alpha}(k, n) = \frac{1}{k} \sum_{i=1}^{k} \log \left( \frac{x(i)}{x(k)} \right)$

where $x(k)$ is the $k$th largest data point.

Looks almost like the MLE, but uses order $k$th order statistic

...how do we choose $k$?

$\hat{\alpha}(k, n) \rightarrow \alpha$ as $n \rightarrow \infty$ if $k(n)/n \rightarrow 0$ & $k(n) \rightarrow \infty$.

Throw away everything except the outliers!
Choosing $k$ in practice: The Hill plot
Choosing $k$ in practice: The Hill plot

- Pareto, $\alpha = 2$
- Exponential
Choosing $k$ in practice: The Hill plot

Pareto, $\alpha = 2$

Mixture, with Pareto-tail, $\alpha = 2$
...but the hill estimator has problems too

This data is from TCP flow sizes!
Identifying power-law distributions
“Listen to your body.”

MLE/WLS

Identifying power-law tails
“Let the tail do the talking.”

Hill estimator

It’s dangerous to rely on any one technique!
(see our forthcoming book for other approaches)
Heavy-tailed phenomena are treated as something mysterious, surprising, & controversial.

1. Properties
2. Emergence
3. Identification
Heavy-tailed phenomena are treated as something **Mysterious, Surprising, & Controversial**

1. **Properties**
   - Pareto
   - Subexponential
   - Weibull
   - LogNormal
   - Regularly Varying

2. **Emergence**

3. **Identification**

Heavy-tailed distributions have many beautiful & strange properties:

1. Scale Invariance → Regularly Varying distributions
2. The “catastrophe principle” → Subexponential distributions
3. Residual lives “blow up” → Long-tailed distributions
Heavy-tailed phenomena are treated as something **Mysterious, Surprising, & Controversial**

1. Properties
2. Emergence
3. Identification

We’ve all been taught that the Normal is “normal” because of the Central Limit Theorem, **BUT**

Heavy-tails are more “normal” than the Normal!
Heavy-tailed phenomena are treated as something mystery, surprising, and controversial.

1. Properties

2. Emergence

3. Identification

- Identifying power-law distributions
  - "Listen to your body"

- Identifying power-law tails
  - "Let the tail do the talking"

- MLE/WLS
- Hill estimator
  - ...and others we didn't talk about
The Fundamentals of Heavy Tails

Properties, Emergence, & Identification

Jayakrishnan Nair, Adam Wierman, Bert Zwart

“The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking […] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world.”

– Clay Shirky, as quoted in Newsweek & the Guardian