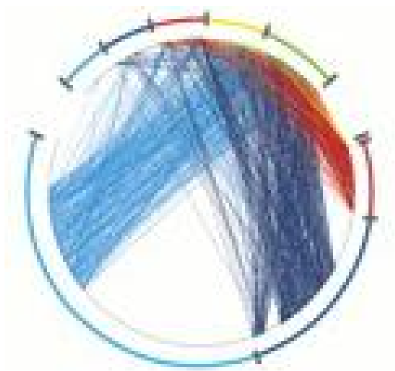
Models and Algorithms for Complex Networks

Power laws and generative processes



Power Laws - Recap

- A (continuous) random variable X follows a power-law distribution if it has density function

$$p(x) = Cx^{-\alpha}$$

- A (continuous) random variable X follows a Pareto distribution if it has cumulative function

$$P[X \geq x] = Cx^{-\beta}$$ power-law with α=1+β

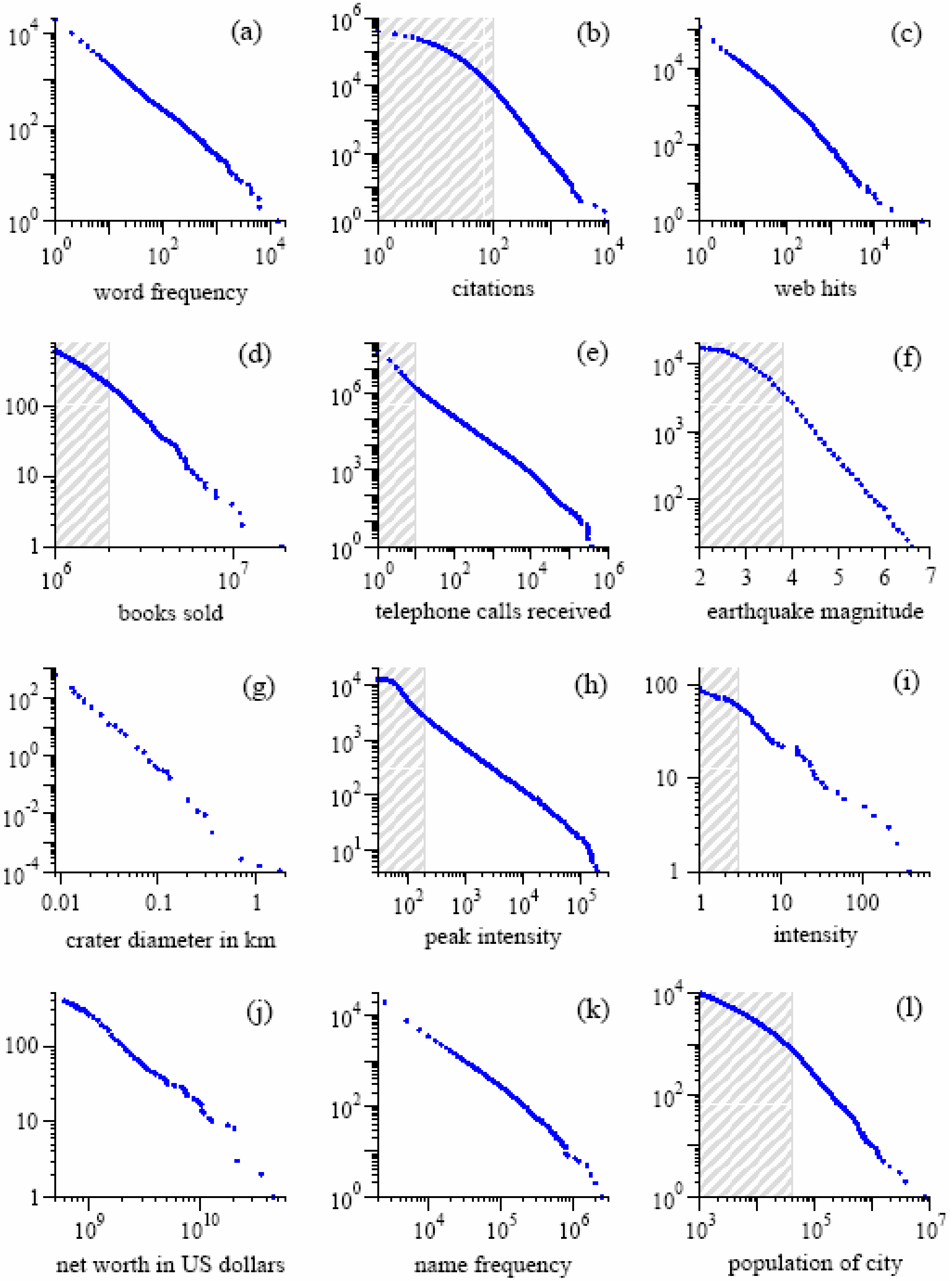
- A (discrete) random variable X follows Zipf's law if the frequency of the r-th largest value satisfies

$$p_r = Cr^{-\gamma}$$ power-law with α=1+1/γ

Power laws are ubiquitous

	quantity	minimum $x_{\min}$	exponent α
(a)	frequency of use of words	1	2.20(1)
(b)	number of citations to papers	100	3.04(2)
(c)	number of hits on web sites	1	2.40(1)
(d)	copies of books sold in the US	2 000 000	3.51(16)
(e)	telephone calls received	10	2.22(1)
(f)	magnitude of earthquakes	3.8	3.04(4)
(g)	diameter of moon craters	0.01	3.14(5)
(h)	intensity of solar flares	200	1.83(2)
(i)	intensity of wars	3	1.80(9)
(j)	net worth of Americans	\$600m	2.09(4)
(k)	frequency of family names	10 000	1.94(1)
(l)	population of US cities	40 000	2.30(5)

TABLE I Parameters for the distributions shown in Fig. 4. The labels on the left refer to the panels in the figure. Exponent values were calculated using the maximum likelihood method of Eq. (5) and Appendix B, except for the moon craters (g), for which only cumulative data were available. For this case the exponent quoted is from a simple least-squares fit and should be treated with caution. Numbers in parentheses give the standard error on the trailing figures.



But not everything is power law

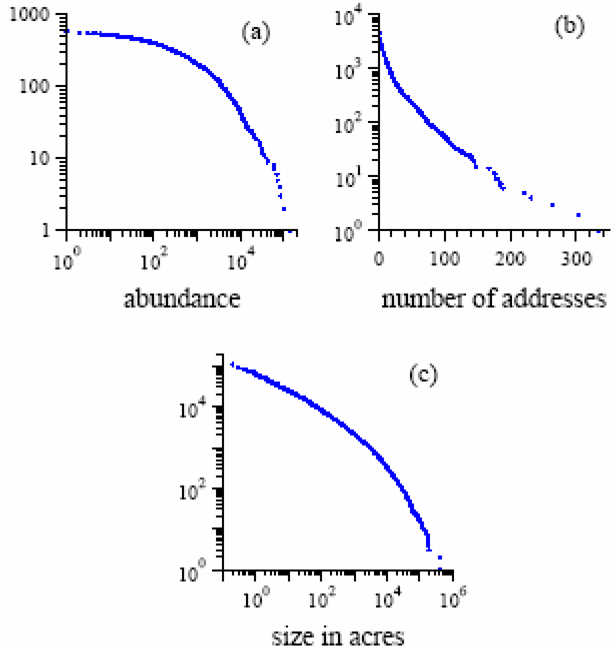
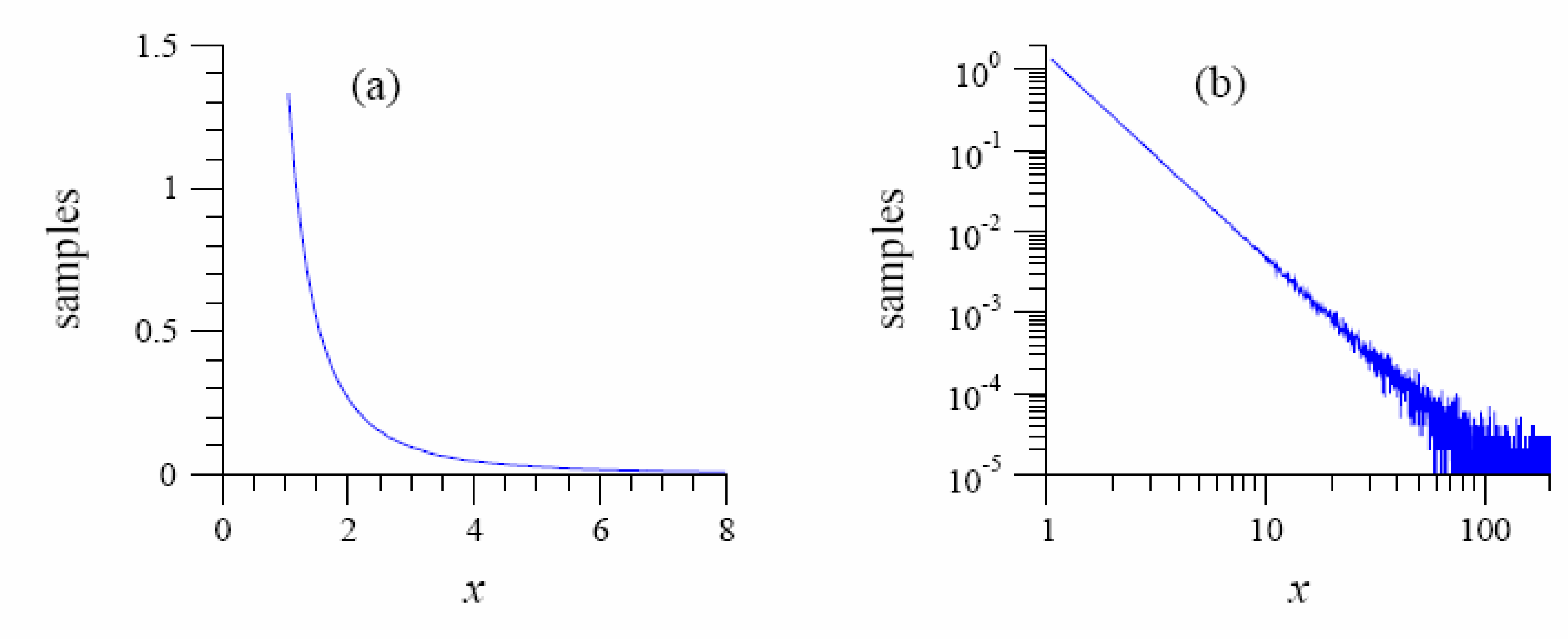


FIG. 5 Cumulative distributions of some quantities whose distributions span several orders of magnitude but that nonetheless do not follow power laws. (a) The number of sightings of 591 species of birds in the North American Breeding Bird Survey 2003. (b) The number of addresses in the email address books of 16 881 users of a large university computer system [34]. (c) The size in acres of all wildfires occurring on US federal land between 1986 and 1996 (National Fire Occurrence Database, USDA Forest Service and Department of the Interior). Note that the horizontal axis is logarithmic in frames (a) and (c) but linear in frame (b).

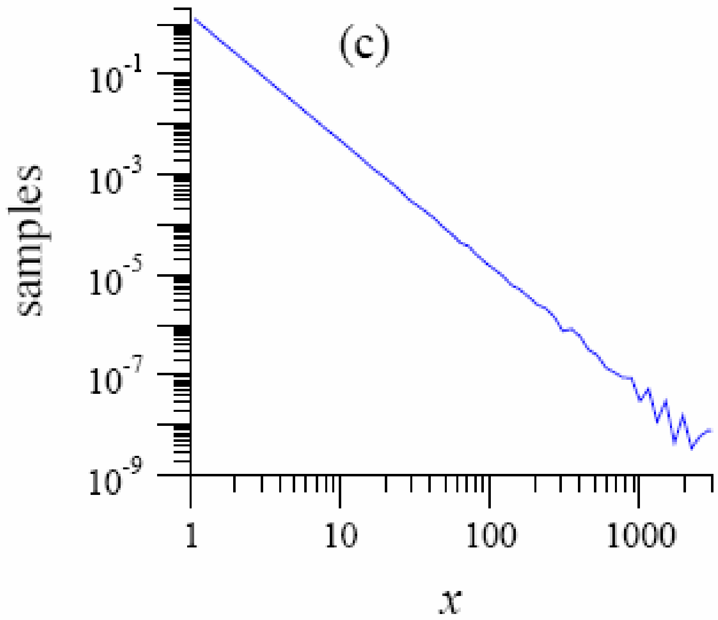
Measuring power laws

Simple log-log plot gives poor estimate



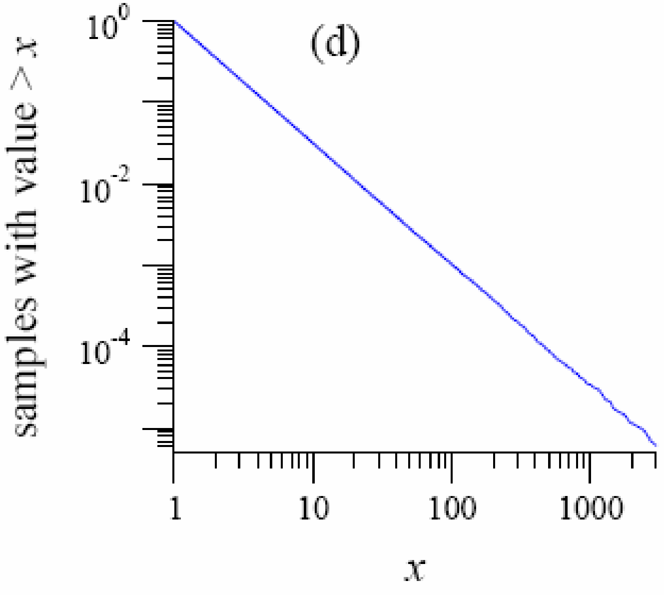
Logarithmic binning

- Bin the observations in bins of exponential size



Cumulative distribution

- Fit a line on the log-log plot of the cumulative distribution
  - it also follows a power-law with exponent α-1



Maximum likelihood estimation

- Assume that the data are produced by a power-law distribution with some exponent α
- Find the exponent that maximizes the probability P(α|x)

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

Divergent(?) mean

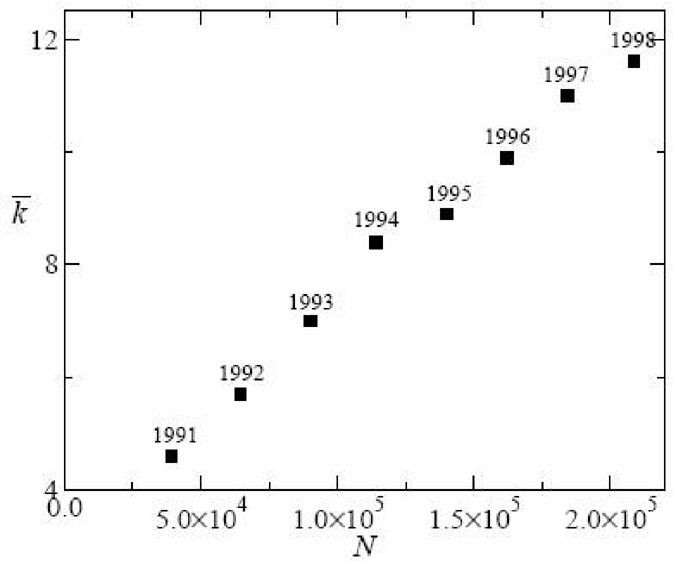
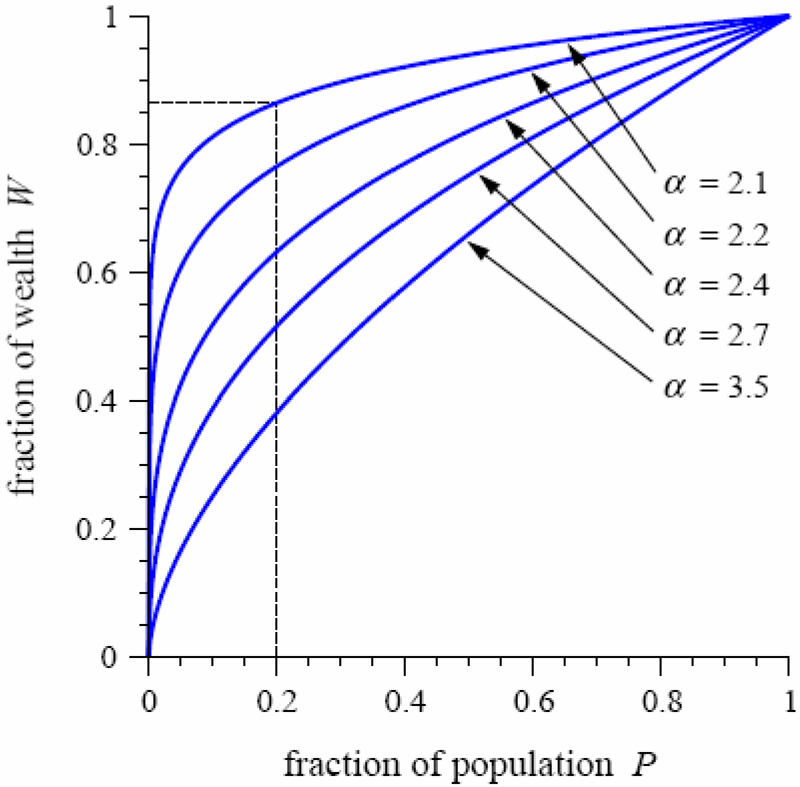


FIG. 3.17. Variation of the mean number of coauthorships (the average degree $\bar{k}$) of the network of coauthorships in neuroscience journals with increasing number of authors, N (according to Barabási, Jeong, Néda, Ravasz, Schubert, and Vicsek 2002).

The 80/20 rule

- Cumulative distribution is top-heavy



Power Laws – Generative processes

§ We have seen that power-laws appear in various natural, or man-made systems

§ What are the processes that generate power-laws?

§ Is there a “universal” mechanism?

Preferential attachment

- The main idea is that "the rich get richer"
  - first studied by Yule for the size of biological genera
  - revisited by Simon
  - reinvented multiple times

- Also known as
  - Gibrat principle
  - cumulative advantage
  - Mathew effect

The Yule process

- The setting:
  - a set of species defines a genus
  - the number of species in genera follows a power-law

- The Yule process:
  - at the n-th step of the process we have n genera
  - m new species are added to the existing genera through speciation evens: an existing species splits into two
  - the generation of the $(m+1)$-th species causes the creation of the $(n+1)$-th genera containing 1 species

- The sizes of genera follows a power law with

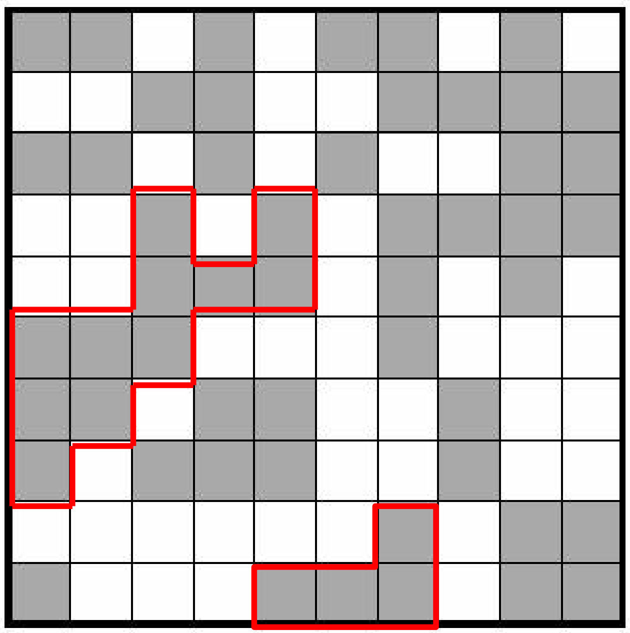
$$p_k \sim k^{-(2+1/m)}$$

Critical phenomena

- When the characteristic scale of a system diverges, we have a phase transition.
- Critical phenomena happen at the vicinity of the phase transition. Power-law distributions appear
- Phase transitions are also referred to as threshold phenomena

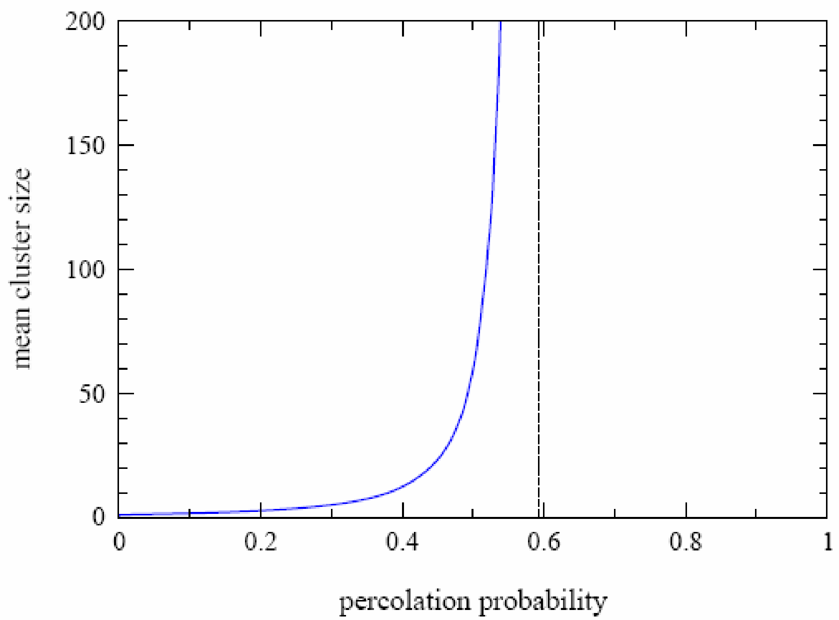
Percolation on a square lattice

- Each cell is occupied with probability p



- What is the mean cluster size?

Critical phenomena and power laws

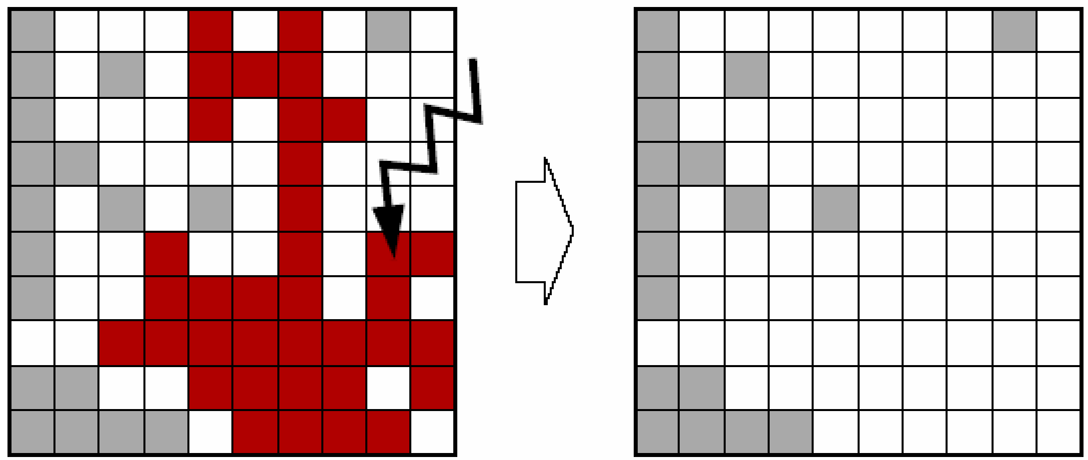


$p_c = 0.5927462\ldots$

- For $p < p_c$ mean size is independent of the lattice size
- For $p > p_c$ mean size diverges (proportional to the lattice size - percolation)
- For $p = p_c$ we obtain a power law distribution on the cluster sizes

Self Organized Criticality

- Consider a dynamical system where trees appear in randomly at a constant rate, and fires strike cells randomly



- The system eventually stabilizes at the critical point, resulting in power-law distribution of cluster (and fire) sizes

The idea behind self-organized criticality (more or less)

- There are two contradicting processes
  - e.g., planting process and fire process
- For some choice of parameters the system stabilizes to a state that no process is a clear winner
  - results in power-law distributions
- The parameters may be tunable so as to improve the chances of the process to survive
  - e.g., customer’s buying propensity, and product quality.

Combination of exponentials

- If variable Y is exponentially distributed

$$p(y) \sim e^{ay}$$

- If variable X is exponentially related to Y

$$X \sim e^{bY}$$

- Then X follows a power law

$$p(x) \sim x^{-(1+a/b)}$$

- Model for population of organisms

Monkeys typing randomly

- Consider the following generative model for a language [Miller 57]
  - The space appears with probability q_s
  - The remaining m letters appear with equal probability $(1-q_s)/m$

- Frequency of words follows a power law!
- Real language is not random. Not all letter combinations are equally probable, and there are not many long words

Least effort principle

- Let C_j be the cost of transmitting the j-th most frequent word

$$C_j \sim \log_m j$$

- The average cost is

$$C = \sum_{j=1}^{n} p_j C_j$$

- The average information content is

$$H = -\sum_{j=1}^{n} p_j \log_2 p_j$$

- Minimizing cost per information unit C/H yields

$$p_j \sim j^{-\alpha}$$

The log-normal distribution

- The variable Y = log X follows a normal distribution

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln x-\mu)^2/2\sigma^2}$$

$$\ln f(x) = -\frac{(\ln x)^2}{2\sigma^2} + \left(\frac{\mu}{\sigma^2} - 1\right)\ln x - \ln\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2}$$

Lognormal distribution

- Generative model: Multiplicative process

$$X_j = F_j X_{j-1}$$

$$\ln X_j = \ln X_0 + \sum_{k=1}^{j} \ln F_k$$

- Central Limit Theorem: If $X_1, X_2, \ldots, X_n$ are i.i.d. variables with mean m and finite variance s, then if $S_n = X_1 + X_2 + \ldots + X_n$

$$\frac{S_n - nm}{\sqrt{ns^2}} \sim N(0,1)$$

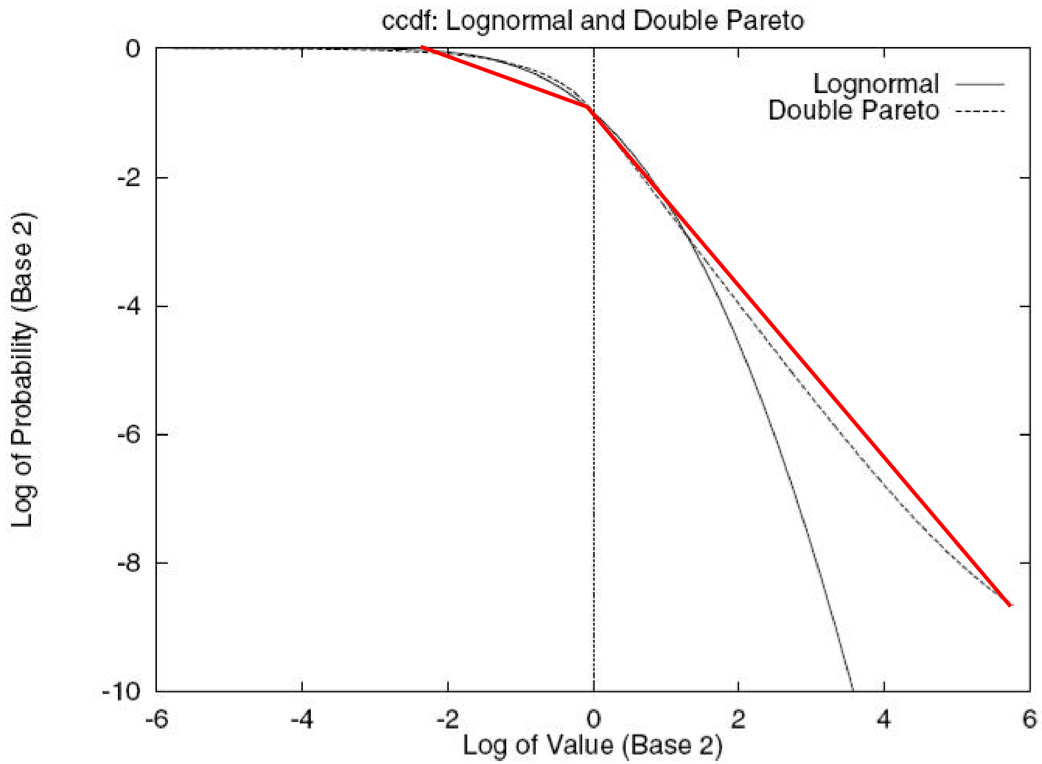
Example – Income distribution

- Start with some income X_0
- At time t with probability 1/3 double the income, with probability 2/3 cut the income in half
- The probability of having income x after n steps follows a log-normal distribution

- BUT... if we have a reflective boundary
  - when reaching income X_0 with probability 2/3 maintain the same income
- then the distribution follows a power-law!

Double Pareto distribution

- Double Pareto: Combination of two Pareto distributions



Double Pareto distribution

- Run the multiplicative process for T steps, where T is an exponentially distributed random variable

References

- M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*
- M. Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions, Internet Mathematics
- Lada Adamic, Zipf, power-laws and Pareto -- a ranking tutorial. http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html