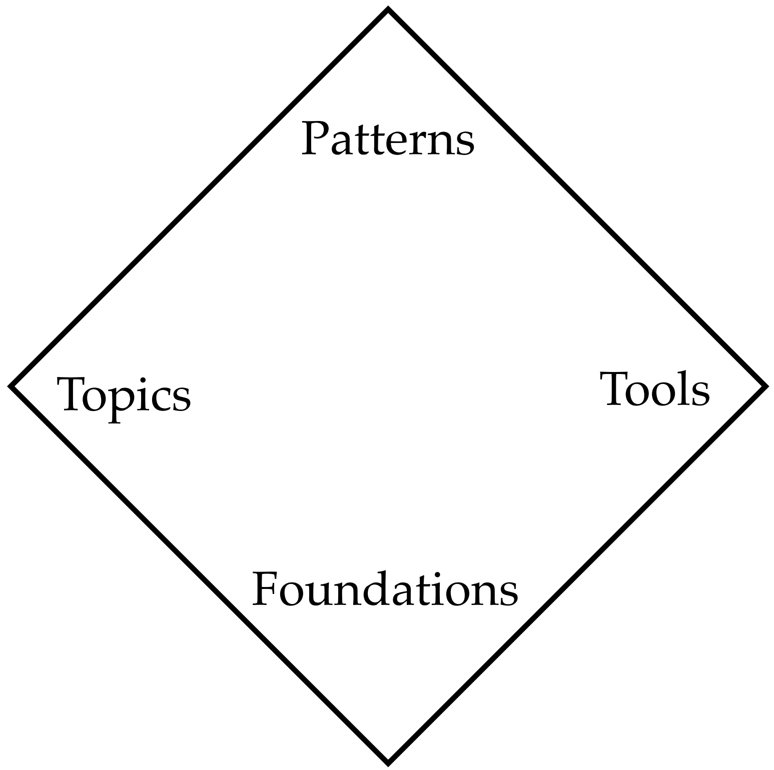
Statistical Analysis of Complex Systems

Cosma Shalizi
Statistics Department, Carnegie Mellon University

Complex Systems Summer School 2006, Beijing

exploitation vs. exploration
increasing returns→skew distributions
stability through hierarchy
local inhibition + long-range activation
etc.

pattern formation
spin glasses
turbulence
social insects
machine learning
optimization
neural networks (real, fake)
evolution (real, fake)
immune system
networks
gene regulation
evolutionary economics
social dynamics
sandpiles
etc.



nonlinear dynamics
time-series analysis
cellular automata
stochastic spatial models
information theory
evolutionary game theory
agent-based models
renormalization group
thermodynamic formalism
heavy-tailed stochastics
(data-mining / statistical learning)
etc.

measures of complexity
organization: represent, detect, use
intrinsic computation, semantics
physics of information
does complexity increase?

After a triangle drawn
by Prof. Michael Cohen
(School of Information,
University of Michigan)

1. Power laws (today)
    - what they are
    - what they mean; what they don't mean
    - how to tell if you have one
2. Statistics for complex systems (Wednesday)
    - Fitting models
    - Inductive complexity
    - Comparison of alternatives
3. Model discovery (Thursday)

Power Laws

What they are
Why they mean
What they don’t mean
How to tell if you have one

What Are They?

Three Kinds of Things Called "Power Laws"

1. Physical laws, like Newton's law of gravitation: $F \propto r^{-2}$
2. Scaling relations, like the "three-quarters" law of biology: $P \propto m^{-3/4}$
3. Statistical distributions, like Zipf's law of city sizes: $\Pr(X \geq x) \propto x^{-1.3}$

We are only going to look at distributions

Power Law (Pareto) Distributions

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

cumulative distribution $\Pr(X \geq x) = \frac{1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-(\alpha-1)}$

parameters

threshold or *minimum* $x_{\min}$

slope or *exponent* α

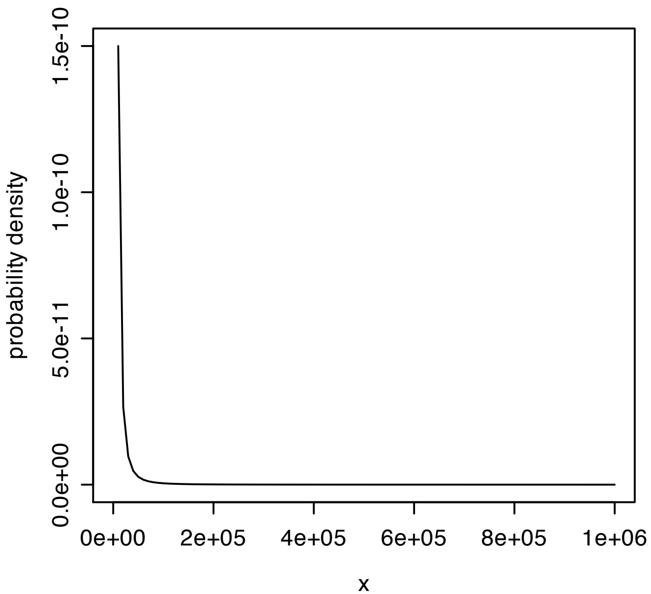
strongly **heavy-tailed** or **right-skewed**

median is much higher than mean

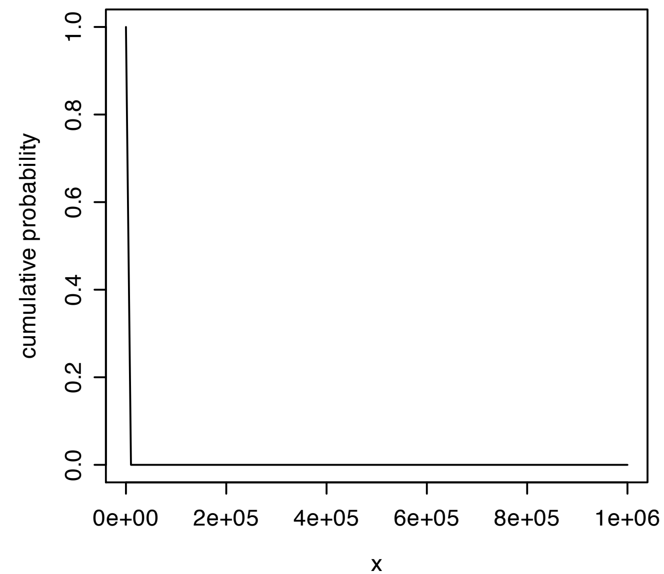
very large fluctuations from mean

"80/20" rule

Plots

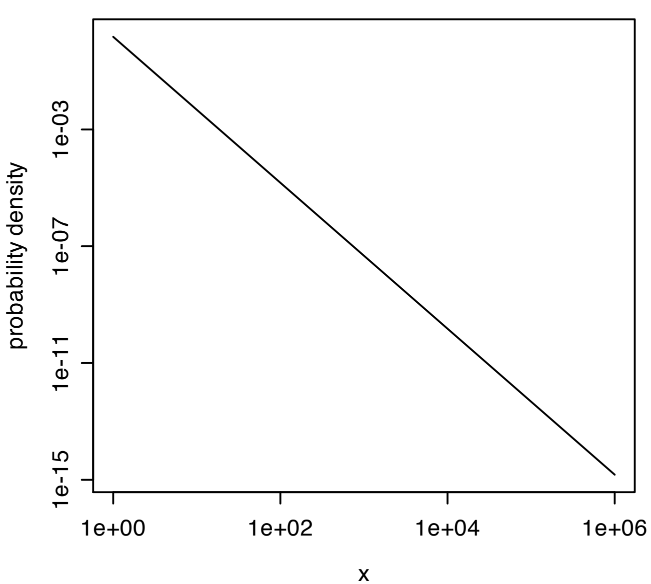


linear

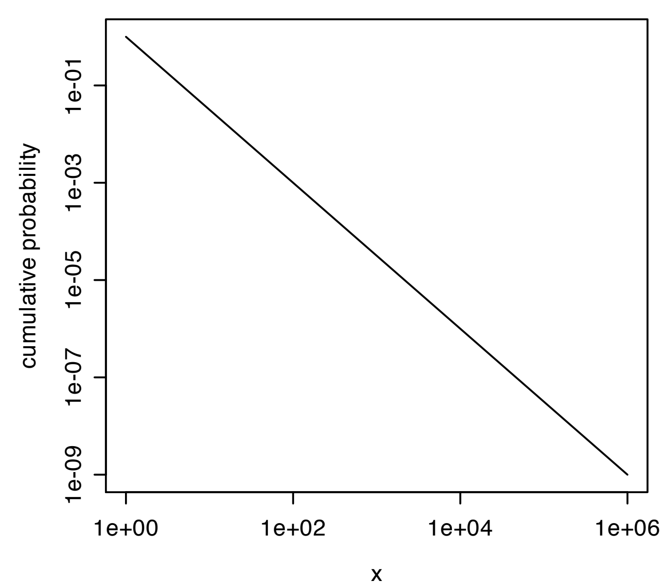


Density

Cumulative Probability



log-log



What Do They Mean?

Why they matter for reality

Many quantities have power law distributions...

- Income and wealth (Pareto; apparently oldest use)
- City sizes (Zipf)
- Word frequencies
- Paper citations
- Earthquake amplitude
- Solar flares
- Family names
- Sizes of genera

... so Gaussian/normal assumptions are badly wrong

... and models need to match reality

Why They Matter: Theory

Different mechanisms produce different distributions

- So: use distribution to learn about mechanism

Averaging many independent variables $\Rightarrow$ Gaussian

- **central limit theorem**
- also true if only approximately independent ("mixing")

Self-reinforcing growth + random seeding $\Rightarrow$ power law

- Simon, 1955 (Carnegie Mellon)
- Barabasi and Albert, 1999

Exponential growth for a random (exponential) time

- Reed and Hughes, 2002

Many other power-law mechanisms!

What Don’t They Mean?

Why do physicists care?

In equilibrium statistical mechanics, Gaussian fluctuations

- Observables average over many, many molecules
- Molecules become independent very quickly (mixing)
- Einstein fluctuation formula based on entropy

Except: power-law fluctuations at critical points

- Correlation length diverges, no mixing
- Usual CLT does not apply
- Power laws come out of lack of scale

So physicists think “power law $\Rightarrow$ critical $\Rightarrow$ cool”

Origin of “self-organized criticality”

BUT THIS IS WRONG, *WRONG*, ***WRONG!***

Why is this wrong?

Because there are many other ways to produce power laws!

Many of them do not involve long-range correlation, memory, or anything else we would call "complex"

Not all complex systems have power laws

Not all power law distributions indicate complexity

How Can You Tell If You Have One?

The bad way: draw a straight line

- Easy
- Common (outside of statistics and economics)
- Worthless

The right way: use maximum likelihood

- A little harder (must learn *some* statistics)
- Not so common (outside of statistics and economics)
- Works

The Bad Way

Plot the cumulative distribution
Fit a line by least squares; slope = α - 1
Use error estimates in slope from fit
Check R^2, the fraction of variance which comes from this line; accept if it is high

Advantages of the Bad Way

- It is easy
  - Excel or Gnuplot will do it for you
- It requires no knowledge
- It sounds reasonable
- Many other people do it
  - vast majority of papers in physics or econophysics on power laws do exactly this

Why then is it bad?

It does not give a proper probability distribution!

If you do have a power law: *bad estimates*

- Asymptotic convergence to true values, but very slow
- Also, the standard error for the slope is wrong

If you don't: *it can't tell the difference*

- Low **power** against heavy-tailed alternatives
- A small part of *any* curve looks like a straight line
- We will see an example

Both of problems have been known to statisticians since the 1960s at least

What then is the right way?

Estimate parameters by *maximum likelihood*
Error bars by *bootstrapping*
Check fit by *Kolmogorov-Smirnov test* (good)
Test against real alternatives (better)

Parameter estimation

Likelihood of a parameter value = probability of data, under that parameter value

Maximum likelihood estimate (MLE) = what parameter value makes the data most probable?

- Curve fitting says "what curve goes closest to my data, in Euclidean distance?"
- MLE says, "what distribution goes closest to my data, in Kullback divergence/relative entropy?"
- This idea leads to **information geometry**

More on MLE

Use log-likelihood instead of likelihood

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

$$= n \log \alpha - 1 + n(\alpha - 1) \log x_{\min} - \alpha \sum_{i=1}^{n} \log x$$

Explicit solutions

$$\widehat{x_{\min}} = x_{(1)}, \text{ smallest value}$$

$$\widehat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \log \frac{x}{x_{\min}} \right]^{-1}$$

For Pareto distribution, MLEs are **consistent**

- i.e., they converge in probability on the true values
- *much* more accurate than line-fitting

They are also **sufficient**

- i.e., they use *all* the information in the data

MLE does not give error estimates

MLE does not check the fit

MLE does not test against alternatives

Bootstrapping

Key technique of modern statistics
- Similar to "surrogate data" in nonlinear dynamics

Fit model to data

Simulate new data set from model

Apply procedure to simulated data

Repeat many times to get typical results
- *if* model is correct

this is *parametric* bootstrap, *nonparametric* is trickier

Bootstrap Error Estimates

Fit Pareto to n data points with MLE
Generate n random values from fitted Pareto
Apply MLE to simulated data, calculate error
Repeat m times to get average error

Works for any distribution, not just Pareto
 but is not always the most efficient way
 ("local asymptotic normality")

Checking the Fit

Goodness-of-fit tests: *if* the model is right, what is the probability of results *like* the data?

Or: probability of results *as far* from expectations as the data

Model here is a distribution, so we need a measure of distance between distributions

Kolmogorov-Smirnov Test

$$F(x) = \text{cumulative distribution according to model}$$
$$\widehat{F}_n(x) = \text{cumulative distribution according to data}$$
$$D = \max_x \left| F(x) - \widehat{F}_n(x) \right|$$

If the model is right, D is usually small, shrinks as n grows

null distribution of D

K&S found null distribution as function of n for *fixed* model F(x)

For model estimated from data, find distribution of D by bootstrapping

More on the KS Test

Does not require binning the data

High probability of detecting bad models

- has high **power** against most alternatives
- **non-parametric** test

Not informative if the fit is bad

- price of being non-parametric

Testing Alternatives

Pareto is only appropriate for heavy-tailed data

Compare it to other heavy-tailed distributions

- i.e., don't bother with Gaussian, exponential, etc.

There are many!

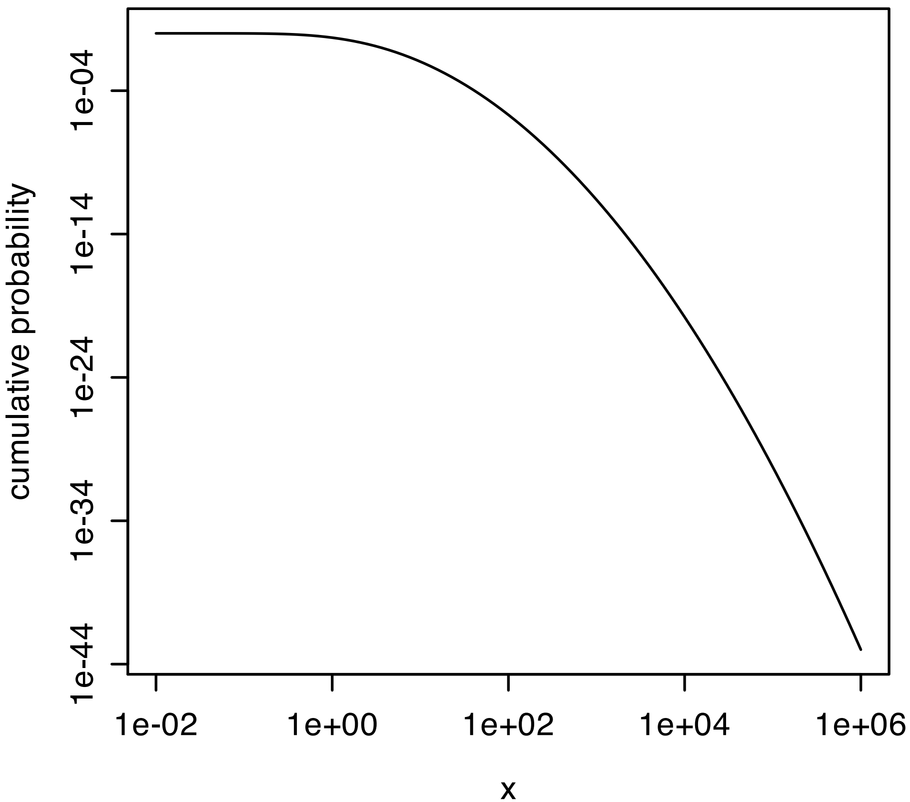
- Weibull distribution
- Stretched exponential
- lognormal
- etc.

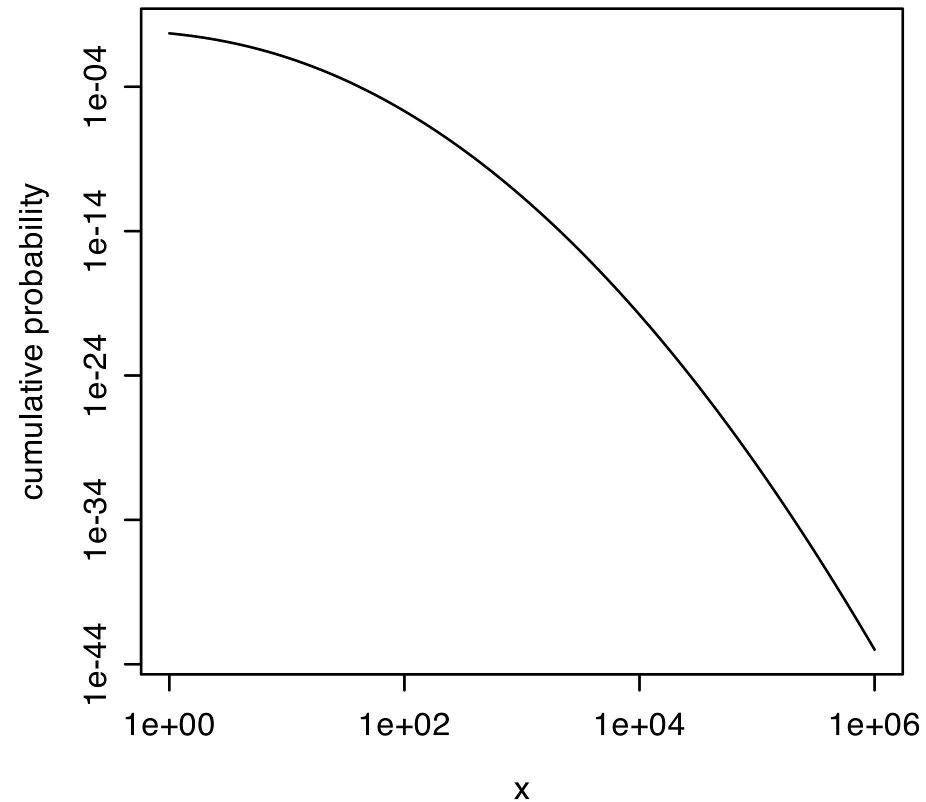
Lognormal is usually the most important

Lognormal Distribution

$$p(x) = \frac{1}{\sqrt{2\pi} s x} e^{-\frac{(\log x - m)^2}{2s^2}}$$

- log(X) has a normal/Gaussian distribution
  - Arises from multiplicative CLT (multiplying many small independent factors)
  - just like Gaussian arises from CLT (adding many small independent variables)
- Extremely common
- Parameters (*m* and *s*) easily found by MLE
- Can be hard to tell from power law





Cumulative distribution of lognormal (log-log plot)
Notice how straight the tail is

Likelihood Ratio Test

Null hypothesis: data comes from Pareto

Alternative hypothesis: some other heavy-tailed distribution, say lognormal

Fit Pareto by MLE

Fit alternative by MLE

$$
\begin{aligned}
L_{\text{null}} &= \text{likelihood of fitted power law model} \\
L_{\text{alt}} &= \text{ of fitted alternative model} \\
T &= \frac{L_{\text{alt}}}{L_{\text{null}}} \\
T \leq c &\Rightarrow \text{accept null hypothesis} \\
T > c &\Rightarrow \text{reject null hypothesis}
\end{aligned}
$$

More on LRT

- Why does this work?
  - If Pareto is right, then eventually T $\to$ 0
  - If alternative is right, then eventually T $\to \infty$
- How can it go wrong?
  - **False alarm** or **Type I** error: reject null (Pareto) when it's right
  - **Miss** or **Type II** error: accept null when it's wrong
- Usual way to pick *c*:
  - Find null distribution of T
  - Find c such that probability of false alarm is some desired small value, the "significance level"
- Alternately: $c = 1$

Severity and Evidence

Accept/reject is not enough

If we accept power law ($T < c$), still need to know

- **power** = probability, under alternative, that $T > c$
- **severity** = probability, under alternative, that $T >$ actual likelihood ratio from data

If we reject power law ($T > c$), need to know

- **significance** = probability, under null, that $T > c$
- **severity** = probability, under null, that $T <$ actual likelihood ratio

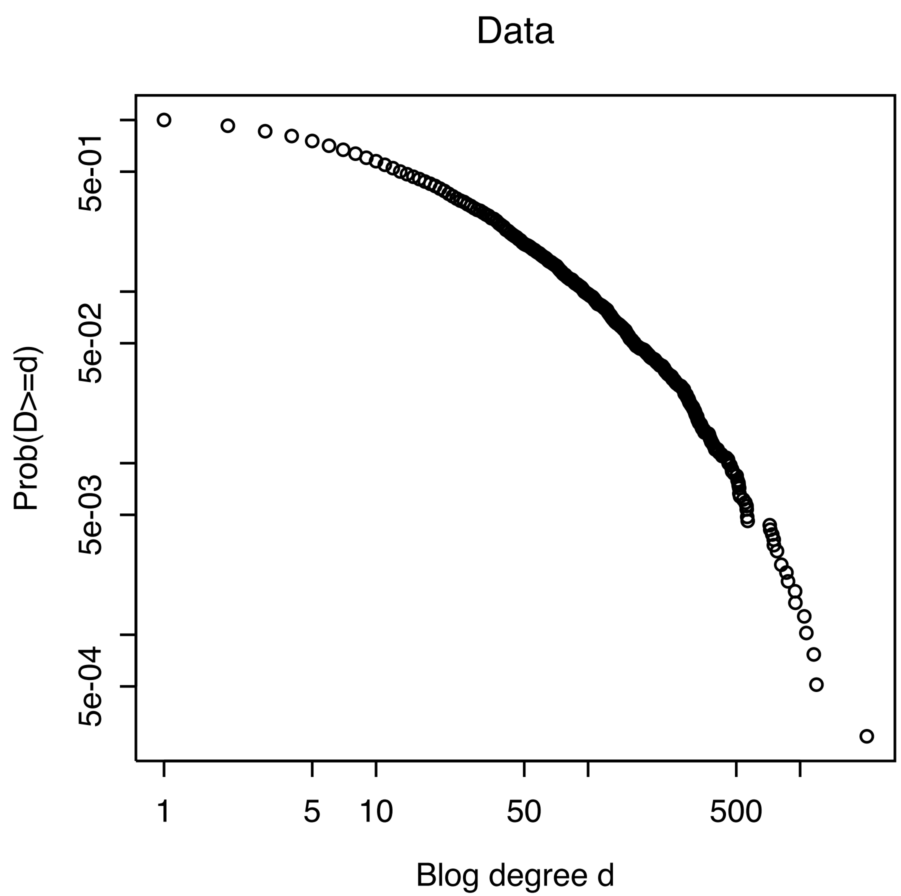
Low significance and high power are both good

High severity: the test provides strong evidence

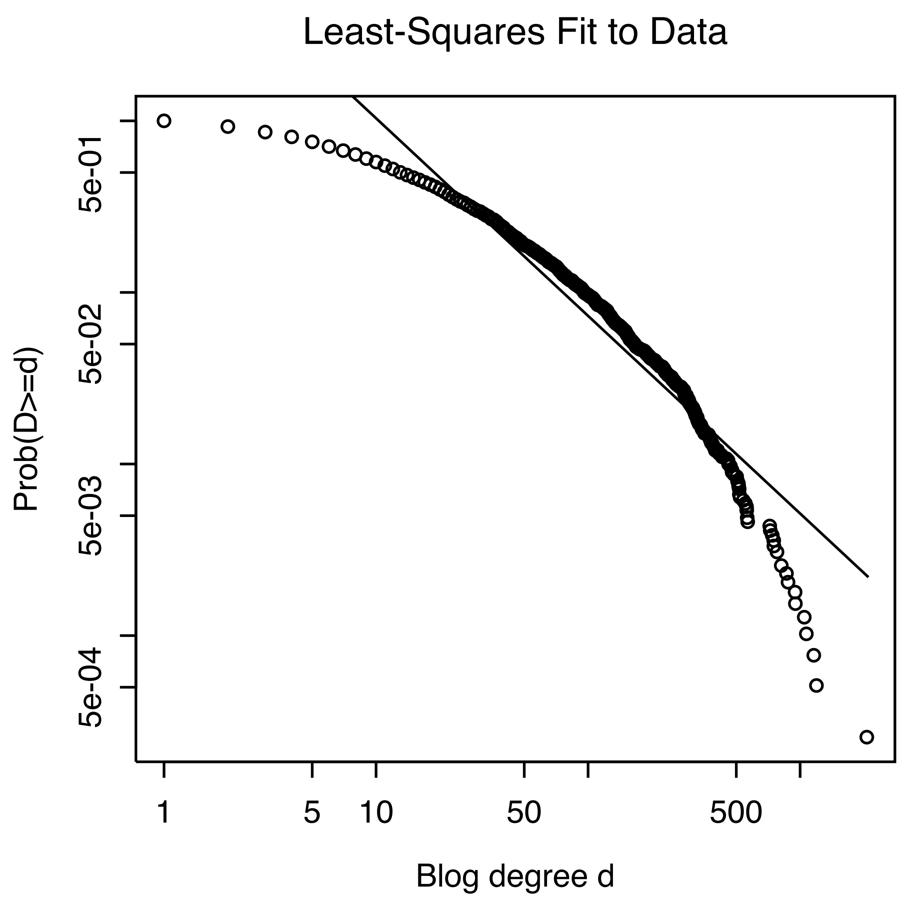
Low severity: the test provides weak evidence

How not to draw a straight line

http://bactra.org/weblog/232.html

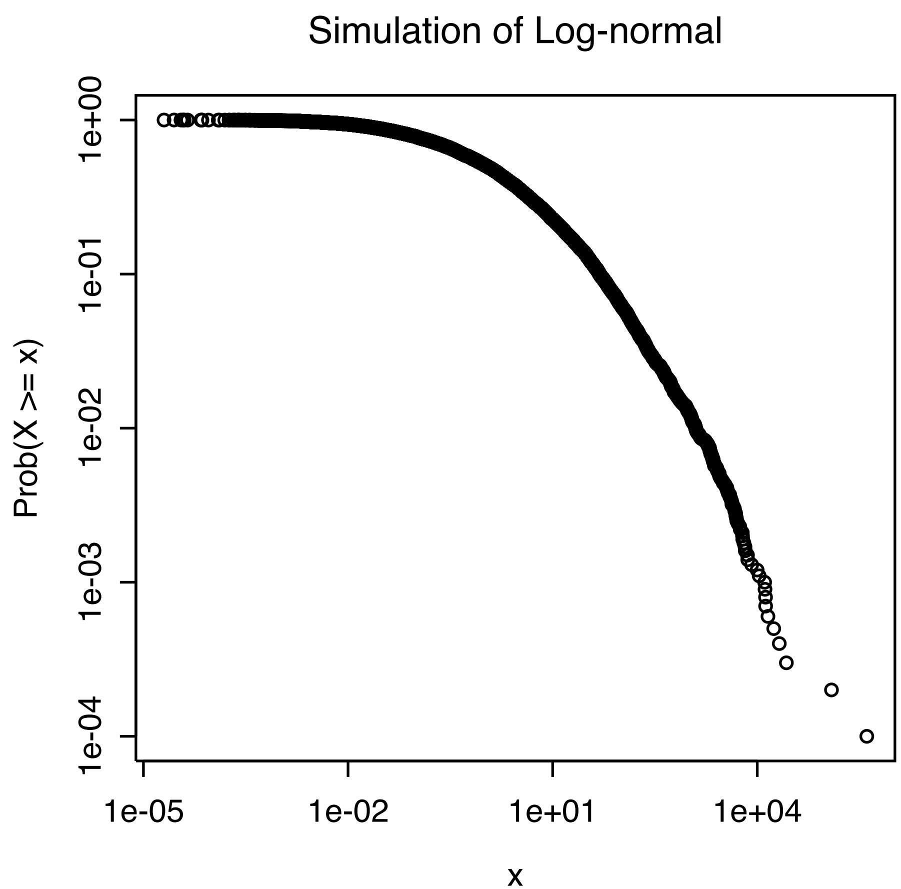


Cumulative degree distribution of English-language political blogs, 2004
(data courtesy H. Farrell & D. Drenzer)
3 orders of magnitude

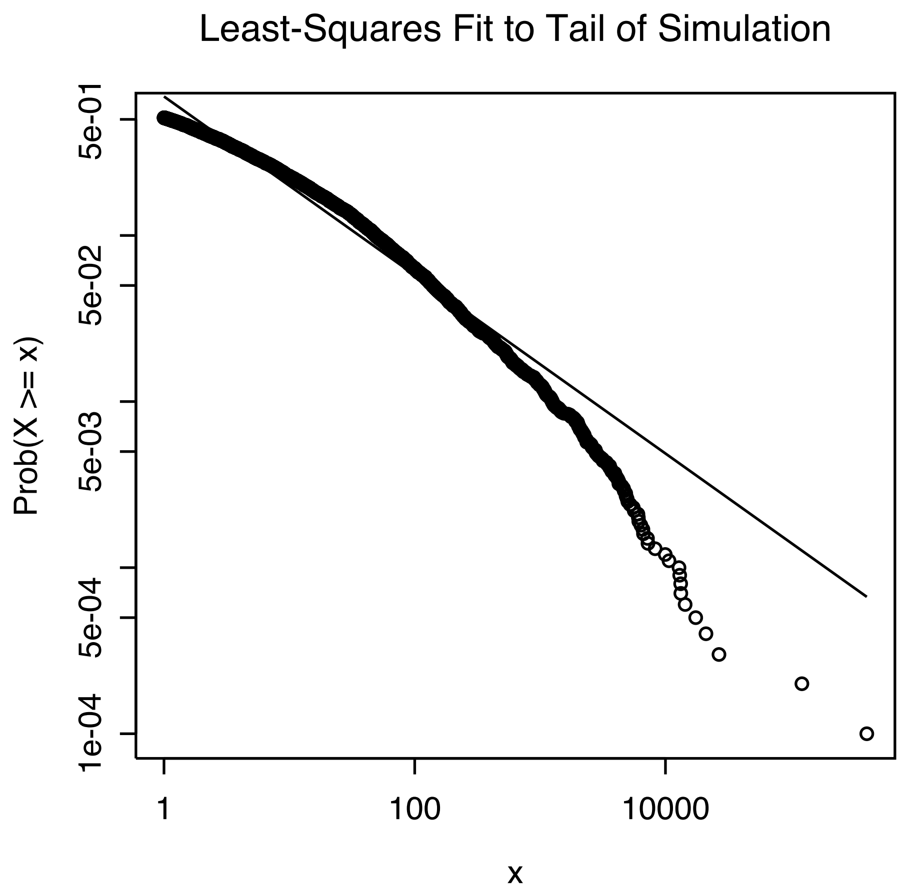


Least-squares fit

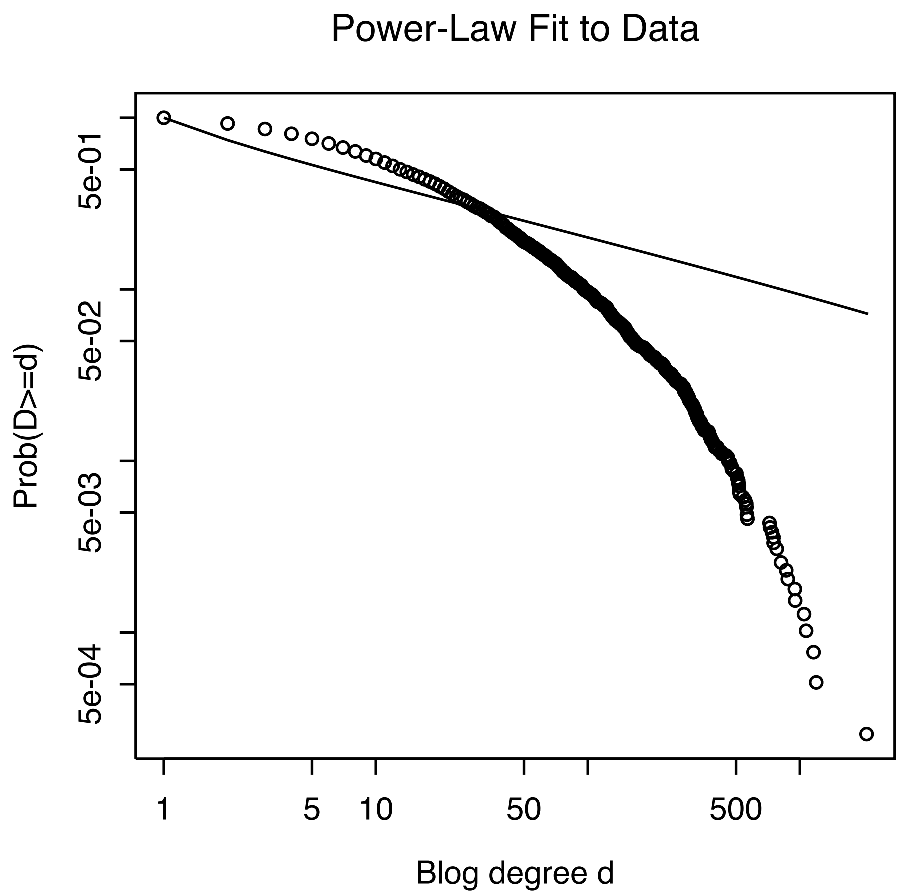
$\alpha = -2.15$, $R^2 = 0.898$



10,000 numbers from a log-normal distribution
m= 0, s= 3

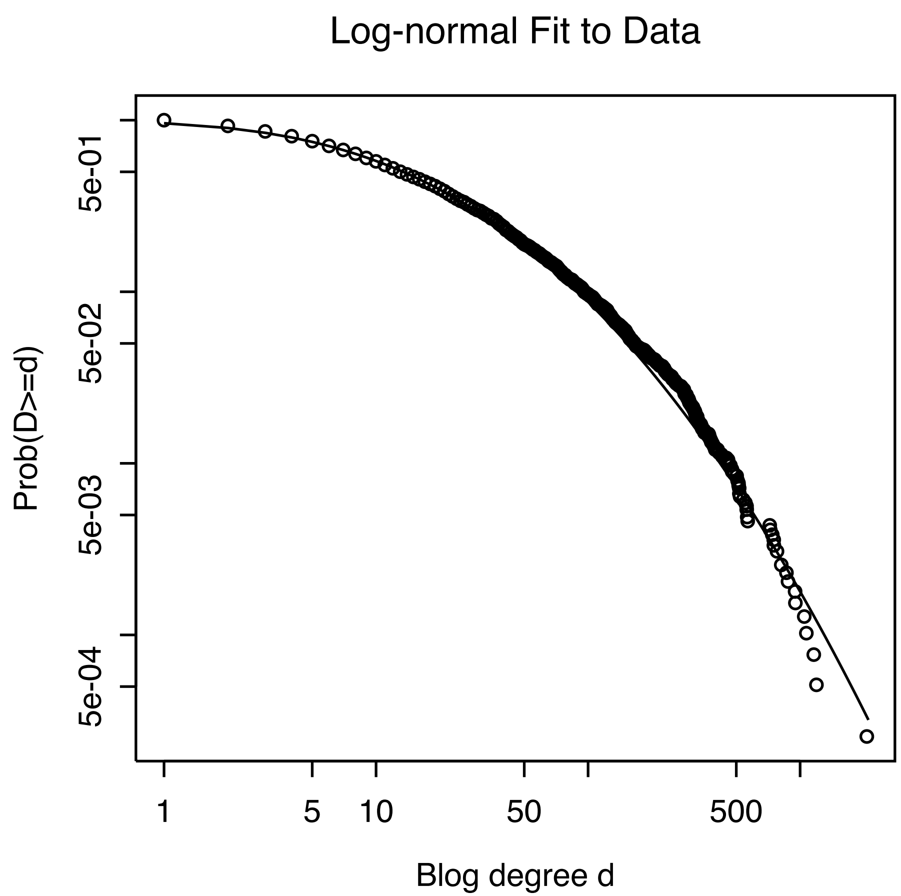


Least-squares fit to random numbers ≥ 1
5112 data points, 4+ orders of magnitude
$R^2 = 0.962$



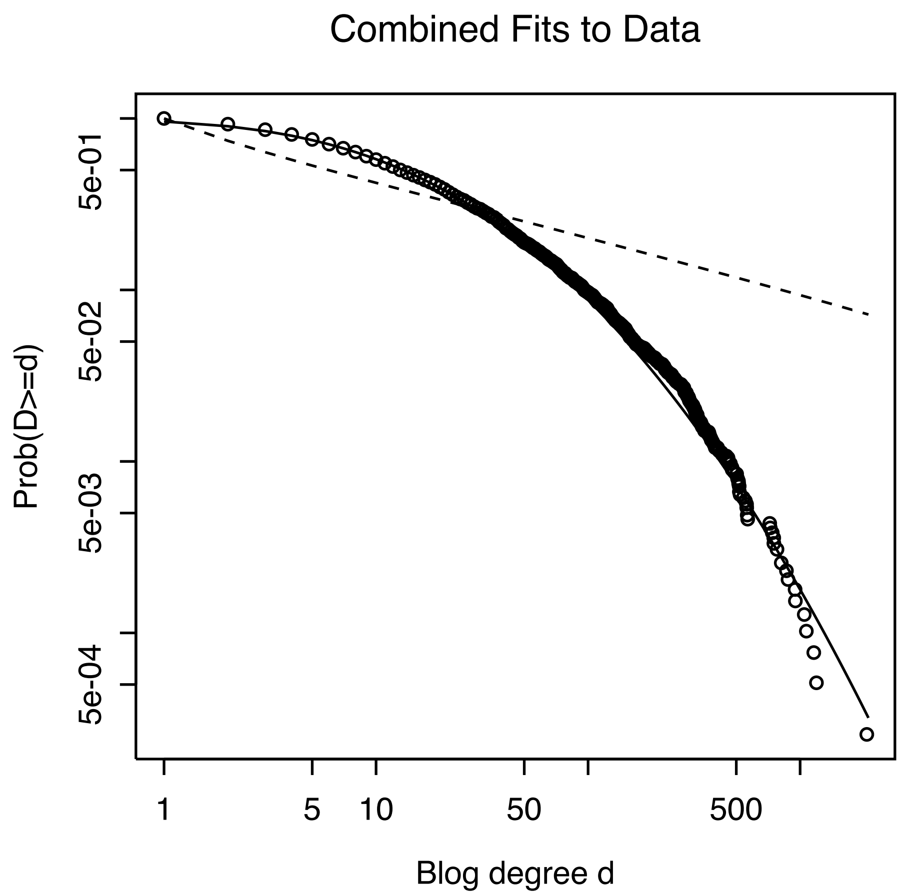
Maximum likelihood fit of power law

α = -1.30, log L = -18481.51



Maximum likelihood fit of log-normal

m = 2.60, s = 1.48, log L = -17218.22



data are

$$e^{-17218.22+18481.51} = e^{1263.29} \approx 13{,}000{,}000$$

times more likely under the log-normal

Morals

1. Power laws are an important kind of heavy-tailed distributions, often seen in complex systems
2. Do not estimate their parameters by line-fitting
3. There are other heavy-tailed distributions, so check before you say something is a power law
4. Many mechanisms can generate power laws
5. Some, but not all, of these mechanisms are complex

References

General References on Power Laws

M. E. J. Newman (2004), "Power laws, Pareto distributions and Zipf's law", http://arxiv.org/abs/cond-mat/0412004

If you read only one thing on power laws, make it this.

Manfred Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*

Fun, recreational-mathematics-level survey of many of the topics related to this school, including power law distributions

Mechanisms for Power Laws

Herbert Simon (1955), "On a Class of Skew Distribution Functions", *Biometrika* **42**: 425--440

Classic paper explaining power laws through “rich get richer” growth

Didier Sornette (1998), "Multiplicative Processes and Power Laws", *Physical Review E* **57**: 4811--4813, http://arxiv.org/abs/cond-mat/9708231

Michael Mitzenmacher (2003), "A Brief History of Generative Models for Power Law and Lognormal Distributions", *Internet Mathematics* **1**: 226--251, http://www.internetmathematics.org/volumes/1/2/pp226_251.pdf

William J. Reed and Barry D. Hughes (2002), "From Gene Families and Genera to Incomes and Internet File Sizes: Why Power Laws are so Common in Nature", *Physical Review E*, **66:** 067103

Critical Fluctuations

Joel Keizer (1987), *Statistical Thermodynamics of Nonequilibrium Processes* (Berlin: Springer-Verlag)

Excellent discussion of critical fluctuations and how they relate to non-equilibrium phenomena.

L. S. Landau and E. M. Lifshitz (1980), *Statistical Physics* (Oxford: Pergamon)

Essential reference on the Einstein fluctuation formula and critical phenomena.

Julia M. Yeomans (1992), *Statistical Mechanics of Phase Transitions* (Oxford: Clarendon Press)

Easier than the above, but less detailed.

Statistical Issues

Michel L. Goldstein, Steven A. Morris and Gary G. Yen (2004), "Fitting to the Power-Law Distribution", http://arxiv.org/abs/cond-mat/0402322

Pedestrian, but accurate, paper on goodness-of-fit testing for power laws.

Norman L. Johnson and Samuel Kotz (1970), *Continuous Univariate Distributions, Part 1* (New York: Wiley)

Standard reference work, discusses the properties of the Pareto distribution, and the pros and cons of various estimators. This volume also includes the log-normal distribution.

Deborah G. Mayo (1996), *Error and the Growth of Experimental Knowledge* (Chicago: University of Chicago Press)

Excellent book on how to really use statistical methods in scientific work. Best general discussion of evidence and severe tests.

Cosma Rohilla Shalizi and M. E. J. Newman (in prep.), “Statistical Inference with Power Law Distributions: Parameter Estimation and Comparison to Alternatives”

Manuscript in preparation; will be accompanied by code for automated hypothesis testing.