Inference with Heavy-Tailed Distributions

Cosma Shalizi

Statistics Department, Carnegie Mellon University

Santa Fe Institute

15 May 2012

Summary

Summary

1. Everything good in the talk I owe to my co-authors, Aaron Clauset and Mark Newman

Summary

1. Everything good in the talk I owe to my co-authors, Aaron Clauset and Mark Newman
2. Power law distributions, $p(x) \propto x^{-\alpha}$, are cool, but not *that* cool

Summary

1. Everything good in the talk I owe to my co-authors, Aaron Clauset and Mark Newman
2. Power law distributions, $p(x) \propto x^{-\alpha}$, are cool, but not *that* cool
3. Most of the studies claiming to find them use unreliable 19th century methods, and have no value as evidence either way

Summary

1. Everything good in the talk I owe to my co-authors, Aaron Clauset and Mark Newman
2. Power law distributions, $p(x) \propto x^{-\alpha}$, are cool, but not *that* cool
3. Most of the studies claiming to find them use unreliable 19th century methods, and have no value as evidence either way
4. Reliable methods exist, and need only very straightforward mid-20th century statistics

Summary

1. Everything good in the talk I owe to my co-authors, Aaron Clauset and Mark Newman
2. Power law distributions, $p(x) \propto x^{-\alpha}$, are cool, but not *that* cool
3. Most of the studies claiming to find them use unreliable 19th century methods, and have no value as evidence either way
4. Reliable methods exist, and need only very straightforward mid-20th century statistics
5. Using reliable methods, *lots* of the claimed power laws disappear, or are at best "not proven"

Summary

1. Everything good in the talk I owe to my co-authors, Aaron Clauset and Mark Newman
2. Power law distributions, $p(x) \propto x^{-\alpha}$, are cool, but not *that* cool
3. Most of the studies claiming to find them use unreliable 19th century methods, and have no value as evidence either way
4. Reliable methods exist, and need only very straightforward mid-20th century statistics
5. Using reliable methods, *lots* of the claimed power laws disappear, or are at best "not proven"

You are now free to tune me out and turn on social media

What Are Power Law Distributions? Why Care?

$$\begin{aligned} p(x) &\propto x^{-\alpha} \text{ (continuous)} \\ \mathrm{P}(X = x) &\propto x^{-\alpha} \text{ (discrete)} \\ \therefore \mathrm{P}(X \geq x) &\propto x^{-(\alpha-1)} \end{aligned}$$

and

$$\log p(x) = \log C - \alpha \log x$$

What Are Power Law Distributions? Why Care?

$$\begin{aligned} p(x) &\propto x^{-\alpha} \text{ (continuous)} \\ \mathrm{P}(X = x) &\propto x^{-\alpha} \text{ (discrete)} \\ \therefore \mathrm{P}(X \geq x) &\propto x^{-(\alpha-1)} \end{aligned}$$

and

$$\log p(x) = \log C - \alpha \log x$$

"Pareto" (continuous), "Zipf" or "zeta" (discrete)

What Are Power Law Distributions? Why Care?

$$\begin{aligned} p(x) &\propto x^{-\alpha} \text{ (continuous)} \\ \mathrm{P}(X = x) &\propto x^{-\alpha} \text{ (discrete)} \\ \therefore \mathrm{P}(X \geq x) &\propto x^{-(\alpha-1)} \end{aligned}$$

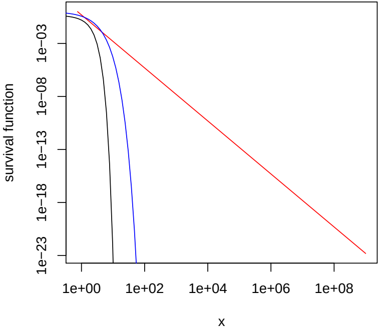
and

$$\log p(x) = \log C - \alpha \log x$$

"Pareto" (continuous), "Zipf" or "zeta" (discrete)
Explicitly:

$$p(x) = \frac{\alpha - 1}{x_{\mathrm{min}}} \left(\frac{x}{x_{\mathrm{min}}} \right)^{-\alpha}$$

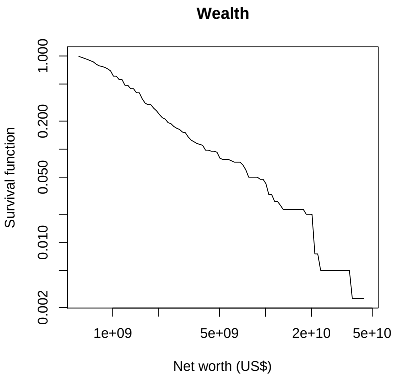
(discrete version involves the Hurwitz zeta function)


survival function
1e−03
1e−08
1e−13
1e−18
1e−23
1e+00
1e+02
1e+04
1e+06
1e+08
x

Money, Words, Cities

The three classic power law distributions

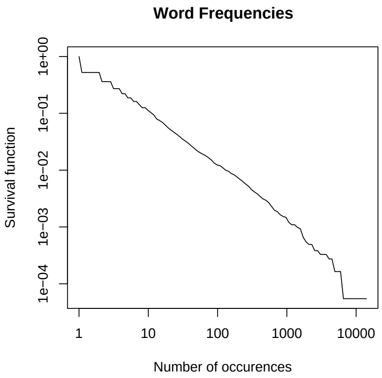
Pareto's law: wealth (richest 400 in US, 2003)



Money, Words, Cities

The three classic power law distributions

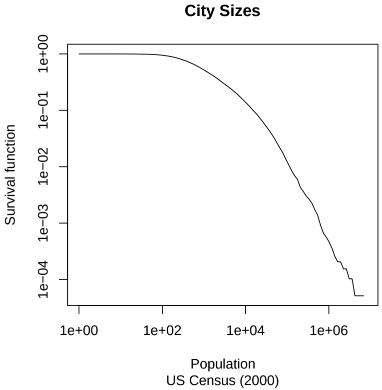
Zipf's law: word frequencies (*Moby Dick*)



Money, Words, Cities

The three classic power law distributions

Zipf's law: city populations



Properties

Highly right skewed

Properties

Highly right skewed
Heavy (fat, long, ...) tails: sub-exponential decay of $p(x)$

Properties

Highly right skewed

Heavy (fat, long, ...) tails: sub-exponential decay of $p(x)$

Extreme inequality ("80/20"): high proportion of summed values comes from small fraction of samples/population

Properties

Highly right skewed

Heavy (fat, long, ...) tails: sub-exponential decay of $p(x)$

Extreme inequality ("80/20"): high proportion of summed values comes from small fraction of samples/population

"Scale-free":

$$p(x|X \geq s) = \frac{\alpha - 1}{s}\left(\frac{x}{s}\right)^{-\alpha}$$

i.e., another power law, same α

Properties

Highly right skewed
Heavy (fat, long, ...) tails: sub-exponential decay of $p(x)$
Extreme inequality ("80/20"): high proportion of summed values comes from small fraction of samples/population
"Scale-free":

$$p(x|X \geq s) = \frac{\alpha - 1}{s}\left(\frac{x}{s}\right)^{-\alpha}$$

i.e., another power law, same α
$\therefore$ no "typical scale"

Properties

Highly right skewed

Heavy (fat, long, ...) tails: sub-exponential decay of $p(x)$

Extreme inequality ("80/20"): high proportion of summed values comes from small fraction of samples/population

"Scale-free":

$$p(x|X \geq s) = \frac{\alpha - 1}{s}\left(\frac{x}{s}\right)^{-\alpha}$$

i.e., another power law, same α

$\therefore$ no "typical scale"

though $x_{\min}$ is the typical value

Origin Myths

Catchy and mysterious origin myth from physics:

- Distinct phases co-exist at phase transitions
- $\therefore$ Each phase can appear by fluctuation inside the other, and vice versa
- $\therefore$ Infinite-range correlations in space and time
- $\therefore$ Central limit theorem breaks down
- but macroscopic physical quantities are still averages
- $\therefore$ they must have a scale-free distribution
- So critical phenomena $\Rightarrow$ power laws

Origin Myths (cont.)

Deflating origin myths:

Piles of papers on my office floor [1, 2, 3]

- I start new piles at rate λ, so age of piles $\sim \mathrm{Exponential}(\lambda)$
- All piles start with size x_{min}
- Once a pile starts, on average it grows exponentially at rate μ
- $X \sim \mathrm{Pareto}(\lambda/\mu + 1, x_{\mathrm{min}})$

Origin Myths (cont.)

Deflating origin myths:

Piles of papers on my office floor [1, 2, 3]

- I start new piles at rate λ, so age of piles $\sim \mathrm{Exponential}(\lambda)$
- All piles start with size x_{min}
- Once a pile starts, on average it grows exponentially at rate μ
- $X \sim \mathrm{Pareto}(\lambda/\mu + 1, x_{\mathrm{min}})$

Mixtures of exponentials work too [4]

There are lots of claims that things follow power laws, especially in the last ≈ 20 years, especially from physicists

There are lots of claims that things follow power laws, especially in the last ≈ 20 years, especially from physicists

> *word frequency, protein interaction degree (yeast), metabolic network degree (*E. coli*), Internet autonomous system network, calls received, intensity of wars, terrorist attack fatalities, bytes per HTTP request, species per genus, # sightings per bird species, population affected by blackouts, sales of best-sellers, population of US cities, area of wildfires, solar flare intensity, earthquake magnitude, religious sect size, surname frequency, individual net worth, citation counts, # papers authored, # hits per URL, in-degree per URL, # entries in e-mail address books, . . .*

$\Rightarrow$ Mason Porter's Power Law Shop


I WENT TO A PHYSICS
CONFERENCE AND ALL
I GOT WAS A LOUSY
POWER LAW.

How do physicists come up with their power laws?

Remember

$$\log p(x) = \log C - \alpha \log x$$

& similarly for the CDF

How do physicists come up with their power laws?

Remember

$$\log p(x) = \log C - \alpha \log x$$

& similarly for the CDF

Suggests:

- Take a log-log plot of the histogram, or of the CDF, and
- Fit an ordinary regression line, then
- Use fitted slope as guess for α, check goodness of fit by R^2

How do physicists come up with their power laws?

Remember

$$\log p(x) = \log C - \alpha \log x$$

& similarly for the CDF

Suggests:

- Take a log-log plot of the histogram, or of the CDF, and
- Fit an ordinary regression line, then
- Use fitted slope as guess for α, check goodness of fit by R^2

This is a clever idea

How do physicists come up with their power laws?

Remember

$$\log p(x) = \log C - \alpha \log x$$

& similarly for the CDF

Suggests:

- Take a log-log plot of the histogram, or of the CDF, and
- Fit an ordinary regression line, then
- Use fitted slope as guess for α, check goodness of fit by R^2

This is a clever idea for the 1890s

How do physicists come up with their power laws?

Remember

$$\log p(x) = \log C - \alpha \log x$$

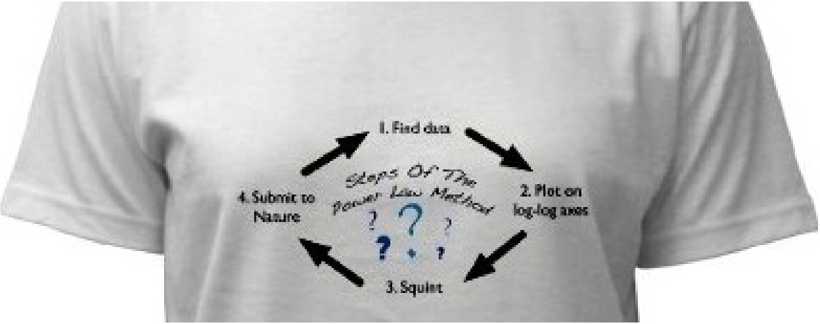
& similarly for the CDF

Suggests:

- Take a log-log plot of the histogram, or of the CDF, and
- Fit an ordinary regression line, then
- Use fitted slope as guess for α, check goodness of fit by R^2

This is a clever idea for the 1890s

Fun fact: "statistical physics" involves no actual statistics


1. Find data
2. Plot on log-log axes
3. Squint
4. Submit to Nature
Steps Of The Power Law Method

Why Is This Bad?

Why Is This Bad?

Histograms: binning always throws away information, adds lots of error

log-sized bins are only infinitessimally better

Why Is This Bad?

Histograms: binning always throws away information, adds lots of error

log-sized bins are only infinitessimally better

CDF or rank-size plot: values are *not independent*; inefficient

Why Is This Bad?

Histograms: binning always throws away information, adds lots of error

log-sized bins are only infinitessimally better

CDF or rank-size plot: values are *not independent*; inefficient

Least-squares line:

- Not a normalized distribution,
- All the inferential assumptions for regression fail
- Always has avoidable error as an estimate of α
- Easily get large R^2 for non-power-law distributions

Some Distributions Which Are Not Power Laws

Log-normal: $\ln X \sim \mathcal{N}(\mu, \sigma^2)$:

Some Distributions Which Are Not Power Laws

Log-normal: $\ln X \sim \mathcal{N}(\mu, \sigma^2)$:

$$p(x) = \frac{1}{(1 - \Phi(\frac{\ln x_{\mathrm{min}} - \mu}{\sigma}))x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

Some Distributions Which Are Not Power Laws

Log-normal: $\ln X \sim \mathcal{N}(\mu, \sigma^2)$:

$$p(x) = \frac{1}{(1 - \Phi(\frac{\ln x_{\mathrm{min}} - \mu}{\sigma}))x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

Stretched exponential/Weibull: $X^{\beta} \sim \mathrm{Exponential}(\lambda)$

Some Distributions Which Are Not Power Laws

Log-normal: $\ln X \sim \mathcal{N}(\mu, \sigma^2)$:

$$p(x) = \frac{1}{(1 - \Phi(\frac{\ln x_{\mathrm{min}} - \mu}{\sigma}))x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

Stretched exponential/Weibull: $X^\beta \sim \mathrm{Exponential}(\lambda)$

$$p(x) = \beta\lambda e^{\lambda x_{\mathrm{min}}^{\beta}} x^{\beta - 1} e^{-\lambda x^\beta}$$

Power law with exponential cut-off ("negative gamma")

Some Distributions Which Are Not Power Laws

Log-normal: $\ln X \sim \mathcal{N}(\mu, \sigma^2)$:

$$p(x) = \frac{1}{(1 - \Phi(\frac{\ln x_{\min} - \mu}{\sigma}))x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

Stretched exponential/Weibull: $X^\beta \sim \mathrm{Exponential}(\lambda)$

$$p(x) = \beta\lambda e^{\lambda x_{\min}^\beta} x^{\beta-1} e^{-\lambda x^\beta}$$

Power law with exponential cut-off ("negative gamma")

$$p(x) = \frac{1/L}{\Gamma(1 - \alpha, x_{\min}/L)} (x/L)^{-\alpha} e^{-x/L}$$

Some Distributions Which Are Not Power Laws

Log-normal: $\ln X \sim \mathcal{N}(\mu, \sigma^2)$:

$$p(x) = \frac{1}{(1 - \Phi(\frac{\ln x_{\mathrm{min}} - \mu}{\sigma}))x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

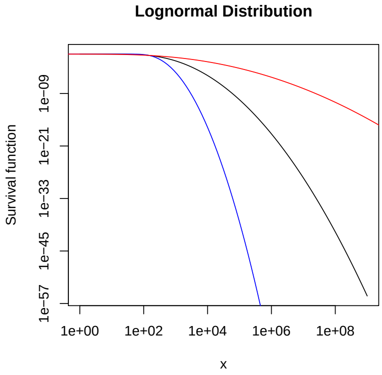
Stretched exponential/Weibull: $X^\beta \sim \mathrm{Exponential}(\lambda)$

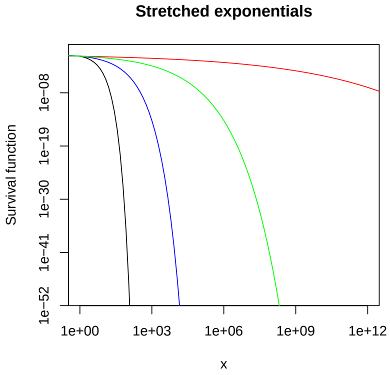
$$p(x) = \beta\lambda e^{\lambda x_{\mathrm{min}}^{\beta}} x^{\beta-1} e^{-\lambda x^\beta}$$

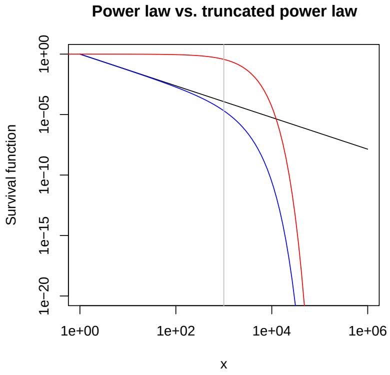
Power law with exponential cut-off ("negative gamma")

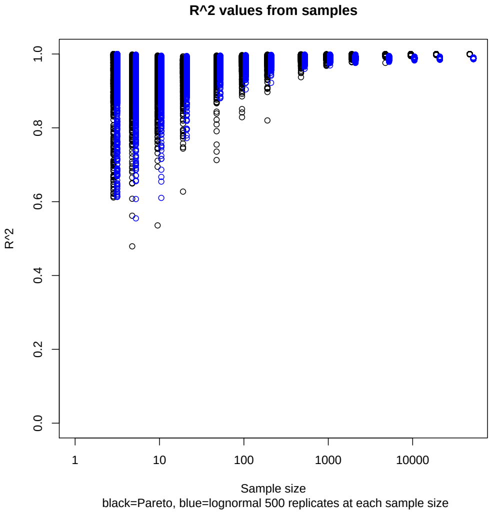
$$p(x) = \frac{1/L}{\Gamma(1-\alpha, x_{\mathrm{min}}/L)}(x/L)^{-\alpha} e^{-x/L}$$

like a power law for $x \ll L$, like an exponential for $x \gg L$


Lognormal Distribution
Survival function
1e−09
1e−21
1e−33
1e−45
1e−57
1e+00
1e+02
1e+04
1e+06
1e+08
x


Stretched exponentials
Survival function
1e−08
1e−19
1e−30
1e−41
1e−52
1e+00
1e+03
1e+06
1e+09
1e+12
x


Power law vs. truncated power law
Survival function
1e+00
1e−05
1e−10
1e−15
1e−20
1e+00
1e+02
1e+04
1e+06
x



R^2 for a log normal (limiting value > 0.9)

Abusing linear regression makes the baby Gauss cry

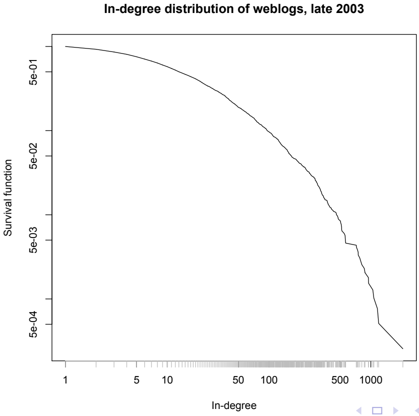
Blogospheric Navel-Gazing

Shirky [5]: in-degree of weblogs follows a power-law, many consequences for media ecology, etc., etc.

Data via [6]

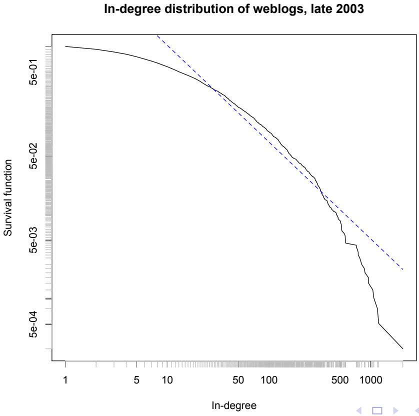
Blogospheric Navel-Gazing

Shirky [5]: in-degree of weblogs follows a power-law, many consequences for media ecology, etc., etc.



Blogospheric Navel-Gazing

Shirky [5]: in-degree of weblogs follows a power-law, many consequences for media ecology, etc., etc.



Estimating the Exponent

Use maximum likelihood

Estimating the Exponent

Use maximum likelihood

$$\mathcal{L}(\alpha, x_{\min}) = n \log \frac{\alpha - 1}{x_{\min}} - \alpha \sum_{i=1}^{n} \log \frac{x_i}{x_{\min}}$$

Estimating the Exponent

Use maximum likelihood

$$\begin{aligned}
\mathcal{L}(\alpha, x_{\min}) &= n \log \frac{\alpha - 1}{x_{\min}} - \alpha \sum_{i=1}^{n} \log \frac{x_i}{x_{\min}} \\
\frac{\partial}{\partial \alpha}\mathcal{L} &= \frac{n}{\alpha - 1} - \sum_{i=1}^{n} \log \frac{x_i}{x_{\min}}
\end{aligned}$$

Estimating the Exponent

Use maximum likelihood

$$
\begin{aligned}
\mathcal{L}(\alpha, x_{\min}) &= n \log \frac{\alpha - 1}{x_{\min}} - \alpha \sum_{i=1}^{n} \log \frac{x_i}{x_{\min}} \\
\frac{\partial}{\partial \alpha}\mathcal{L} &= \frac{n}{\alpha - 1} - \sum_{i=1}^{n} \log \frac{x_i}{x_{\min}} \\
\widehat{\alpha} &= 1 + \frac{n}{\sum_{i=1}^{n} \log x_i / x_{\min}}
\end{aligned}
$$

Properties of the MLE

Consistent: $\widehat{\alpha} \rightarrow \alpha$

Properties of the MLE

Consistent: $\widehat{\alpha} \rightarrow \alpha$
Standard error: $\mathrm{Var}\left[\widehat{\alpha}\right] = n^{-1}(\alpha - 1)^2 + O(n^{-2})$

Properties of the MLE

Consistent: $\widehat{\alpha} \rightarrow \alpha$

Standard error: $\mathrm{Var}\left[\widehat{\alpha}\right] = n^{-1}(\alpha - 1)^2 + O(n^{-2})$

Efficient: no consistent alternative with less variance

In particular, dominates regression

Properties of the MLE

Consistent: $\widehat{\alpha} \rightarrow \alpha$
Standard error: $\mathrm{Var}\left[\widehat{\alpha}\right] = n^{-1}(\alpha - 1)^2 + O(n^{-2})$
Efficient: no consistent alternative with less variance

In particular, dominates regression

Asymptotically Gaussian: $\widehat{\alpha} \rightsquigarrow \mathcal{N}(\alpha, \frac{(\alpha-1)^2}{n})$

Properties of the MLE

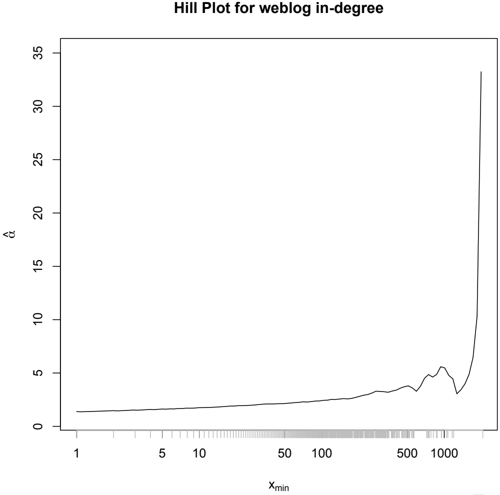
Consistent: $\widehat{\alpha} \rightarrow \alpha$
Standard error: $\mathrm{Var}\left[\widehat{\alpha}\right] = n^{-1}(\alpha-1)^2 + O(n^{-2})$
Efficient: no consistent alternative with less variance
In particular, dominates regression
Asymptotically Gaussian: $\widehat{\alpha} \rightsquigarrow \mathcal{N}(\alpha, \frac{(\alpha-1)^2}{n})$
Ancient: Worked out in the 1950s [7, 8]

Properties of the MLE

Consistent: $\widehat{\alpha} \rightarrow \alpha$
Standard error: $\mathrm{Var}\left[\widehat{\alpha}\right] = n^{-1}(\alpha - 1)^2 + O(n^{-2})$
Efficient: no consistent alternative with less variance
In particular, dominates regression
Asymptotically Gaussian: $\widehat{\alpha} \rightsquigarrow \mathcal{N}(\alpha, \frac{(\alpha-1)^2}{n})$
Ancient: Worked out in the 1950s [7, 8]
Computationally trivial

$\widehat{\alpha}$ depends on $x_{\min}$; "Hill" plot [9]

$\widehat{\alpha}$ depends on $x_{\min}$; "Hill" plot [9]



Estimating the Scaling Region

Maximizing likelihood over x_{min} leads to trouble (try it and see)

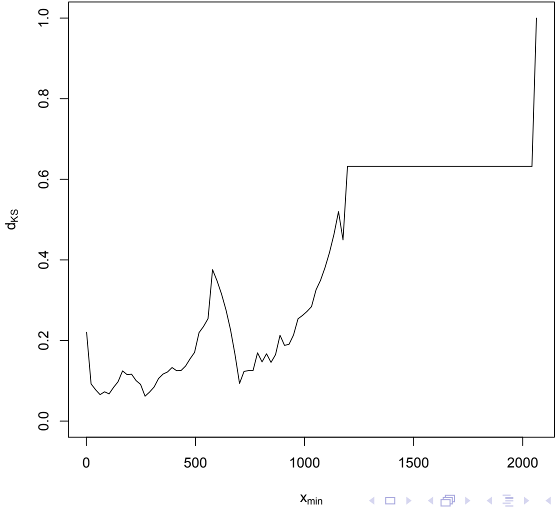
Estimating the Scaling Region

Maximizing likelihood over x_{min} leads to trouble (try it and see)
Only want the scaling region in the tail anyway

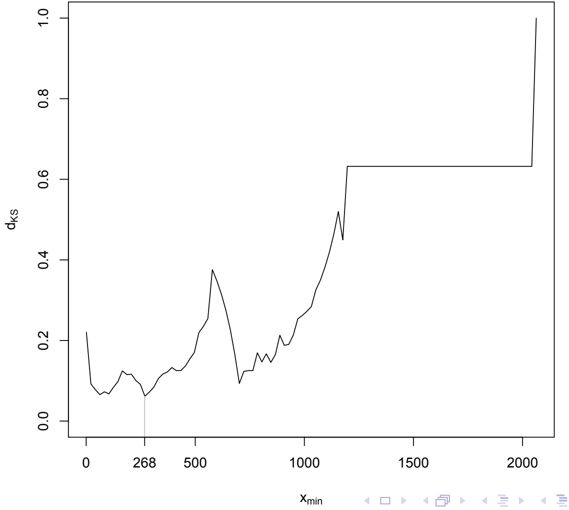
Estimating the Scaling Region

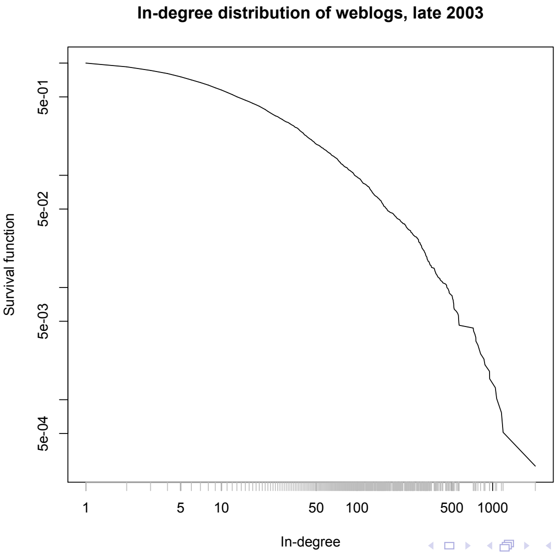
Maximizing likelihood over $x_{\min}$ leads to trouble (try it and see)
Only want the scaling region in the tail anyway
Minimize discrepancy between fitted and empirical distributions [10]:

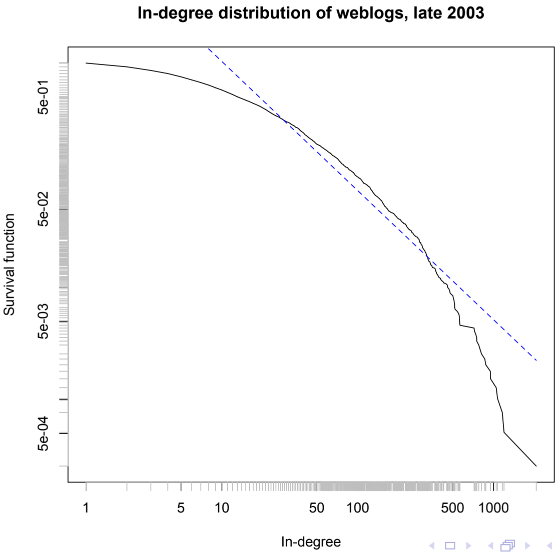
$$\begin{aligned}\widehat{x_{\min}} &= \operatorname*{argmin}_{x_{\min}} \max_{x \geq x_{\min}} |\hat{P}_n(x) - P(x; \widehat{\alpha}, x_{\min})| \\ &= \operatorname*{argmin}_{x_{\min}} d_{KS}(\hat{P}_n, P(\widehat{\alpha}, x_{\min}))\end{aligned}$$

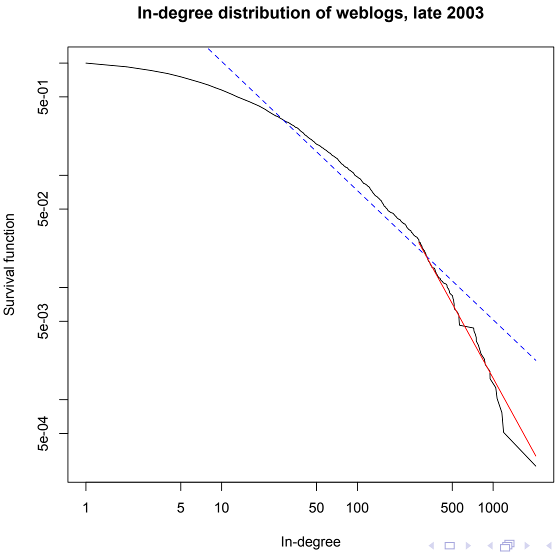

1.0
0.8
0.6
0.4
0.2
0.0
dKS
0
500
1000
1500
2000
xmin

top 2.8%




In-degree distribution of weblogs, late 2003
Survival function
5e-01
5e-02
5e-03
5e-04
1
5
10
50
100
500
1000
In-degree


In-degree distribution of weblogs, late 2003
Survival function
5e-01
5e-02
5e-03
5e-04
1
5
10
50
100
500
1000
In-degree


In-degree distribution of weblogs, late 2003
Survival function
5e-01
5e-02
5e-03
5e-04
1
5
10
50
100
500
1000
In-degree

Goodness-of-Fit

How can we tell if it's a good fit or not, if we can't use R^2?

Goodness-of-Fit

How can we tell if it's a good fit or not, if we can't use R^2?

You shouldn't use R^2 that way for a regression

Goodness-of-Fit

How can we tell if it's a good fit or not, if we can't use R^2?

You shouldn't use R^2 that way for a regression

Use a goodness-of-fit test!

Goodness-of-Fit

How can we tell if it's a good fit or not, if we can't use R^2?

You shouldn't use R^2 that way for a regression

Use a goodness-of-fit test!

Kolmogorov-Smirnov statistic is nice: for CDFs P, Q

$$d_{KS}(P,Q) = \max_x |P(x) - Q(x)|$$

Compare empirical CDF to theoretical one

Goodness-of-Fit

How can we tell if it's a good fit or not, if we can't use R^2?

You shouldn't use R^2 that way for a regression

Use a goodness-of-fit test!

Kolmogorov-Smirnov statistic is nice: for CDFs P, Q

$$d_{KS}(P,Q) = \max_x |P(x) - Q(x)|$$

Compare empirical CDF to theoretical one

Tabulated p-values, *assuming* the theoretical CDF isn't estimated

Goodness-of-Fit

How can we tell if it's a good fit or not, if we can't use R^2?

You shouldn't use R^2 that way for a regression

Use a goodness-of-fit test!

Kolmogorov-Smirnov statistic is nice: for CDFs P, Q

$$d_{KS}(P,Q) = \max_x |P(x) - Q(x)|$$

Compare empirical CDF to theoretical one

Tabulated p-values, *assuming* the theoretical CDF isn't estimated

Analytic corrections via heroic probability theory [11, pp. 99ff]

Goodness-of-Fit

How can we tell if it's a good fit or not, if we can't use R^2?

You shouldn't use R^2 that way for a regression

Use a goodness-of-fit test!

Kolmogorov-Smirnov statistic is nice: for CDFs P, Q

$$d_{KS}(P,Q) = \max_x |P(x) - Q(x)|$$

Compare empirical CDF to theoretical one

Tabulated p-values, *assuming* the theoretical CDF isn't estimated

Analytic corrections via heroic probability theory [11, pp. 99ff]

or, use the bootstrap, like a civilized person

Given: n data points $x_{1:n}$

1. Estimate α and $x_{\min}$; $n_{\mathrm{tail}} = \#$ of data points $\geq x_{\min}$
2. Calculate d_{KS} for data and best-fit power law $= d^*$
3. Draw n random values $b_1, \ldots b_n$ as follows:
    1. with probability n_{tail}/n, draw from power-law
    2. otherwise, pick one of the $x_i < x_{\min}$ uniformly
4. Find $\widehat{\alpha}$, $\widehat{x_{\min}}$, d_{KS} for $b_{1:n}$
5. Repeat many times to get distribution of d_{KS} values
6. p-value $=$ fraction of simulations where $d \geq d^*$

For the blogs: $p = 6.6 \times 10^{-2}$

Testing Against Alternatives

Compare against alternatives: more statistical power, more substantive information

Testing Against Alternatives

Compare against alternatives: more statistical power, more substantive information

$*$IC is sub-optimal here

Testing Against Alternatives

Compare against alternatives: more statistical power, more substantive information
$*$IC is sub-optimal here
Better: Vuong's normalized log-likelihood-ratio test [12]

Testing Against Alternatives

Compare against alternatives: more statistical power, more substantive information
$*$IC is sub-optimal here
Better: Vuong's normalized log-likelihood-ratio test [12]
Two models, θ, ψ

$$\mathcal{R}(\psi, \theta) = \log p_{\psi}(x_{1:n}) - \log p_{\theta}(x_{1:n})$$

$\mathcal{R}(\psi, \theta) > 0$ means: the data were more likely under ψ than under θ
How much more likely do they need to be?

Distribution of Likelihood Ratios: Fixed Models

Assume $X_1, X_2, \ldots$ all IID, with true distribution ν
Fix θ and ψ; what is distribution of $n^{-1}\mathcal{R}(\psi, \theta)$?

Distribution of Likelihood Ratios: Fixed Models

Assume $X_1, X_2, \ldots$ all IID, with true distribution ν
Fix θ and ψ; what is distribution of $n^{-1}\mathcal{R}(\psi,\theta)$?

$$\begin{aligned} n^{-1}\mathcal{R}(\psi,\theta) &= \frac{\log p_{\psi}(x_{1:n}) - \log p_{\theta}(x_{1:n})}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\psi}(x_i)}{p_{\theta}(x_i)} \end{aligned}$$

Distribution of Likelihood Ratios: Fixed Models

Assume $X_1, X_2, \ldots$ all IID, with true distribution ν
Fix θ and ψ; what is distribution of $n^{-1}\mathcal{R}(\psi,\theta)$?

$$\begin{aligned} n^{-1}\mathcal{R}(\psi,\theta) &= \frac{\log p_{\psi}(x_{1:n}) - \log p_{\theta}(x_{1:n})}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{p_{\psi}(x_i)}{p_{\theta}(x_i)} \end{aligned}$$

mean of IID terms so use law of large numbers:

$$\frac{1}{n}\mathcal{R}(\psi,\theta) \to \mathbf{E}_{\nu}\left[\log\frac{p_{\psi}(X)}{p_{\theta}(X)}\right] = D(\nu\|\theta) - D(\nu\|\psi)$$

Distribution of Likelihood Ratios: Fixed Models

Assume $X_1, X_2, \ldots$ all IID, with true distribution ν
Fix θ and ψ; what is distribution of $n^{-1}\mathcal{R}(\psi,\theta)$?

$$n^{-1}\mathcal{R}(\psi,\theta) = \frac{\log p_\psi(x_{1:n}) - \log p_\theta(x_{1:n})}{n}$$
$$= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\psi(x_i)}{p_\theta(x_i)}$$

mean of IID terms so use law of large numbers:

$$\frac{1}{n}\mathcal{R}(\psi,\theta) \rightarrow \mathbf{E}_\nu\left[\log \frac{p_\psi(X)}{p_\theta(X)}\right] = D(\nu\|\theta) - D(\nu\|\psi)$$

$\mathcal{R}(\psi,\theta) > 0 \approx \psi$ diverges less from ν than θ does

Use CLT:

$$\frac{1}{\sqrt{n}}\mathcal{R}(\psi,\theta) \rightsquigarrow \mathcal{N}(\sqrt{n}(D(\nu\|\theta) - D(\nu\|\psi)), \omega^2_{\psi,\theta})$$

where

$$\omega^2_{\psi,\theta} = \mathrm{Var}\left[\log\frac{p_\psi(X)}{p_\theta(X)}\right]$$

so if the models are equally good, we get a mean-zero Gaussian but if one is better $\mathcal{R}(\psi,\theta) \rightarrow \pm\infty$, depending

Distribution of $\mathcal{R}$ with Estimated Models

two classes of models Ψ, Θ; $\hat{\psi}, \hat{\theta}$ = ML *estimated* models

$\hat{\psi} \rightarrow \psi^*$, $\hat{\theta} \rightarrow \theta^*$: converging to **pseudo-truth**; $\psi^* \neq \theta^*$

some regularity assumptions

Distribution of $\mathcal{R}$ with Estimated Models

two classes of models Ψ, Θ; $\hat{\psi}, \hat{\theta}$ = ML *estimated* models
$\hat{\psi} \rightarrow \psi^*$, $\hat{\theta} \rightarrow \theta^*$: converging to **pseudo-truth**; $\psi^* \neq \theta^*$
some regularity assumptions
Everything works out as if no estimation:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\mathcal{R}(\hat{\psi}, \hat{\theta}) &\rightsquigarrow \mathcal{N}(\sqrt{n}(D(\nu\|\theta^*) - D(\nu\|\psi^*)), \omega^2_{\psi^*,\theta^*}) \\
\frac{1}{n}\mathcal{R}(\hat{\psi}, \hat{\theta}) &\rightarrow D(\nu\|\theta^*) - D(\nu\|\psi^*) \\
\widehat{\omega}^2 \equiv \mathrm{Var}_{\mathrm{sample}}\left[\log\frac{p_\psi(X)}{p_\theta(X)}\right] &\rightarrow \omega^2_{\psi^*,\theta^*}
\end{aligned}$$

Vuong's Test for Non-Nested Model Classes

Assume all conditions from before

Vuong's Test for Non-Nested Model Classes

Assume all conditions from before

If the two models are really equally close to the truth,

$$\frac{\mathcal{R}}{\sqrt{n\widehat{\omega}^2}} \rightsquigarrow \mathcal{N}(0,1)$$

but if one is better, normalized log likelihood ratio goes to $\pm\infty$, telling you which is better

Vuong's Test for Non-Nested Model Classes

Assume all conditions from before
If the two models are really equally close to the truth,

$$\frac{\mathcal{R}}{\sqrt{n\widehat{\omega}^2}} \rightsquigarrow \mathcal{N}(0,1)$$

but if one is better, normalized log likelihood ratio goes to $\pm\infty$, telling you which is better

- Don't need to adjust for parameter #, but any $o(n)$ adjustment is fine; [13] is probably better than $*$IC
- Does *not* assume that truth is in either Ψ or Θ
- *Does* assume $\psi^* \neq \theta^*$

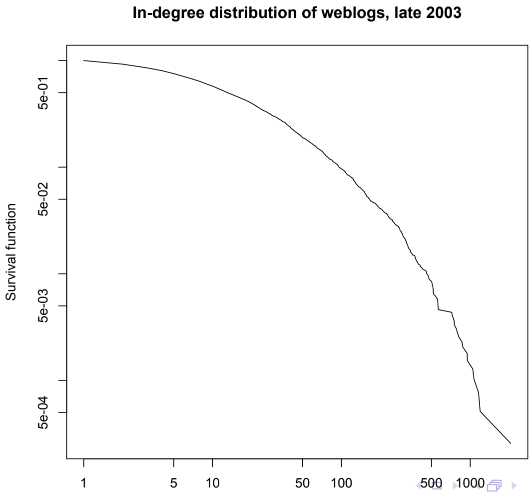
Back to Blogs

Fit a log-normal to the same tail (to give the advantage to power law)

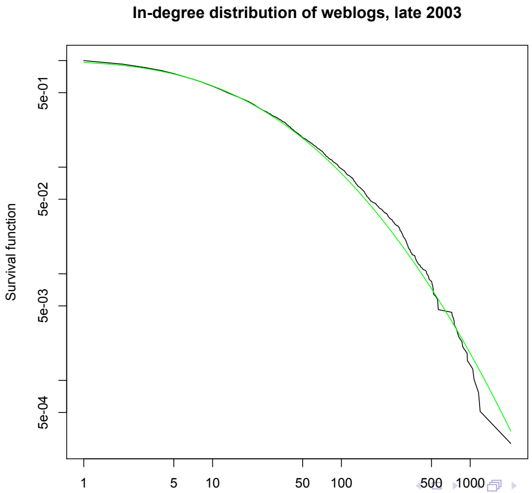
$$\begin{aligned}
\mathcal{R}(\mathrm{power\ law}, \mathrm{log}-\mathrm{normal}) &= -0.85 \\
\widehat{\omega} &= 0.098 \\
\frac{\mathcal{R}}{\sqrt{n\widehat{\omega}^2}} &= -0.83
\end{aligned}$$

so the log-normal fits better, but not by much — we'd see fluctuations at least that big 41% of the time if they were equally good

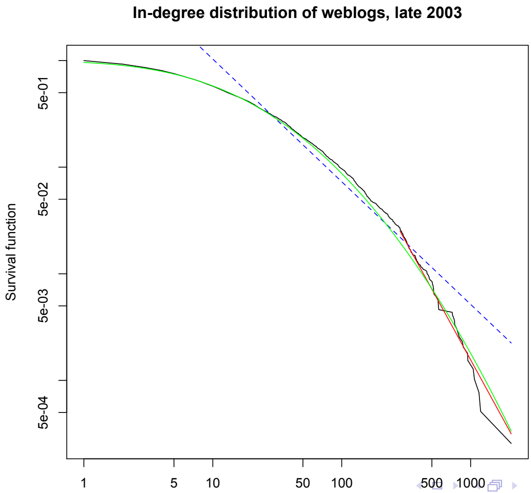
Fitting a log-normal to the complete data



Fitting a log-normal to the complete data



Fitting a log-normal to the complete data



Visualization

Beyond the log-log plot: Handcock and Morris's relative distribution [14, 15]
Compare two whole distributions, not just mean/variance etc.

Visualization

Beyond the log-log plot: Handcock and Morris's relative distribution [14, 15]

Compare two whole distributions, not just mean/variance etc.

Have a **reference distribution**, CDF F_0 (or just a **reference sample**) and a **comparison sample** $y_1, \ldots y_n$

Construct **relative data**

$$r_i = F_0(y_i)$$

relative CDF:

$$G(r) = F(F_0^{-1}(r))$$

relative density

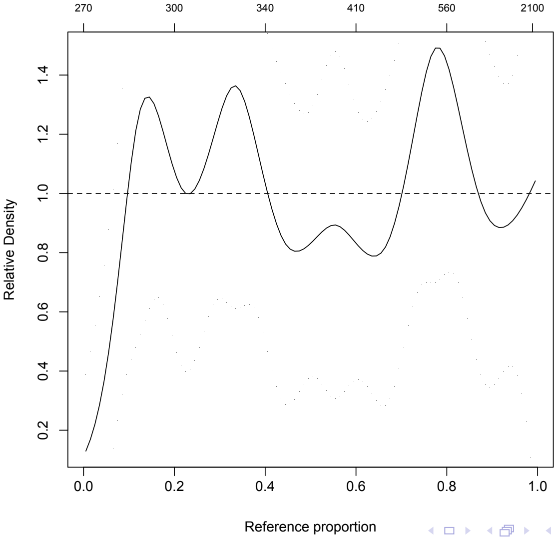
$$g(r) = \frac{f(F_0^{-1}(r))}{f_0(F_0^{-1}(r))}$$

- Relative data are uniform $\Leftrightarrow$ distributions are the same
- $g(r)$ tells us *where* and *how* the distributions differ
- Can estimate $G(r)$ by empirical CDF of r_i
- Can estimate $g(r)$ by non-parametric density estimation on r_i
- Invariant under any monotone transformation of the data (multiplication, taking logs, etc.)
- Related to Neyman's smooth test of goodness-of-fit
- Can adjust for covariates flexibly [15]

R package: `reldist`, from CRAN

Relative Distribution with Power Laws

1. Estimate power law distribution from data
2. Use that as the reference distribution


270
300
340
410
560
2100
Relative Density
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.2
0.4
0.6
0.8
1.0
Reference proportion

How Bad Is the Literature?

[10] looked at 24 claimed power laws

How Bad Is the Literature?

[10] looked at 24 claimed power laws

> *word frequency, protein interaction degree (yeast), metabolic network degree (*E. coli*), Internet autonomous system network, calls received, intensity of wars, terrorist attack fatalities, bytes per HTTP request, species per genus, # sightings per bird species, population affected by blackouts, sales of best-sellers, population of US cities, area of wildfires, solar flare intensity, earthquake magnitude, religious sect size, surname frequency, individual net worth, citation counts, # papers authored, # hits per URL, in-degree per URL, # entries in e-mail address books*

How Bad Is the Literature?

[10] looked at 24 claimed power laws

> *word frequency, protein interaction degree (yeast), metabolic network degree (*E. coli*), Internet autonomous system network, calls received, intensity of wars, terrorist attack fatalities, bytes per HTTP request, species per genus, # sightings per bird species, population affected by blackouts, sales of best-sellers, population of US cities, area of wildfires, solar flare intensity, earthquake magnitude, religious sect size, surname frequency, individual net worth, citation counts, # papers authored, # hits per URL, in-degree per URL, # entries in e-mail address books*

Of these, the *only* clear power law is word frequency

How Bad Is the Literature?

[10] looked at 24 claimed power laws

> *word frequency, protein interaction degree (yeast), metabolic network degree (*E. coli*), Internet autonomous system network, calls received, intensity of wars, terrorist attack fatalities, bytes per HTTP request, species per genus, # sightings per bird species, population affected by blackouts, sales of best-sellers, population of US cities, area of wildfires, solar flare intensity, earthquake magnitude, religious sect size, surname frequency, individual net worth, citation counts, # papers authored, # hits per URL, in-degree per URL, # entries in e-mail address books*

Of these, the *only* clear power law is word frequency
The rest: indistinguishable from log-normal and/or stretched exponential; and/or cut-off significantly better than pure power law; and/or goodness-of-fit is just horrible

What's Bad About Hallucinating Power Laws?

Scientists should not try to explain things which don't happen

What's Bad About Hallucinating Power Laws?

Scientists should not try to explain things which don't happen

e.g., years of theorizing why biochemical networks are scale-free [16, 17, 18], when they aren't [19, 20]

What's Bad About Hallucinating Power Laws?

Scientists should not try to explain things which don't happen

e.g., years of theorizing why biochemical networks are scale-free [16, 17, 18], when they aren't [19, 20]

Decision-makers waste resources planning for power laws which don't exist

Does It Really Matter Whether It's a *Power Law*?

Maybe all that matters is that the distribution has a heavy tail

Probably true for Shirky

Does It Really Matter Whether It's a *Power Law*?

Maybe all that matters is that the distribution has a heavy tail

Probably true for Shirky

Then *don't* say that it's a power law

Does It Really Matter Whether It's a *Power Law*?

Maybe all that matters is that the distribution has a heavy tail

Probably true for Shirky

Then *don't* say that it's a power law

Do look at density estimation methods for heavy-tailed distributions [21, 22]

- Data-independent transformation from $[0, \infty)$ to $[0, 1]$
- Nonparametric density estimate on $[0, 1]$
- Inverse transform

The Correct Line

1. Lots of distributions give straightish log-log plots
2. Regression on log-log plots is bad; don't do it, and don't believe those who do it.
3. Use maximum likelihood to estimate the scaling exponent
4. Use goodness of fit to estimate the scaling region
5. Use goodness of fit tests to check goodness of fit
6. Use Vuong's test to check alternatives
7. Ask yourself whether you really care

[1] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955. URL http://www.jstor.org/pss/2333389.

[2] Yuji Ijiri and Herbert A. Simon. *Skew Distributions and the Sizes of Business Firms*. North-Holland, Amsterdam, 1977. With Charles P. Bonini and Theodore A. van Wormer.

[3] William J. Reed and Barry D. Hughes. From gene families and genera to incomes and Internet file sizes: Why power laws are so common in nature. *Physical Review E*, 66:067103, 2002. doi: 10.1103/PhysRevE.66.067103.

[4] B. A. Maguire, E. S. Pearson, and A. H. A. Wynn. The time intervals between industrial accidents. *Biometrika*, 39: 168–180, 1952. URL http://www.jstor.org/pss/2332475.

[5] Clay Shirky. Power laws, weblogs, and inequality. In Mitch Ratcliffe and Jon Lebkowsky, editors, *Extreme Democracy*,

forthcoming. URL http://www.shirky.com.writings/powerlaw_weblog.html.

[6] Henry Farrell and Daniel Drezner. The power and politics of blogs. *Public Choice*, 134:15–30, 2008. URL http://www.utsc.utoronto.ca/~farrell/blogpaperfinal.pdf.

[7] A. N. M. Muniruzzaman. On measures of location and dispersion and tests of hypotheses in a Pareto population. *Bulletin of the Calcutta Statistical Association*, 7:115–123, 1957.

[8] H. L. Seal. The maximum likelihood fitting of the discrete Pareto law. *Journal of the Institute of Actuaries*, 78:115–121, 1952. URL http://www.actuaries.org.uk/files/pdf/library/JIA-078/0115-0121.pdf.

[9] B. M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3:1163–1174, 1975.

URL
http://projecteuclid.org/euclid.aos/1176343247.

[10] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51: 661–703, 2009. URL http://arxiv.org/abs/0706.1062.

[11] David Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984. URL http://www.stat.yale.edu/~pollard/1984book/.

[12] Quang H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989. URL http://www.jstor.org/pss/1912557.

[13] Hwan-sik Choi and Nicholas M. Kiefer. Differential geometry and bias correction in nonnested hypothesis testing. Online preprint, 2006. URL http://www.arts.cornell.edu/econ/kiefer/GeometryMS6.pdf.

[14] Mark S. Handcock and Martina Morris. Relative distribution methods. *Sociological Methodology*, 28:53–97, 1998. URL `http://www.jstor.org/pss/270964`.

[15] Mark S. Handcock and Martina Morris. *Relative Distribution Methods in the Social Sciences*. Springer-Verlag, Berlin, 1999.

[16] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999. URL `http://arxiv.org/abs/cond-mat/9910332`.

[17] Ricard V. Solé, Romualdo Pastor-Satorras, Eric Smith, and Thomas B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5:43–54, 2002. URL `http://arxiv.org/abs/cond-mat/0207311`.

[18] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *Complexus*, 1:38–44, 2003. URL `http://arxiv.org/abs/cond-mat/0108043`.

[19] Raya Khanin and Ernst Wit. How scale-free are biological networks? *Journal of Computational Biology*, 13:810–818, 2006. doi: 10.1089/cmb.2006.13.810. URL http://iwi.eldoc.ub.rug.nl/root/2006/JCompBiolKhanin/.

[20] Adrían López García de Lomana, Qasim K. Beg, G. de Fabritiis, and Jordi Villà-Freixa. Statistical analysis of global connectivity and activity distributions in cellular networks. *Journal of Computational Biology*, 17:869–878, 2010. doi: 10.1089/cmb.2008.0240. URL http://arxiv.org/abs/1004.3138.

[21] Natalia M. Markovitch and Udo R. Krieger. Nonparametric estimation of long-tailed density functions and its application to the analysis of World Wide Web traffic. *Performance Evaluation*, 42:205–222, 2000. doi: 10.1016/S0166-5316(00)00031-6.

[22] Natalia Markovich. *Nonparametric Analysis of Univariate*

Heavy-Tailed Data: Research and Practice. John Wiley, New York, 2007.