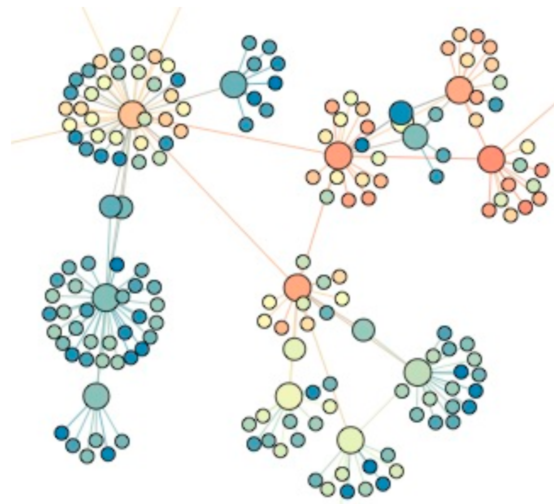


SNA 3D: Power laws

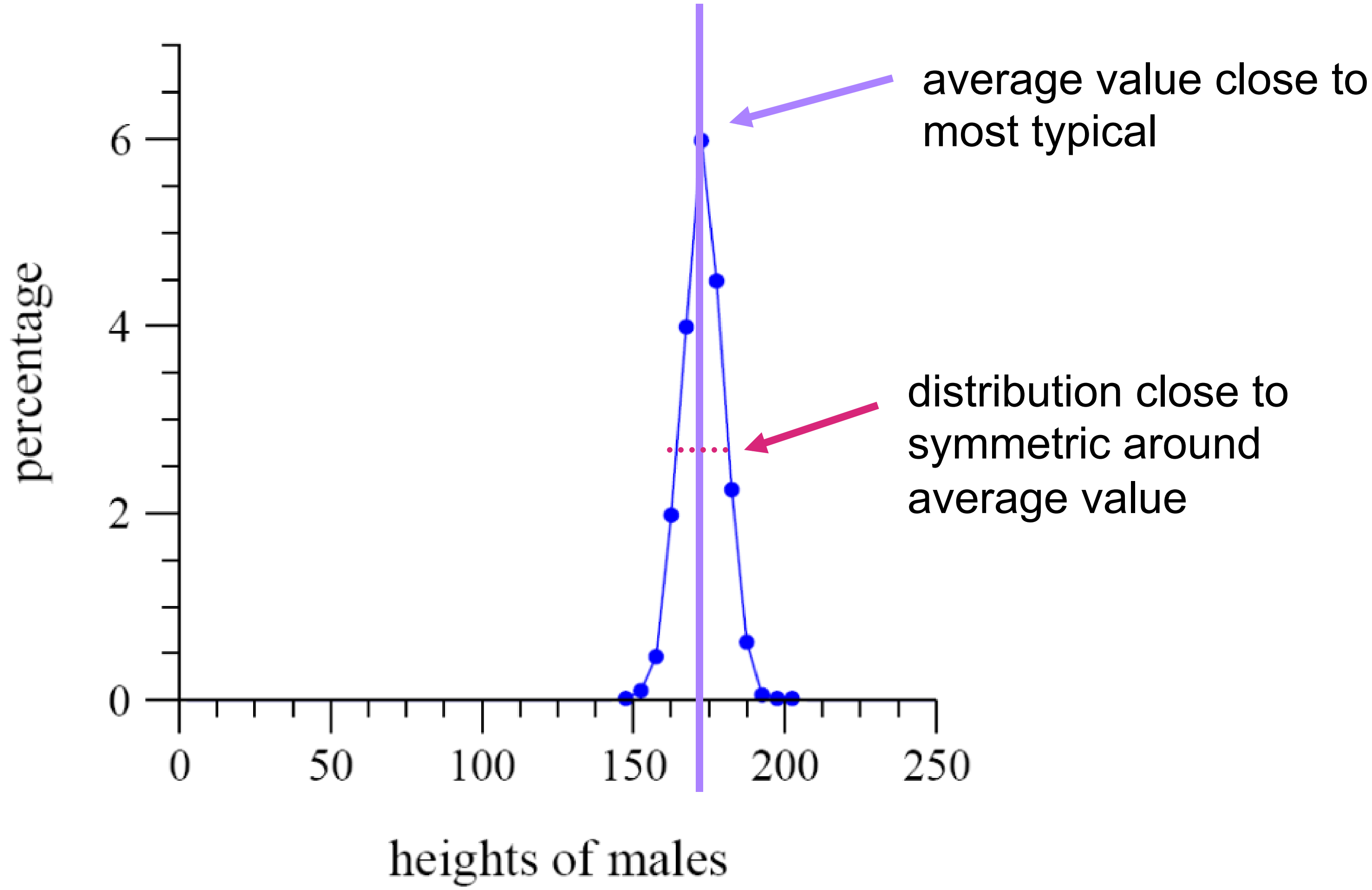
Lada Adamic

school of information

Heavy tails: right skew

- Right skew
  - normal distribution (not heavy tailed)
    - e.g. heights of human males: centered around 180cm (5' 11'')
  - Zipf's or power-law distribution (heavy tailed)
    - e.g. city population sizes: NYC 8 million, but many, many small towns

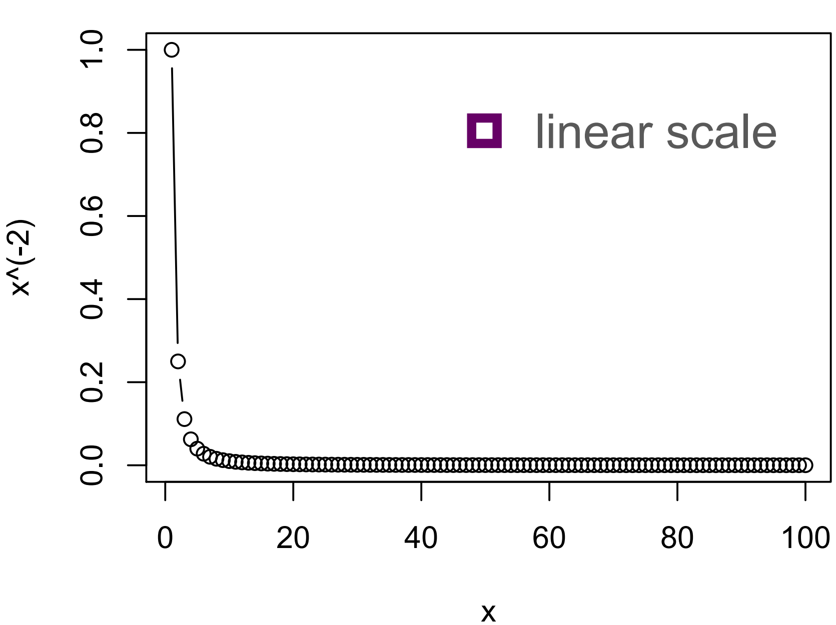
Normal distribution (human heights)

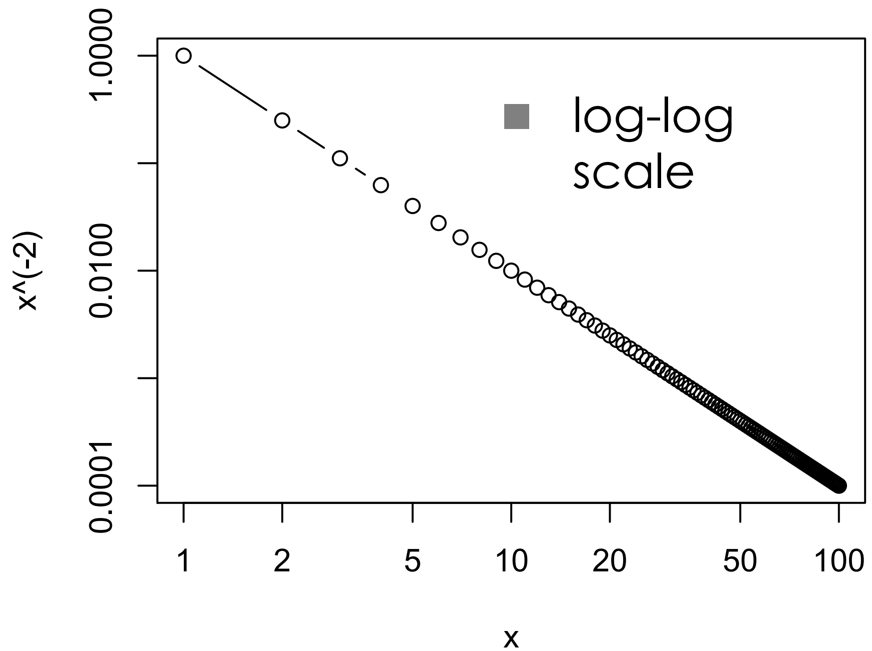


Heavy tails: max to min ratio

- High ratio of max to min
  - human heights
    - tallest man: 272cm (8' 11"), shortest man: (1' 10") *ratio: 4.8*
      from the Guinness Book of world records
  - city sizes
    - NYC: pop. 8 million, Duffield, Virginia pop. 52, *ratio: 150,000*

Power-law distribution

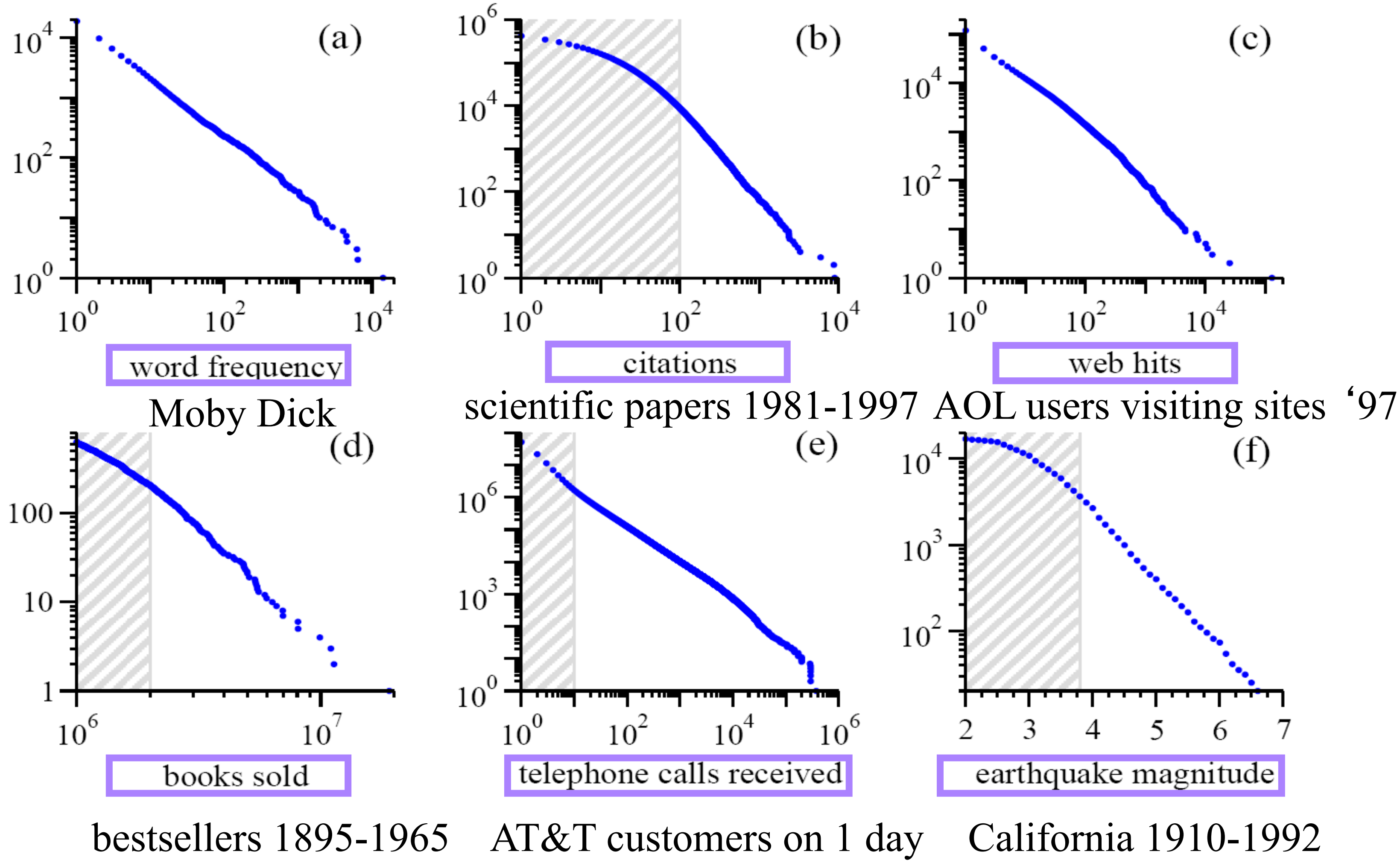




- high skew (asymmetry)
- straight line on a log-log plot

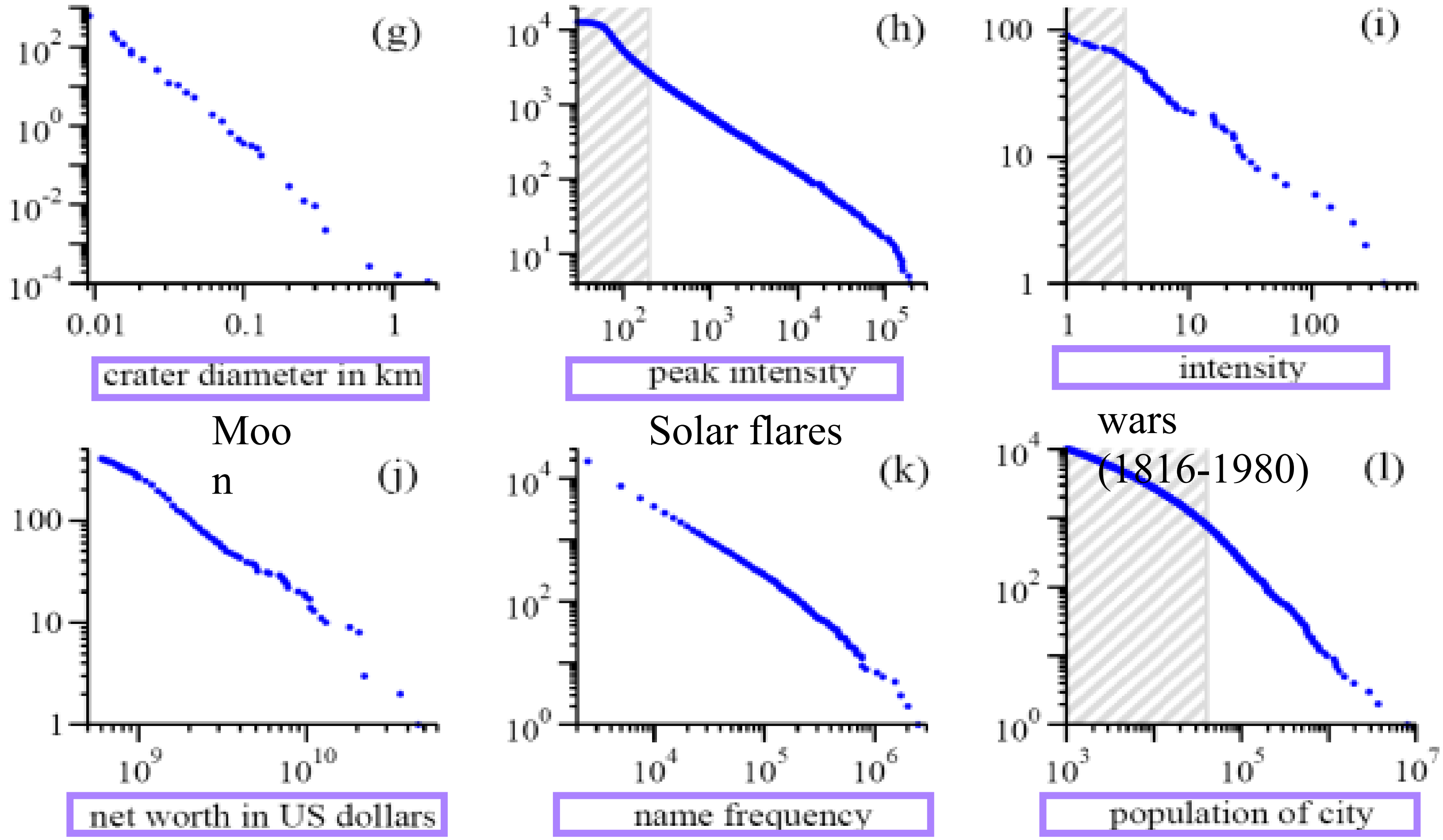
Power laws are seemingly everywhere

note: these are cumulative distributions, more about this in a bit…



Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

Yet more power laws



Moon

Solar flares

wars (1816-1980)

richest individuals 2003

US family names 1990

US cities 2003

Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

Power law distribution

- Straight line on a log-log plot

$$\ln(p(x)) = c - \alpha \ln(x)$$

- Exponentiate both sides to get that *p(x)*, the probability of observing an item of size 'x' is given by

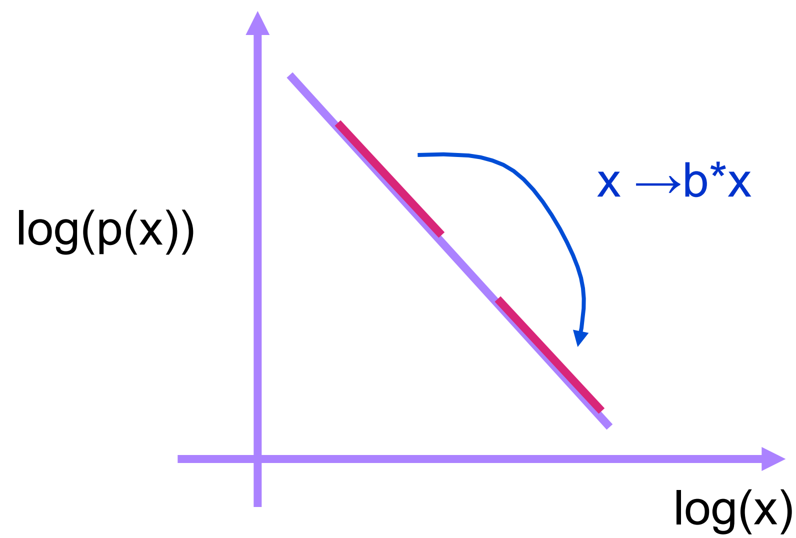
$$p(x) = Cx^{-\alpha}$$

normalization constant (probabilities over all *x* must sum to 1)

power law exponent α

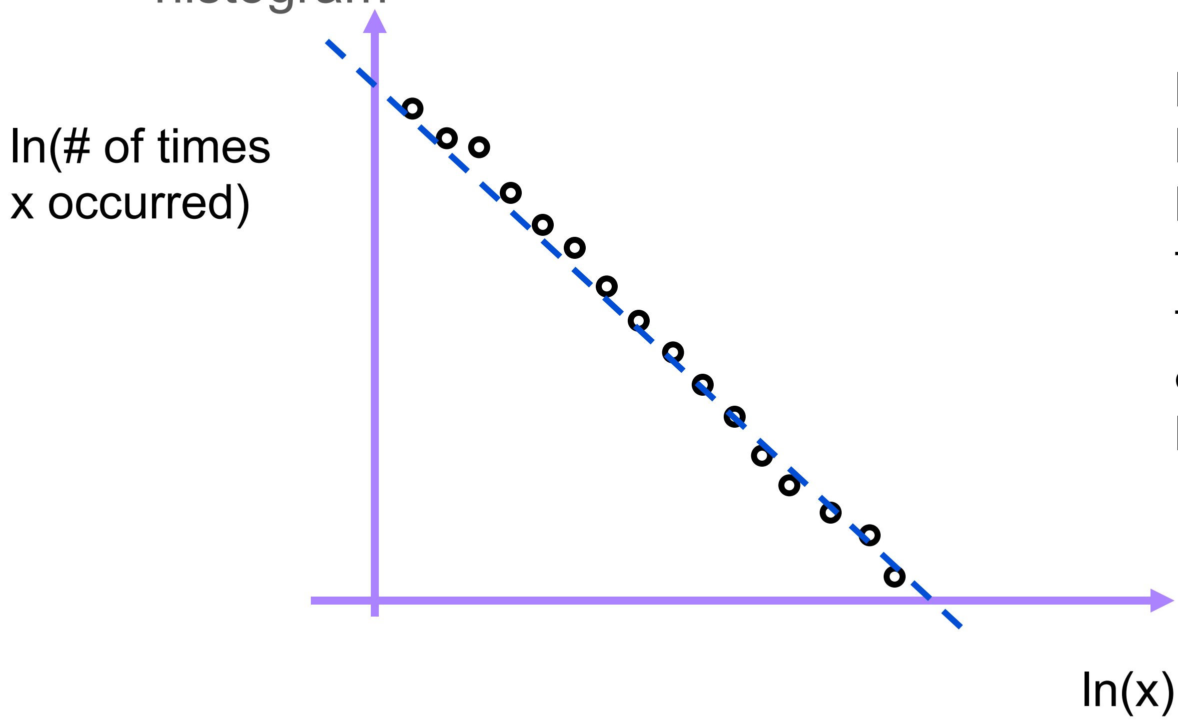
What does it mean to be scale free?

- A power law looks the same no mater what scale we look at it on (2 to 50 or 200 to 5000)
- Only true of a power-law distribution!
- $p(bx) = g(b)\ p(x)$ – shape of the distribution is unchanged except for a multiplicative constant
- $p(bx) = (bx)^{-\alpha} = b^{-\alpha}\ x^{-\alpha}$



Fitting power-law distributions

- Most common and not very accurate method:
  - Bin the different values of x and create a frequency histogram



ln(x) is the natural logarithm of x,
but any other base of the logarithm will give the same exponent of α because
$\log_{10}(x) = \ln(x)/\ln(10)$

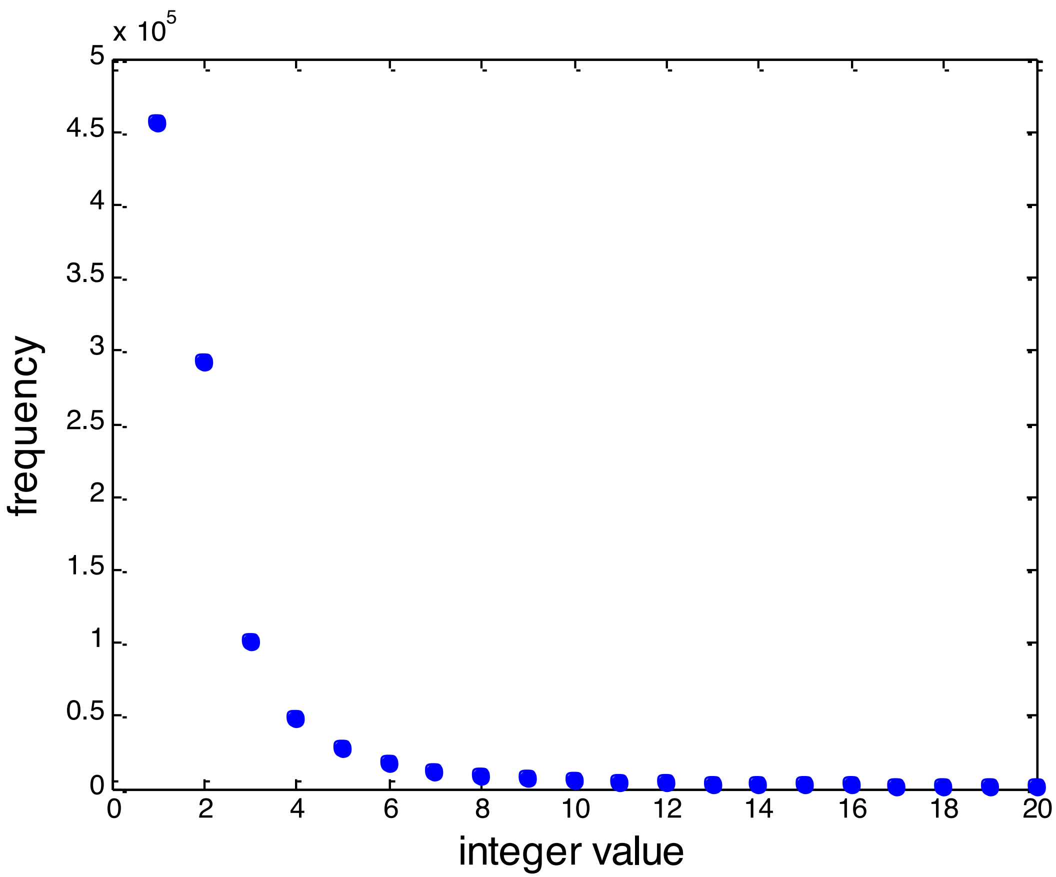
x can represent various quantities, the indegree of a node, the magnitude of an earthquake, the frequency of a word in text

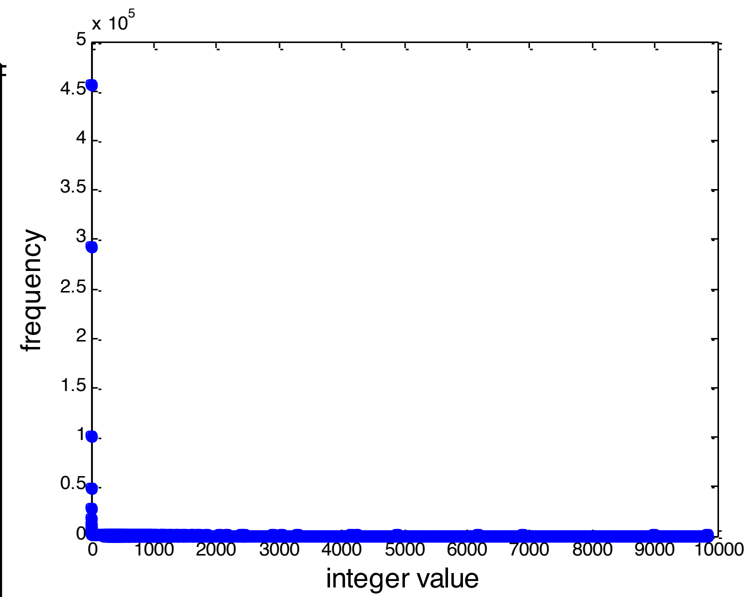
Example on an artificially generated data set

- Take 1 million random numbers from a distribution with $\alpha = 2.5$
- Can be generated using the so-called 'transformation method'
- Generate random numbers r on the unit interval $0 \le r < 1$
- then $x = (1-r)^{-1/(\alpha-1)}$ is a random power law distributed real number in the range $1 \le x < \infty$

Linear scale plot of straight bin of the data

- Number of times 1 or 3843 or 99723 occured
- Power-law relationship not as apparent
- Only makes sense to look at smallest bins



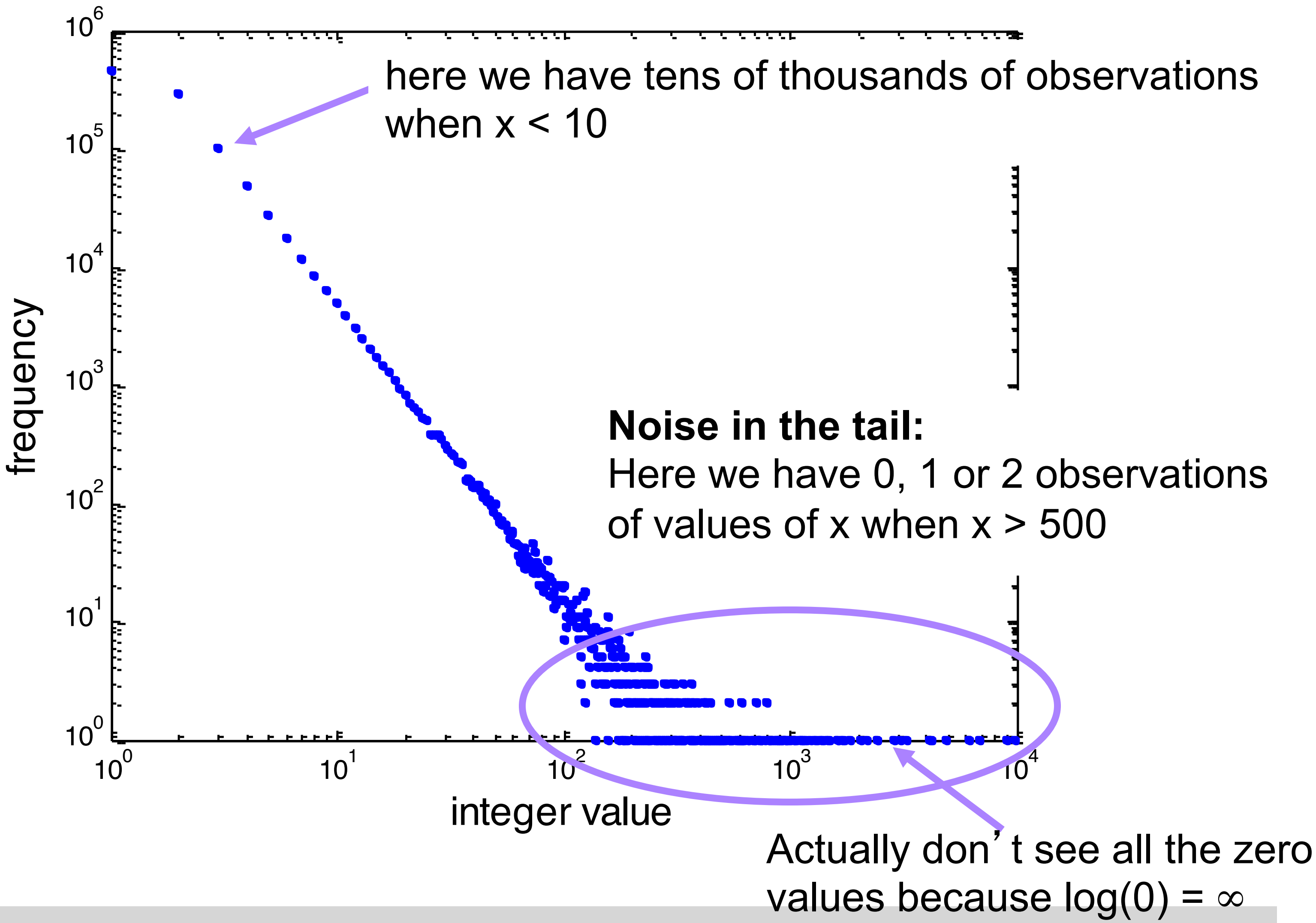


whole range

first few bins

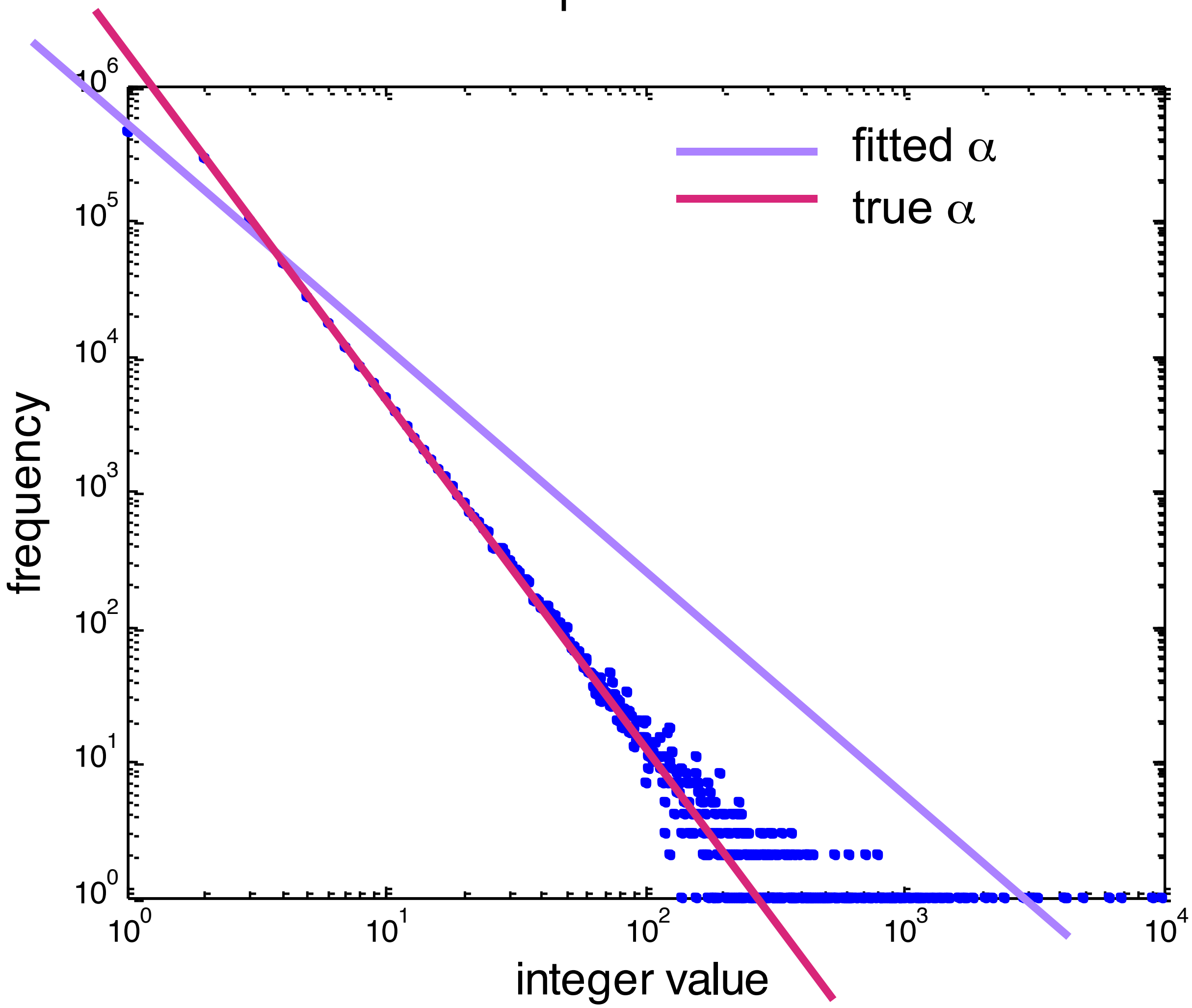
Log-log scale plot of simple binning of the data

- Same bins, but plotted on a log-log scale



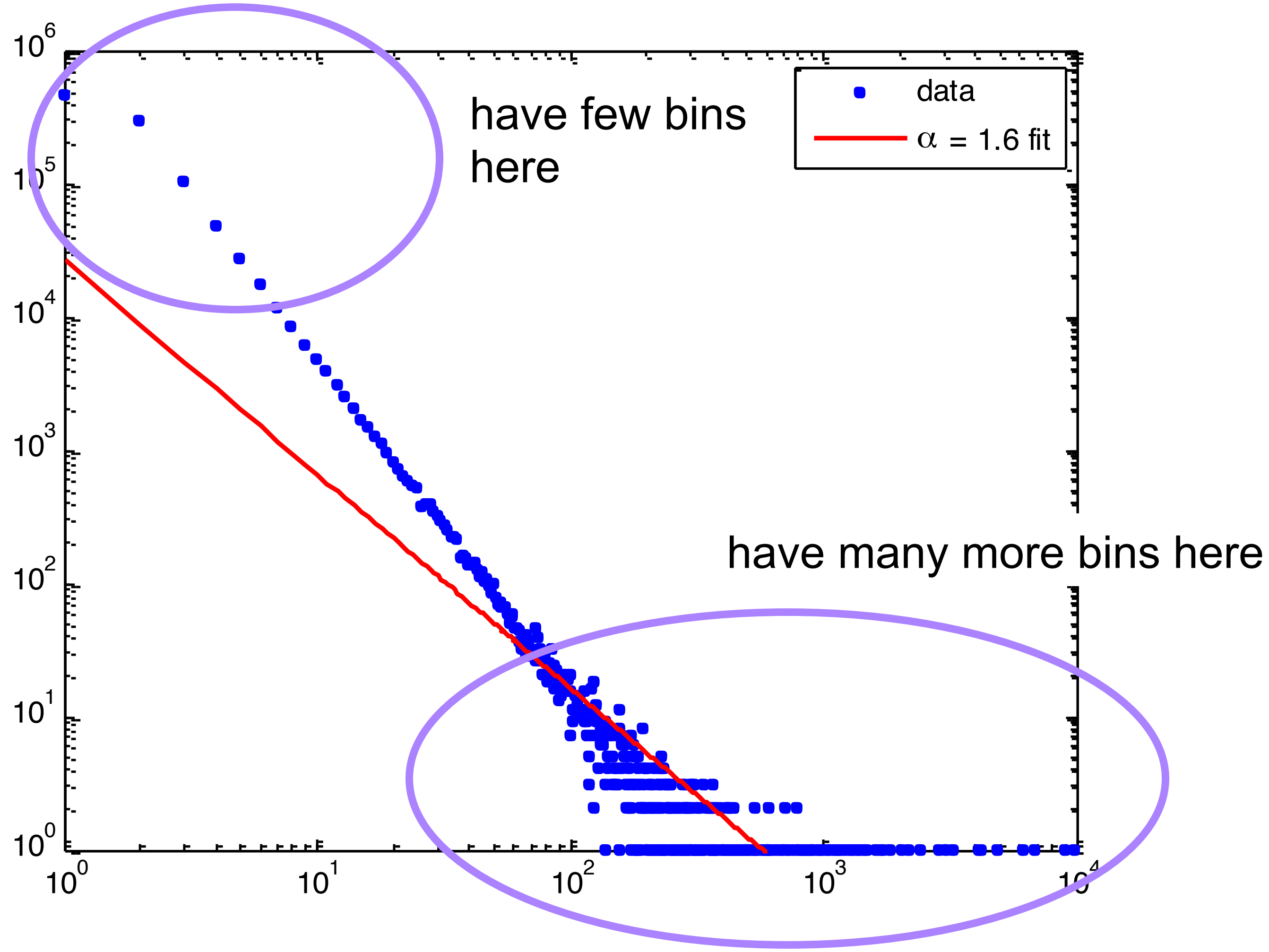
Log-log scale plot of straight binning of the data

- Fitting a straight line to it via least squares regression will give values of the exponent α that are too low



What goes wrong with straightforward binning

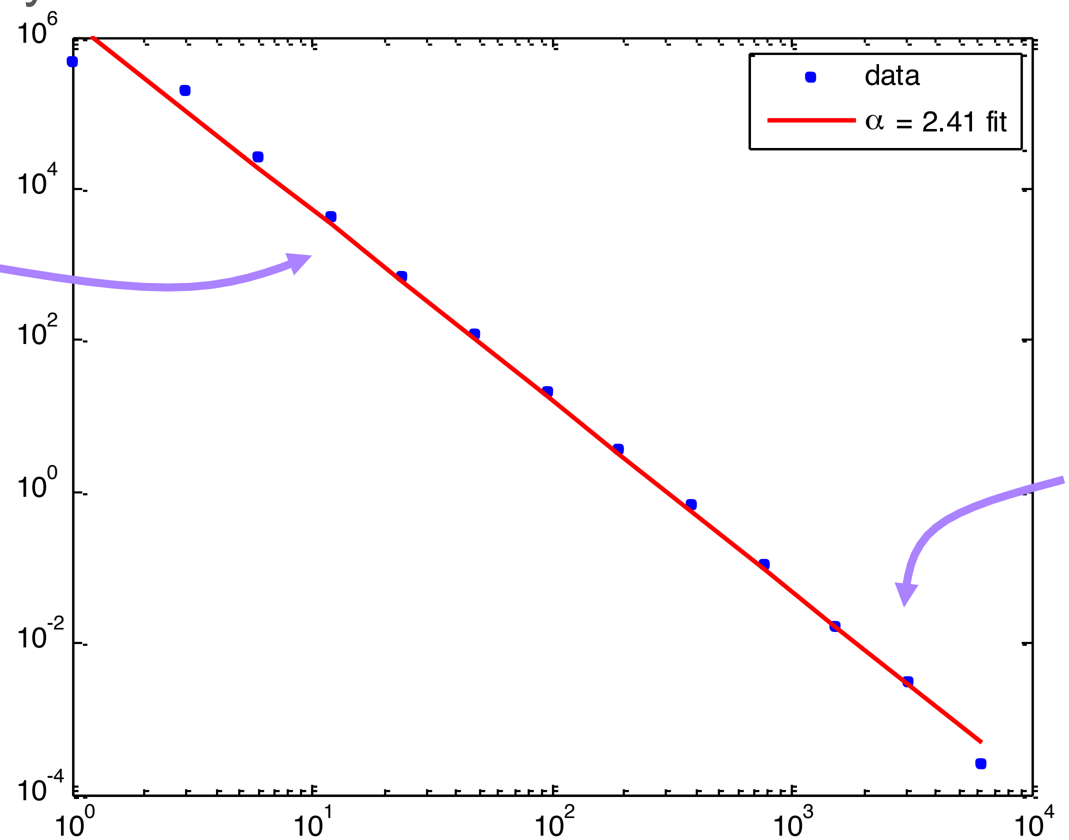
- Noise in the tail skews the regression result



First solution: logarithmic binning

- bin data into exponentially wider bins:
  - 1, 2, 4, 8, 16, 32, …
- normalize by the width of the bin

evenly spaced datapoints



less noise in the tail of the distribution

- disadvantage: binning smoothes out data but also loses information

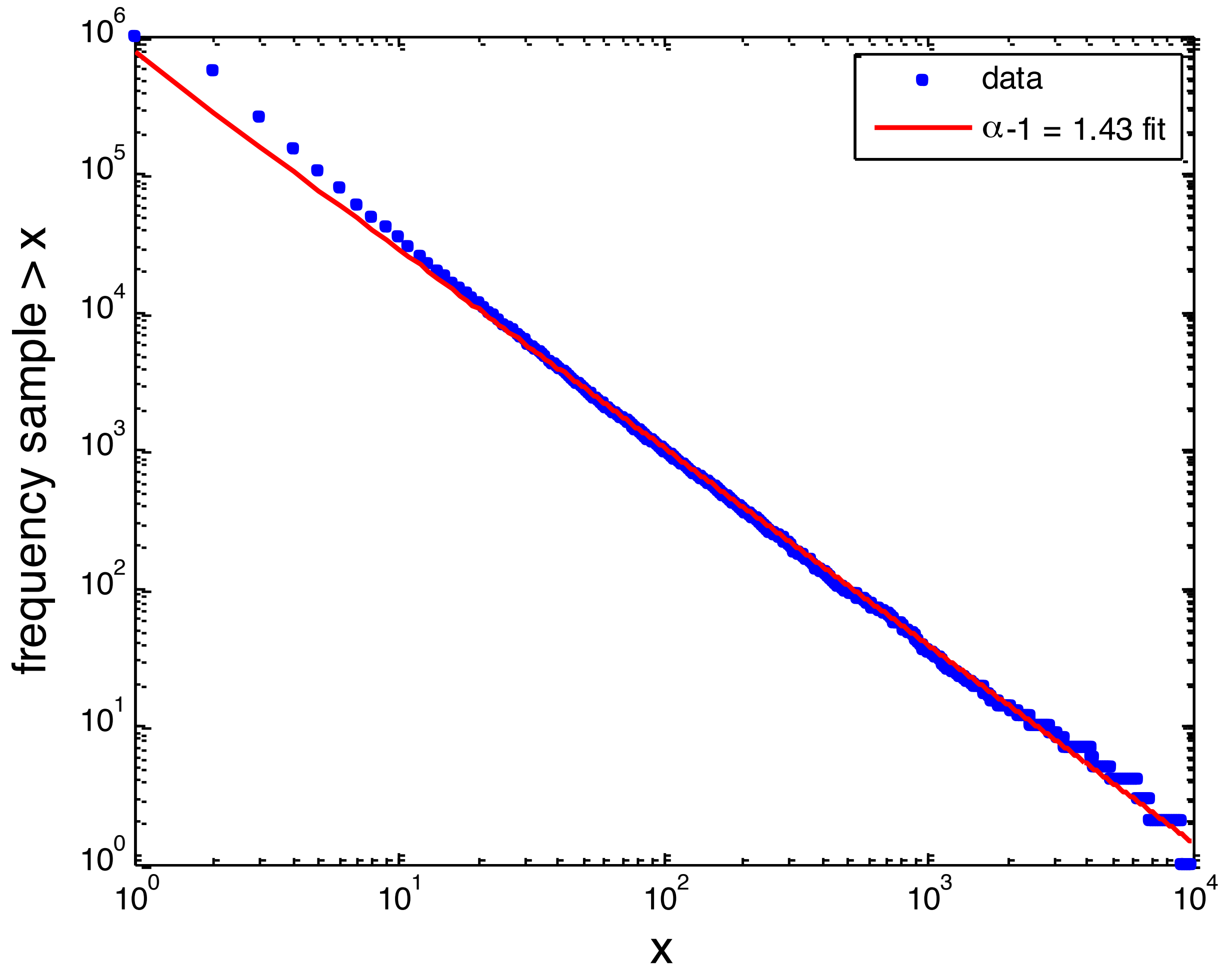
Second solution: cumulative binning

- No loss of information
  - No need to bin, has value at each observed value of x
- But now have cumulative distribution
  - i.e. how many of the values of x are at least X
  - The cumulative probability of a power law probability distribution is also power law but with an exponent α **- 1**

$$\int c x^{-\alpha} = \frac{c}{1-\alpha} x^{-(\alpha-1)}$$

Fitting via regression to the cumulative distribution

- fitted exponent (2.43) much closer to actual (2.5)

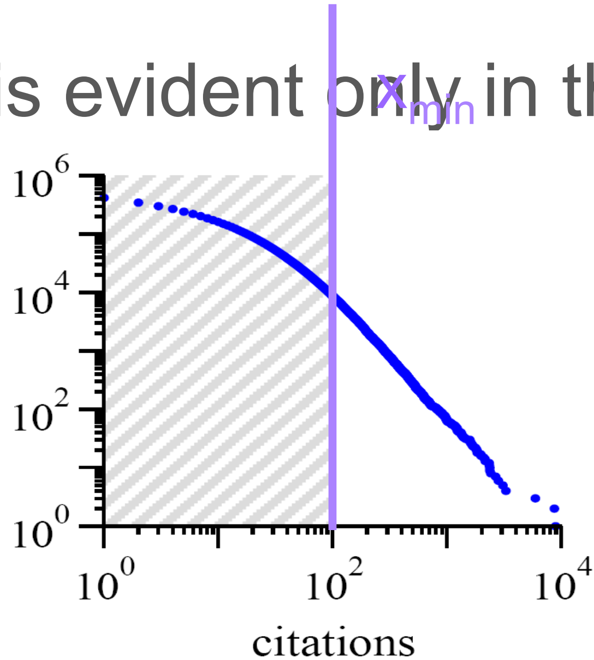


Where to start fitting?

- some data exhibit a power law only in the tail
- after binning or taking the cumulative distribution you can fit to the tail
- so need to select an x_{min} the value of x where you think the power-law starts
- certainly x_{min} needs to be greater than 0, because $x^{-\alpha}$ is infinite at $x = 0$

Example:

- Distribution of citations to papers
- power law is evident only in the tail (x_{min} > 100 citations)



Source:**MEJ Newman, 'Power laws, Pareto distributions and Zipf's law',** *Contemporary Physics* **46**, 323–351 (2005)

Maximum likelihood fitting – best

- You have to be sure you have a power-law distribution (this will just give you an exponent but not a goodness of fit)

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

- x_i are all your datapoints, and you have *n* of them
- for our data set we get $\alpha = 2.503$ – pretty close!

Some exponents for real world data

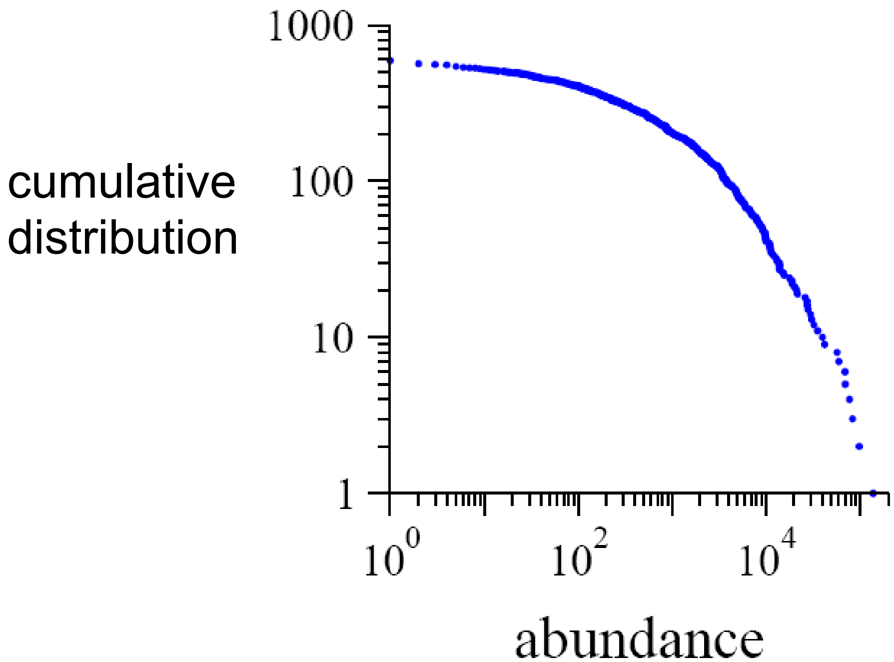
	x_{min}	exponent α
frequency of use of words	1	2.20
number of citations to papers	100	3.04
number of hits on web sites	1	2.40
copies of books sold in the US	2 000 000	3.51
telephone calls received	10	2.22
magnitude of earthquakes	3.8	3.04
diameter of moon craters	0.01	3.14
intensity of solar flares	200	1.83
intensity of wars	3	1.80
net worth of Americans	$600m	2.09
frequency of family names	10 000	1.94
population of US cities	40 000	2.30

Many real world networks are power law

	exponent α (in/out degree)
film actors	2.3
telephone call graph	2.1
email networks	1.5/2.0
sexual contacts	3.2
WWW	2.3/2.7
internet	2.5
peer-to-peer	2.1
metabolic network	2.2
protein interactions	2.4

Hey, not everything is a power law

- number of sightings of 591 bird species in the North American Bird survey in 2003.



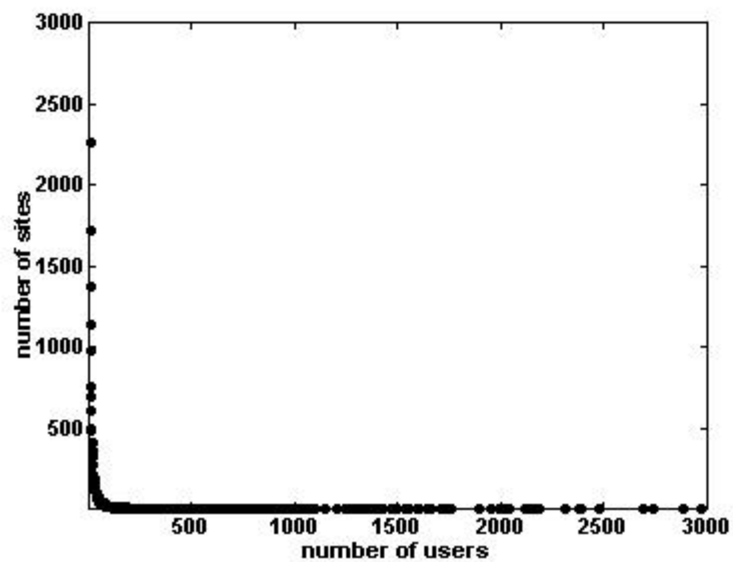
- another example:
  - size of wildfires (in acres)

Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

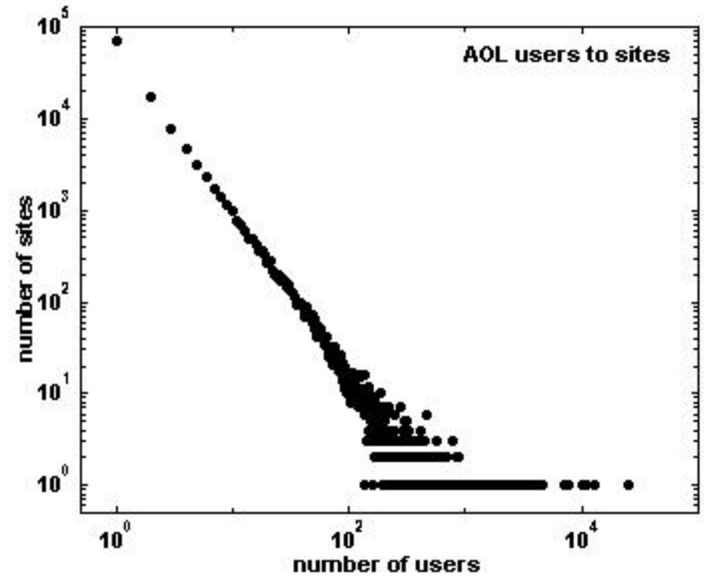
Not every network is power law distributed

- reciprocal, frequent email communication
- power grid
- Roget's thesaurus
- company directors…

Example on a real data set: number of AOL visitors to different websites back in 1997



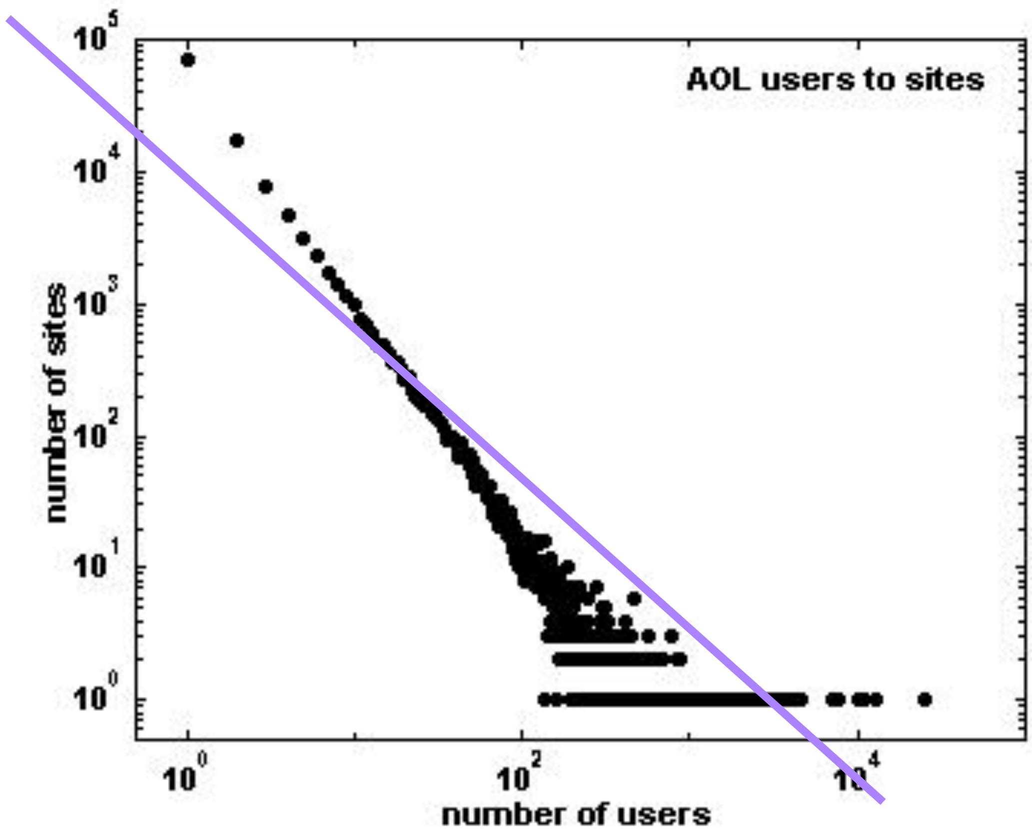
simple binning on a linear scale



simple binning on a log-log scale

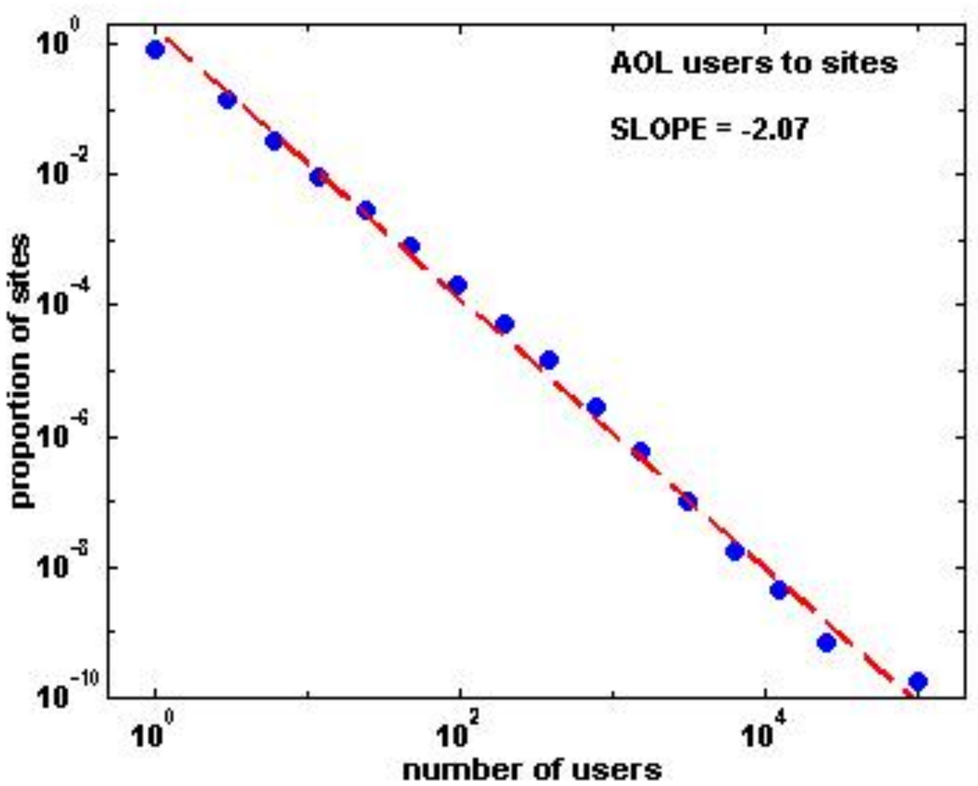
trying to fit directly…

- direct fit is too shallow: α = 1.17…



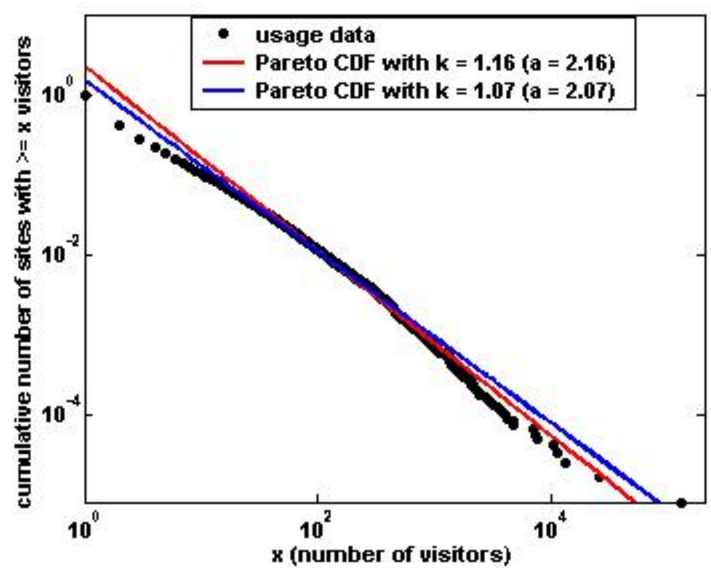
Binning the data logarithmically helps

- select exponentially wider bins
  - 1, 2, 4, 8, 16, 32, ….



Or we can try fitting the cumulative distribution

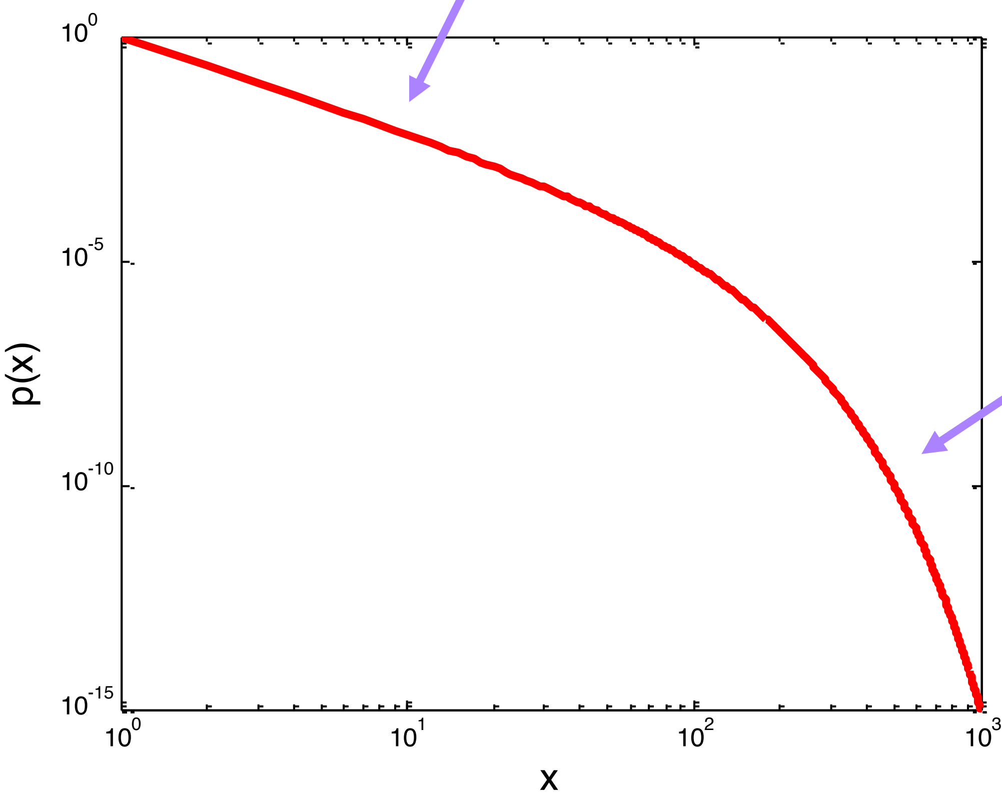
- Shows perhaps 2 separate power-law regimes that were obscured by the exponential binning
- Power-law tail may be closer to 2.4



Another common distribution: power-law with an exponential cutoff

- $p(x) \sim x^{-a} e^{-x/\kappa}$

starts out as a power law



ends up as an exponential

but could also be a lognormal or double exponential…

Zipf &Pareto:
what they have to do with power-laws

- Zipf
  - George Kingsley Zipf, a Harvard linguistics professor, sought to determine the 'size' of the 3rd or 8th or 100th most common word.
  - Size here denotes the frequency of use of the word in English text, and not the length of the word itself.
  - Zipf's law states that the size of the r'th largest occurrence of the event is inversely proportional to its rank:

    $\boldsymbol{y \sim r^{-\beta}}$, with β close to unity.

So how do we go from Zipf to Pareto?

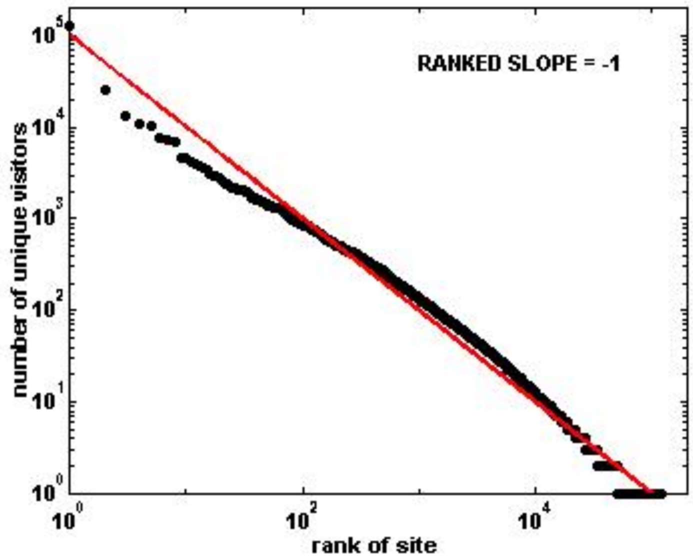
- The phrase "The ***r*** th largest city has ***n*** inhabitants" is equivalent to saying "***r*** cities have ***n*** or more inhabitants".

- This is exactly the definition of the Pareto distribution, except the x and y axes are flipped. Whereas for Zipf, ***r*** is on the x-axis and ***n*** is on the y-axis, for Pareto, ***r*** is on the y-axis and ***n*** is on the x-axis.

- Simply inverting the axes, we get that if the rank exponent is β, i.e.
  $n \sim r^{-\beta}$ for Zipf,   (n = income, r = rank of person with income n)
  then the Pareto exponent is $1/\beta$ so that
  $r \sim n^{-1/\beta}$ (n = income, r = number of people whose income is n or higher)

Zipf's law & AOL site visits

- Deviation from Zipf's law
  - slightly too few websites with large numbers of visitors



Zipf' s Law and city sizes (~1930) [2]

Rank(k)	City	Population (1990)	Zips' s Law $10{,}000{,}000/k$	Modified Zipf' s law: (Mandelbrot) $5{,}000{,}000\big/\left(k-\frac{2}{5}\right)^{3/4}$
1	Now York	7,322,564	10,000,000	7,334,265
7	Detroit	1,027,974	1,428,571	1,214,261
13	Baltimore	736,014	769,231	747,693
19	Washington DC	606,900	526,316	558,258
25	New Orleans	496,938	400,000	452,656
31	Kansas City	434,829	322,581	384,308
37	Virgina Beach	393,089	270,270	336,015
49	Toledo	332,943	204,082	271,639
61	Arlington	261,721	163,932	230,205
73	Baton Rouge	219,531	136,986	201,033
85	Hialeah	188,008	117,647	179,243
97	Bakersfield	174,820	103,270	162,270

not any more

slide: Luciano Pietronero

80/20 rule

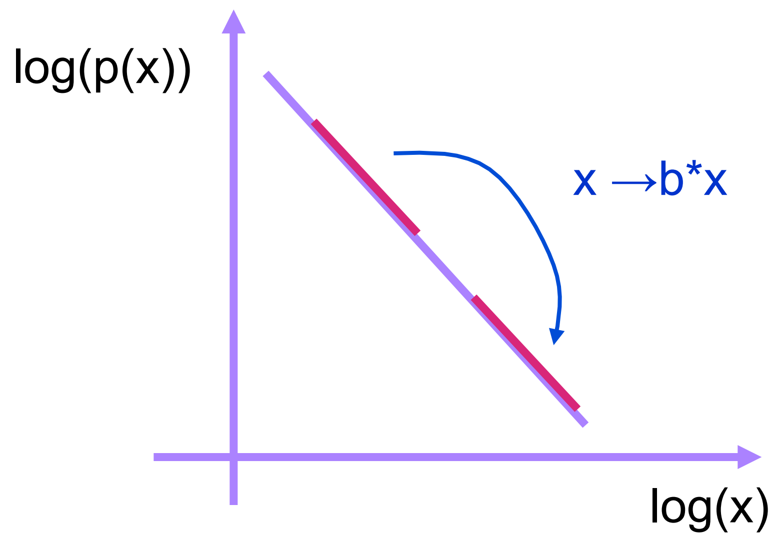
- The fraction W of the wealth in the hands of the richest P of the the population is given by

$$W = P^{(\alpha-2)/(\alpha-1)}$$

- Example: US wealth: $\alpha = 2.1$
  - richest 20% of the population holds 86% of the wealth

What does it mean to be scale free?

- A power law looks the same no mater what scale we look at it on (2 to 50 or 200 to 5000)
- Only true of a power-law distribution!
- p(bx) = g(b) p(x) – shape of the distribution is unchanged except for a multiplicative constant
- $p(bx) = (bx)^{-\alpha} = b^{-\alpha} x^{-\alpha}$



Wrap up on power-laws

- Power-laws are cool and intriguing
- But make sure your data is actually power-law before boasting