Statistical Laws of Extreme Events

Eduardo G. Altmann

School of Mathematics and Statistics
University of Sydney

git clone git@github.com:edugalt/ccc16.git

https://github.com/edugalt/ccc16/archive/master.zip

Complexity, Criticality and Computation (C3) Research Camp 2016

The probability of Trump wining is 20%

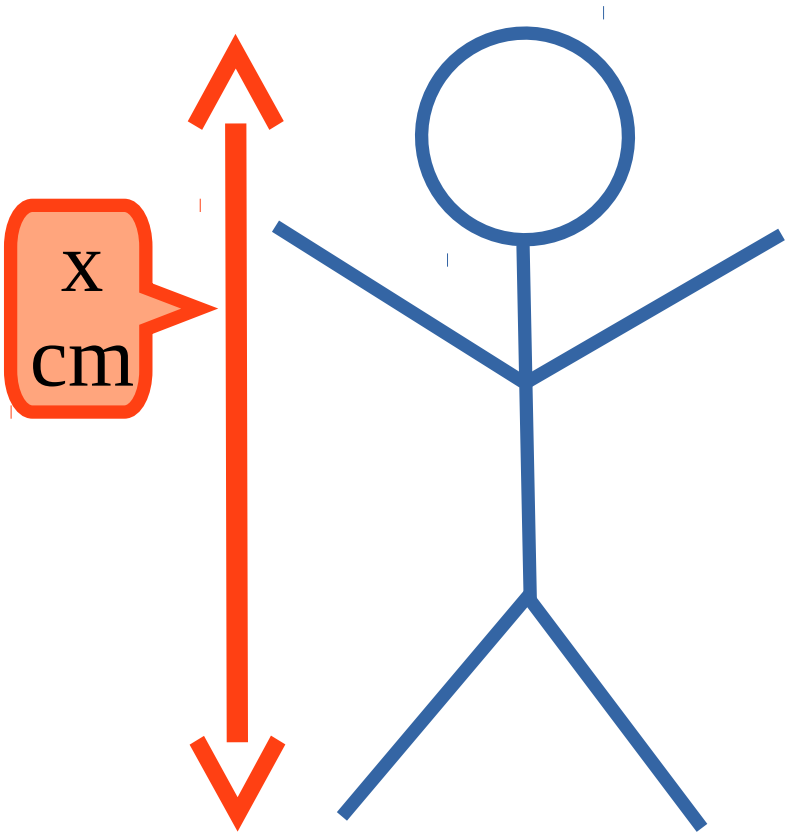
How much does a 6 pack of beer cost in Australia?

Answer 1: Which beer? Where? How do you want to pay?

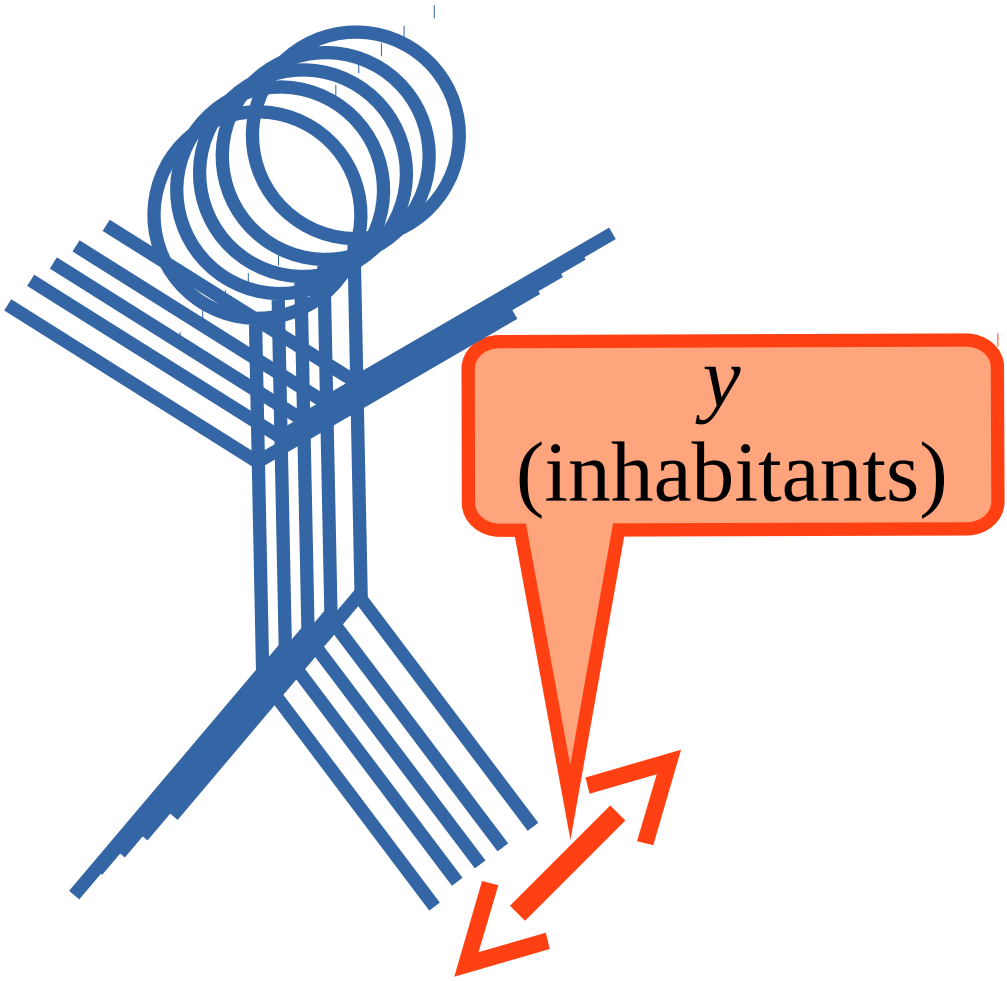
Answer 2: $19.3 \pm 1.8\ A\$$

Quiz

- how tall is my brother?



- what is the population of the city he lives?



Body sizes

City sizes

Height H is approx. Gaussian distributed

Brazil: $\langle H \rangle = 172\text{cm}$

U.S.A: $\langle H \rangle = 176\text{cm} \quad \sigma_H \approx 4\text{cm} \quad \frac{\sigma_H}{\langle H \rangle} = 0.02$

City population is highly non-Gaussian, varying from thousands to millions

Model 1:

$$\langle P \rangle = \frac{1}{N} \sum_{i=1}^{N} P_i$$

$$\frac{\sigma_P}{\langle P \rangle} = \frac{198.2}{34.0} = 5.8$$

Model 2:

$$\tilde{P} = \frac{1}{\sum w_i} \sum_{i=1}^{N} w_i P_i = \left\langle \frac{P^2}{\langle P \rangle} \right\rangle$$

$$w_i = P_i$$

$$\frac{\sigma_{P^2/\langle P \rangle}}{\tilde{P}} = \frac{49010}{1190} = 41.2$$

My brother: $H^* = 196 \text{ cm} \approx \langle H \rangle + 5\sigma_H$

$P^* \approx 11$ Millions

Hands on the city data!

git clone git@github.com:edugalt/ccc16.git

https://github.com/edugalt/ccc16/archive/master.zip

Definitions

Fraction of cities with population x:
Probability of finding an item with *x* counts:

$$P(x) \equiv \lim_{N\to\infty} \frac{\sum_{i=1}^{N} \delta(x_i = x)}{N} \approx \frac{N_x}{N}$$

n-th moment of the distribution:

$$\mu_n \equiv \sum_{i=1}^{\infty} x^n P(x) \equiv \langle x^n \rangle$$

Mean and Variance:

$$\mathbb{E}(x) = \mu_1,\ \mathbb{V}(x) \equiv \mu_2 - \mu_1^2, \sigma \equiv \sqrt{\mathbb{V}}$$

Definition: A probability distribution *P(x)* is said to have fat tails if

$$\exists \alpha \text{ such that } \mu_{\alpha'} \to \infty \text{ for all } \alpha' > \alpha$$

Example: power-law distribution $P(x) \sim x^{-\alpha}$, with $\alpha > 1$ and $x > 1$

Remarks: - for finite data, the moments of *P(x)* never diverge; for finite systems, *P(x)* often has cut-offs
- characteristic properties of fat-tailed data are a straight line in *P(x)* in log-log plots over several decades and a large (increasing with *N*) ratio of moments:

$$\frac{\mu_\alpha}{\mu_{\alpha-1}} \gg 1 \text{ often } \frac{\mathbb{E}(x)}{\mathbb{V}(x)} \gg 1$$

- binomial distribution B(N,p) has $\frac{\mathbb{E}(x)}{\mathbb{V}(x)} = (1-p)$

Properties of the power-law distribution: $P(x) \sim x^{-\alpha}, \text{ for large x}$

1. Linear in log-log scale: $\log P(x) = -\alpha \log x + \text{cst.}$

2. Scale invariant $P(sx) = s^{-\alpha} P(x) \sim x^{-\alpha}$

3. Power-law complementary cumulative distribution is also a power law: $P(x > X) \equiv \sum_{X}^{\infty} P(x) \sim X^{-(\alpha-1)}$

Expected number of items (out of N) with $x>X$: $r_i = N P(x > X_i) \sim N X_i^{-(\alpha-1)}$

$$\Downarrow$$

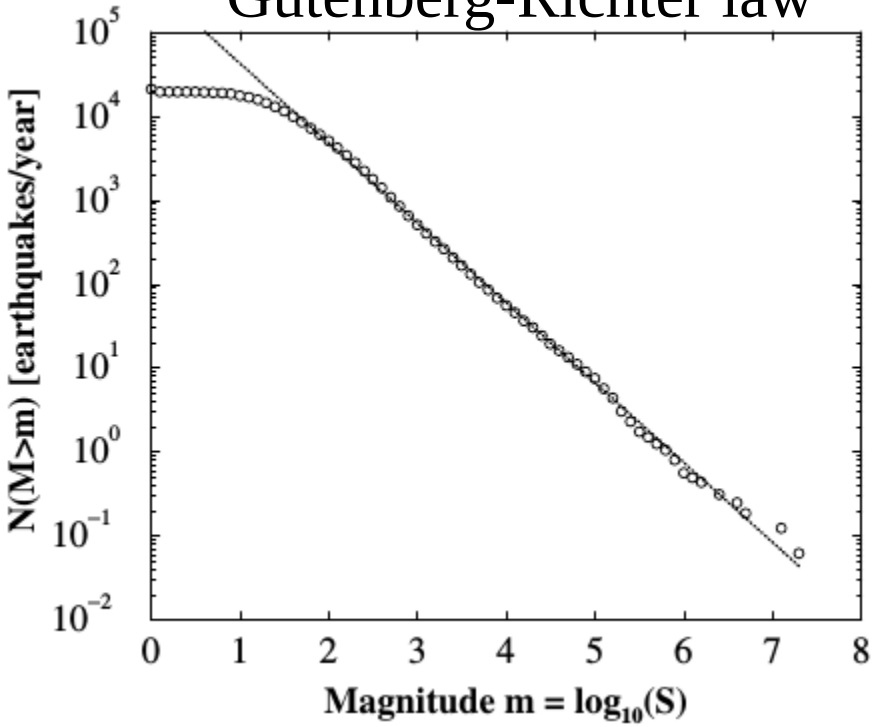
$X(r) \sim r^{-\frac{1}{\alpha-1}}$, for small ranks r

4. Power-law rank frequency distribution with exponent: $\beta = 1/(\alpha - 1) \qquad \alpha > 1 \Rightarrow \beta > 0$

Examples of data following fat-tailed distributions

Earthquakes, Avalanches, and Natural Disasters

Gutenberg-Richter law



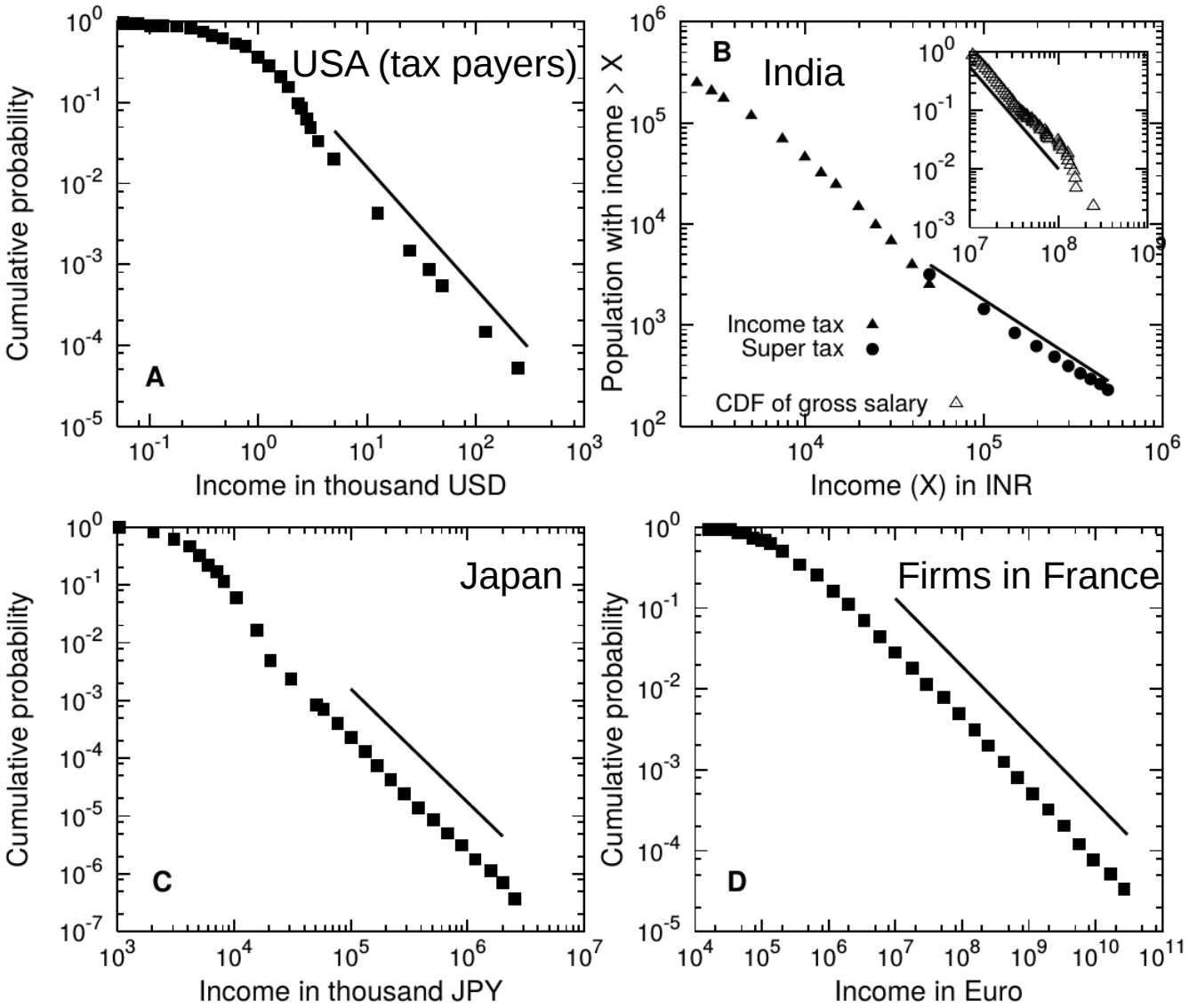
Bak et al. (2002)

The 1755 Lisbon earthquake [Wikipedia]

Inequality in economical systems

Fraction of the population with income larger than *x:*

$$P(X > x) \sim x^{-\alpha}, \text{ with } \alpha > 0$$

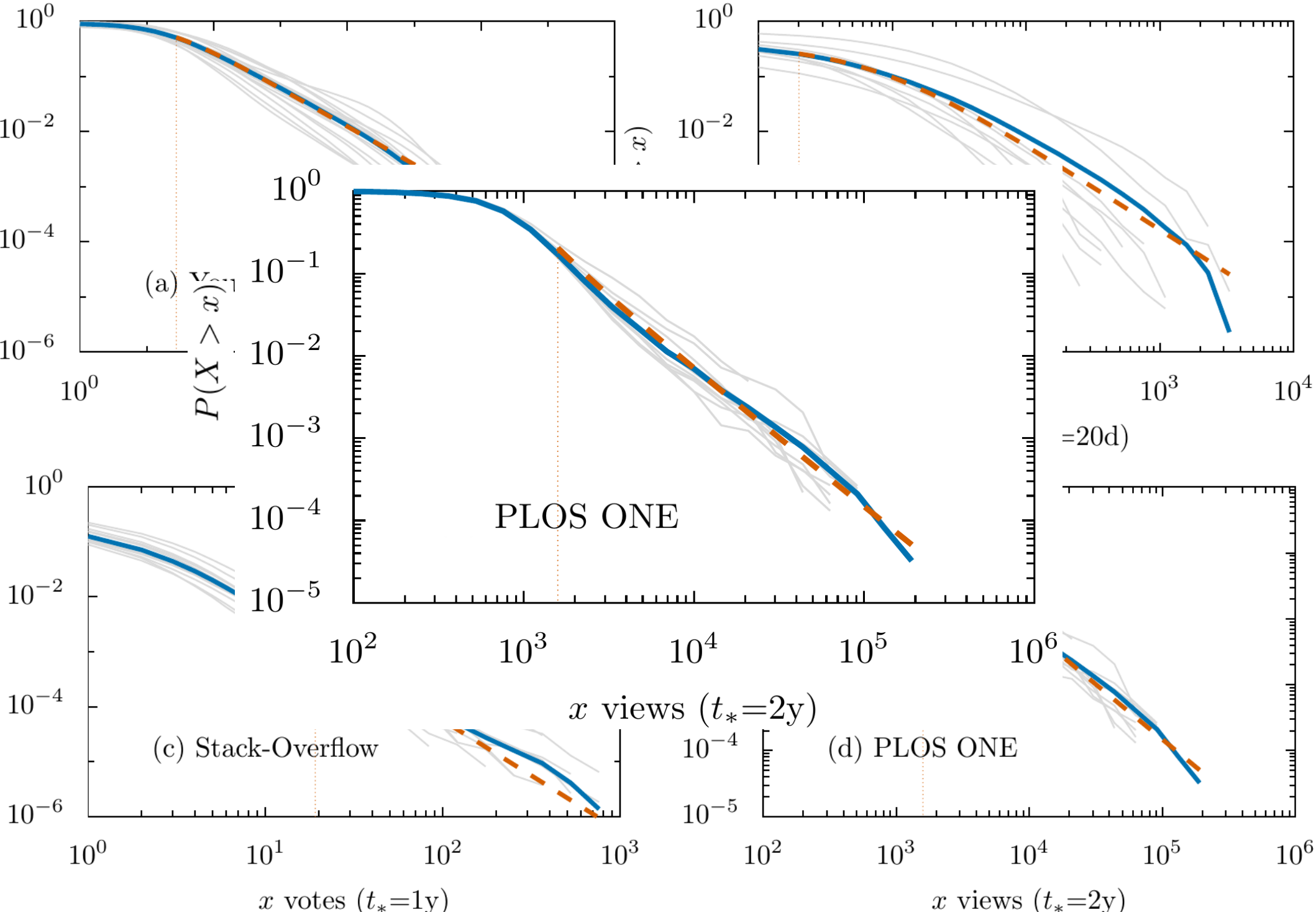


Vilfredo Pareto (1848-1923)

20-80 rule
"20% of the people have 80% of the wealth"

from Chatterjee et al. 2007

Inequality in the economy of attention

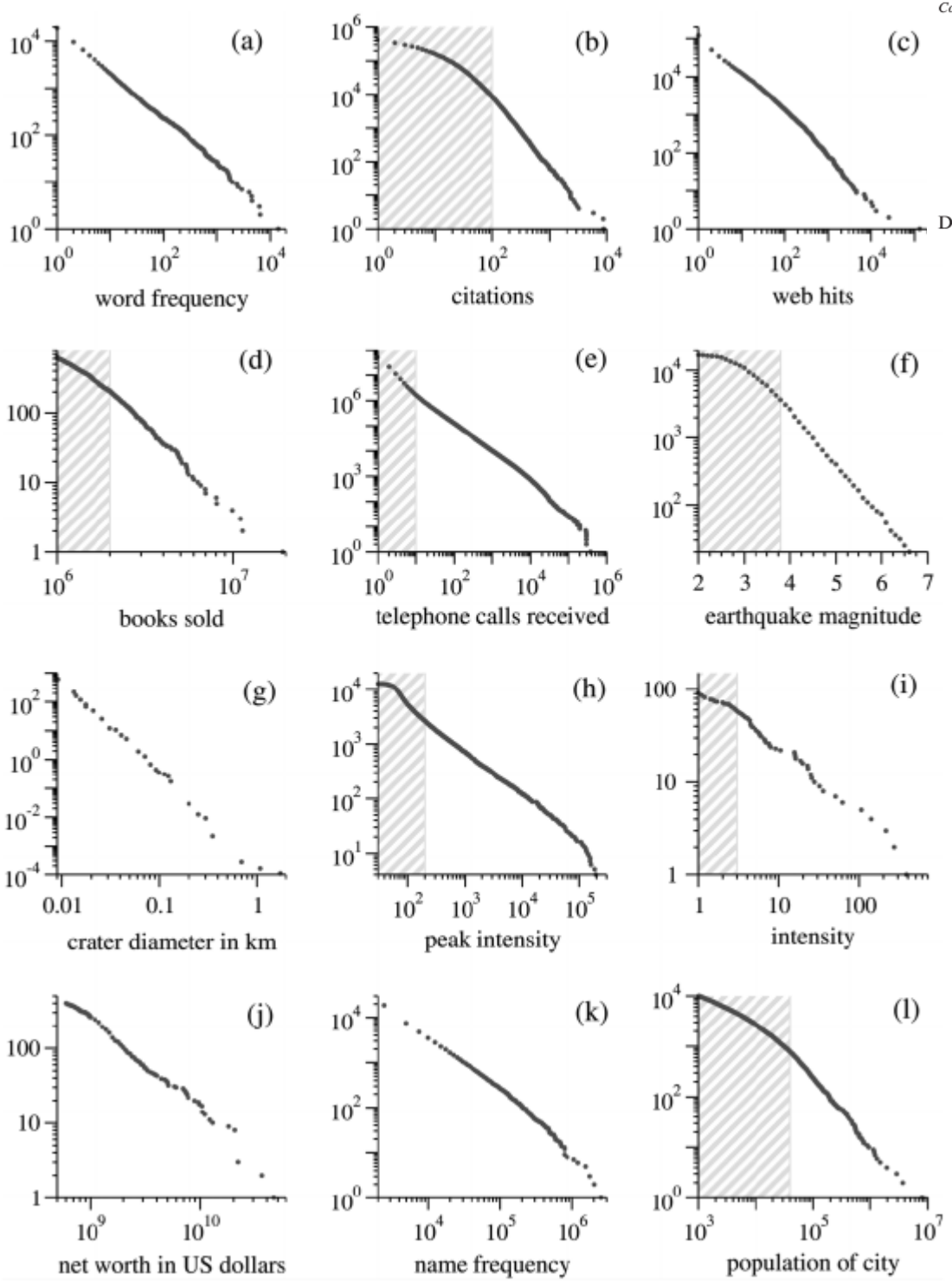


Contemporary Physics, Vol. 46, No. 5, September–October 2005, 323–351

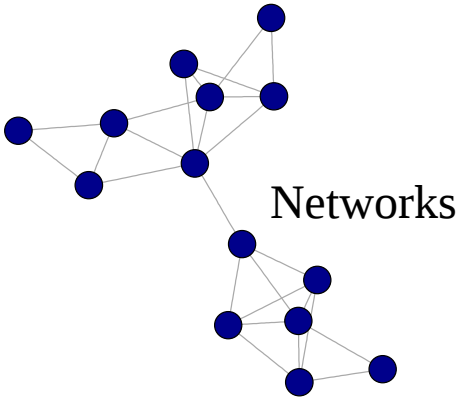
Power laws, Pareto distributions and Zipf's law

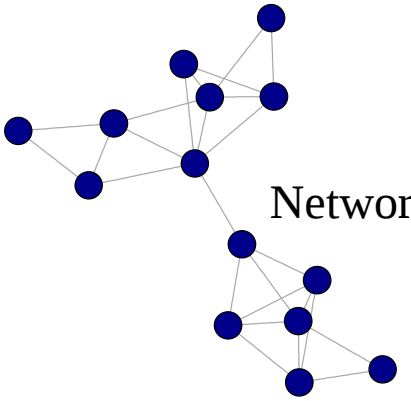
M.E.J. NEWMAN*

Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109, USA



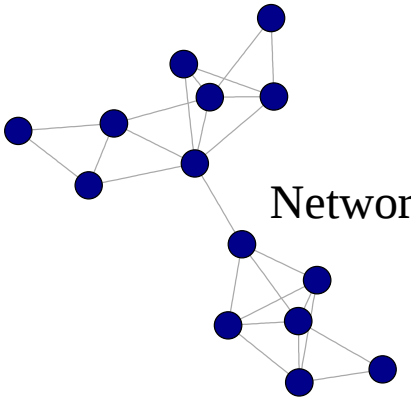
Cumulative distributions or 'rank/frequency plots' of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponents in table 1. Source references for the data are given in the text. (a) Numbers of occurrences of words in the novel Moby Dick by Hermann Melville. (b) Numbers of citations to scientific papers published in 1981, from time of publication until June 1997. (c) Numbers of hits on web sites by 60000 users of the America Online Internet service for the day of 1 December 1997. (d) Numbers of copies of bestselling books sold in the US between 1895 and 1965. (e) Number of calls received by AT&T telephone customers in the US for a single day. (f) Magnitude of earthquakes in California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the maximum amplitude of the earthquake, and hence the distribution obeys a power law even though the horizontal axis is linear. (g) Diameter of craters on the moon. Vertical axis is measured per square kilometre. (h) Peak gamma-ray intensity of solar flares in counts per second, measured from Earth orbit between February 1980 and November 1989. (i) Intensity of wars from 1816 to 1980, measured as battle deaths per 10000 of the population of the participating countries. (j) Aggregate net worth in dollars of the richest individuals in the US in October 2003. (k) Frequency of occurrence of family names in the US in the year 1990. (l) Populations of US cities in the year 2000.


Networks



Networks

Why my friends have more friends than I do?

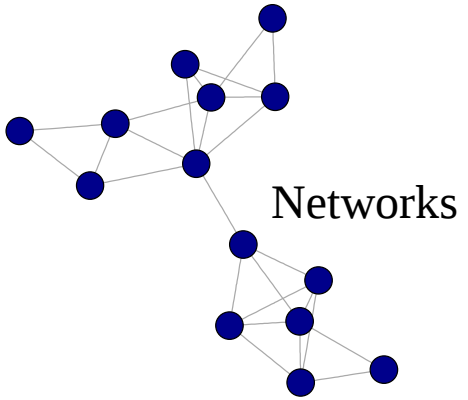


Hands on social network data!

git clone git@github.com:edugalt/ccc16.git

https://github.com/edugalt/ccc16/archive/master.zip

Solving the friendship paradox



$$N \equiv \text{Number of nodes}$$

$$L \equiv \text{ Number of links}$$

$$z_i \equiv \text{ degree of node } i$$

$$P(z) \equiv \frac{1}{N}\sum_i \delta(z - z_i) = \text{ degree distribution}$$

Average degree of nodes:
(your expected degree)

$$\langle z \rangle = \sum_i P(z) z_i = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{2L}{N}$$

Probability of finding node *i* via neighbour (a link):

$$w_i = \frac{z_i}{\sum_i z_i} = \frac{z_i}{2L}$$

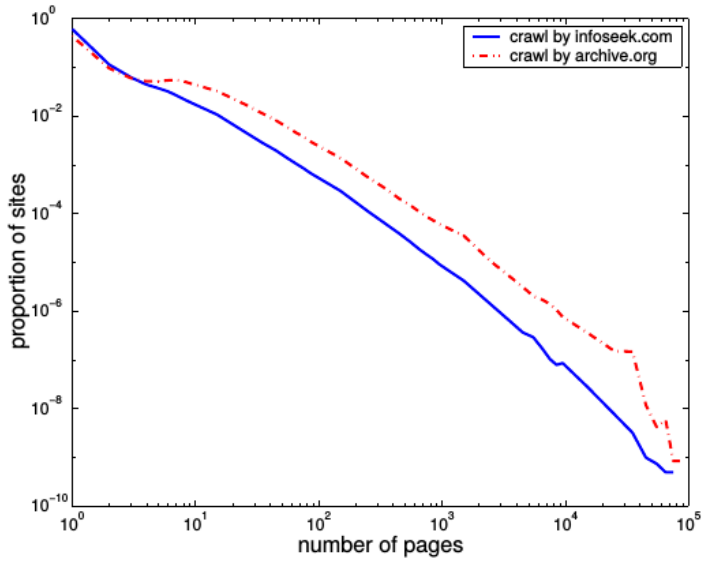
Average degree of your friends:
(your expected degree)

$$\bar{z} = \sum_i w_i z_i = \frac{1}{2L}\sum_{i=1}^{N} z_i^2 = \frac{\langle z^2 \rangle}{\langle z \rangle} = \langle z \rangle + \frac{\sigma_z^2}{\langle z \rangle}$$

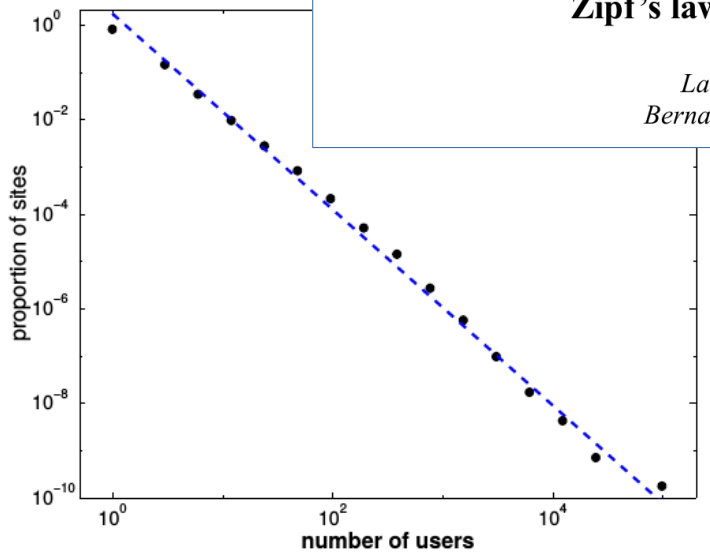
Glottometrics 3, 2002,143-150

Zipf's law and the Internet

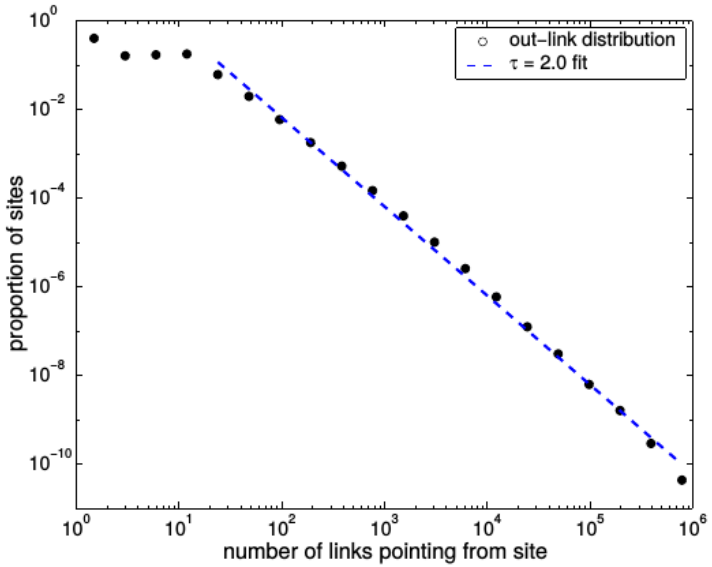
Lada A. Adamic[1]
Bernardo A. Huberman



c)



d)



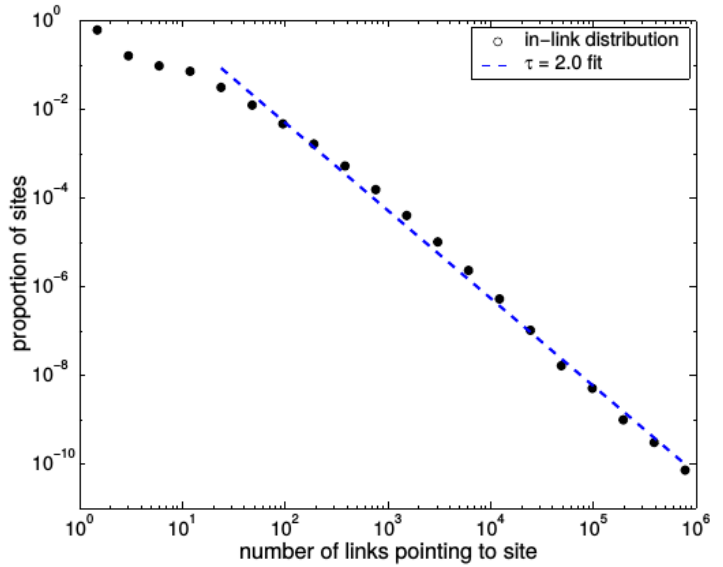


Figure 1. Fitted power law distributions of the number of site a) pages, b) visitors, c) out links, and d) in links, measured in 1997.

Language

How many words exist?

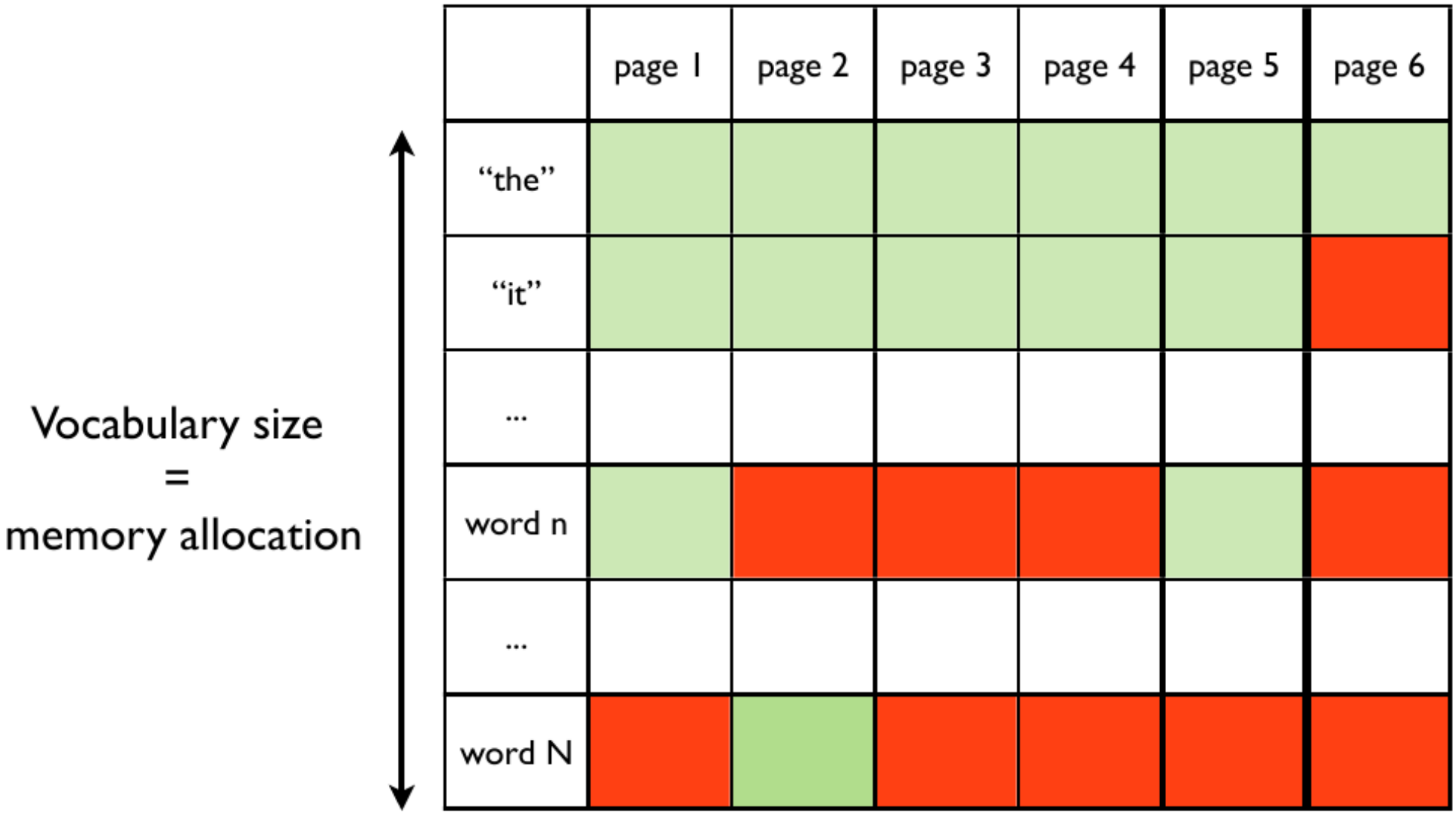
Report on the state of the German language (March 2013)
German Academy for Language and Literature
Union of German Academies of Sciences and Humanities

Year	1905-1914	1948-1957	1995-2004
# distinct words	3.715.000	5.045.000	5.238.000

Quantitative Analysis of Culture Using Millions of Digitized Books
Michel et. al., Science (2011) [*English*]

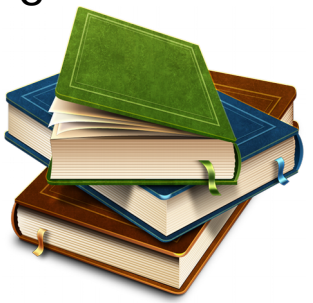
Year	1900	1950	2000
# distinct words	544.000	597.000	1.022.000

Application: invert indexing



Further applications: vocabulary richness of texts / authors
(different document lengths)

Languag
e

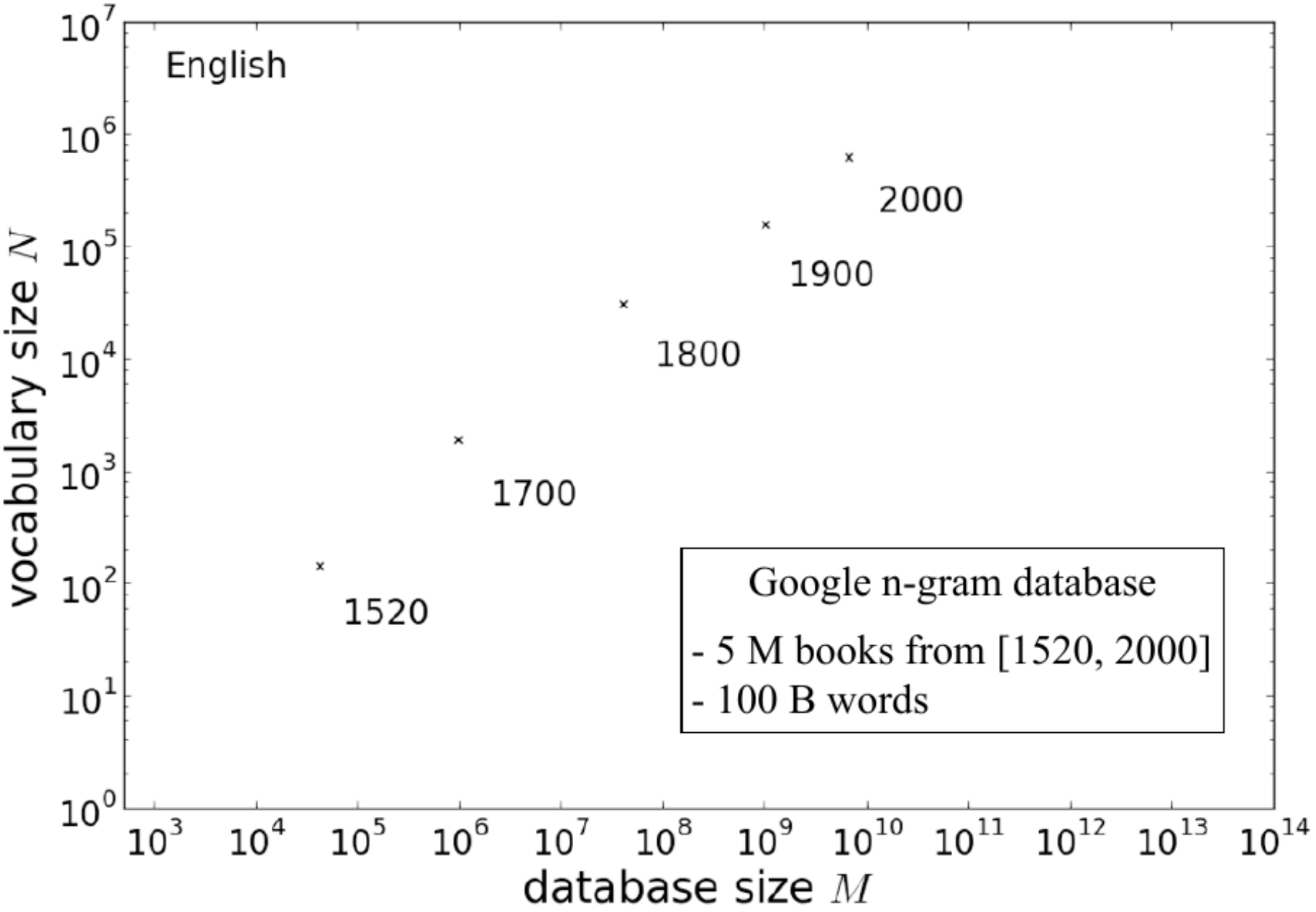


Hands on language!

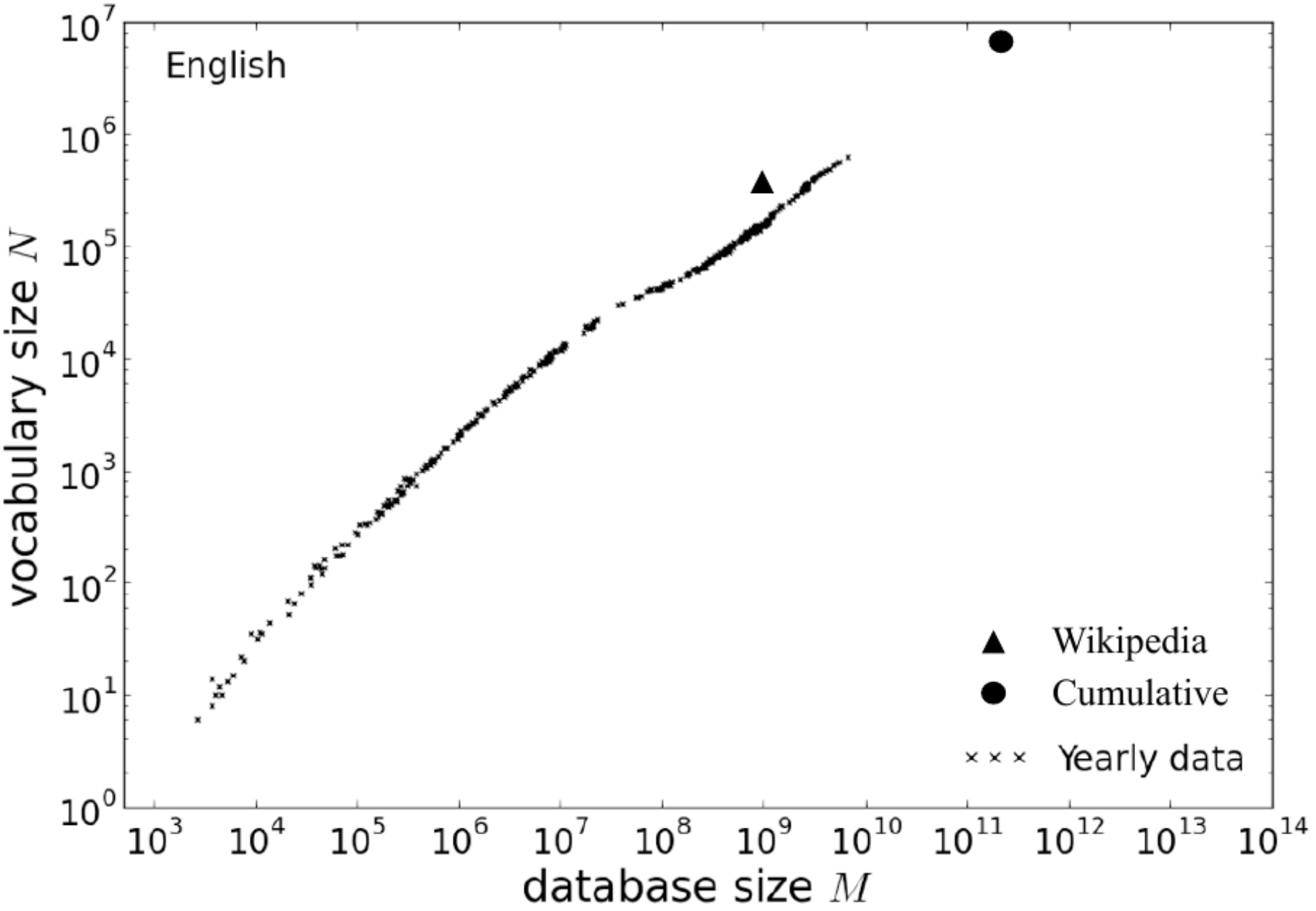
git clone git@github.com:edugalt/ccc16.git

https://github.com/edugalt/ccc16/archive/master.zip

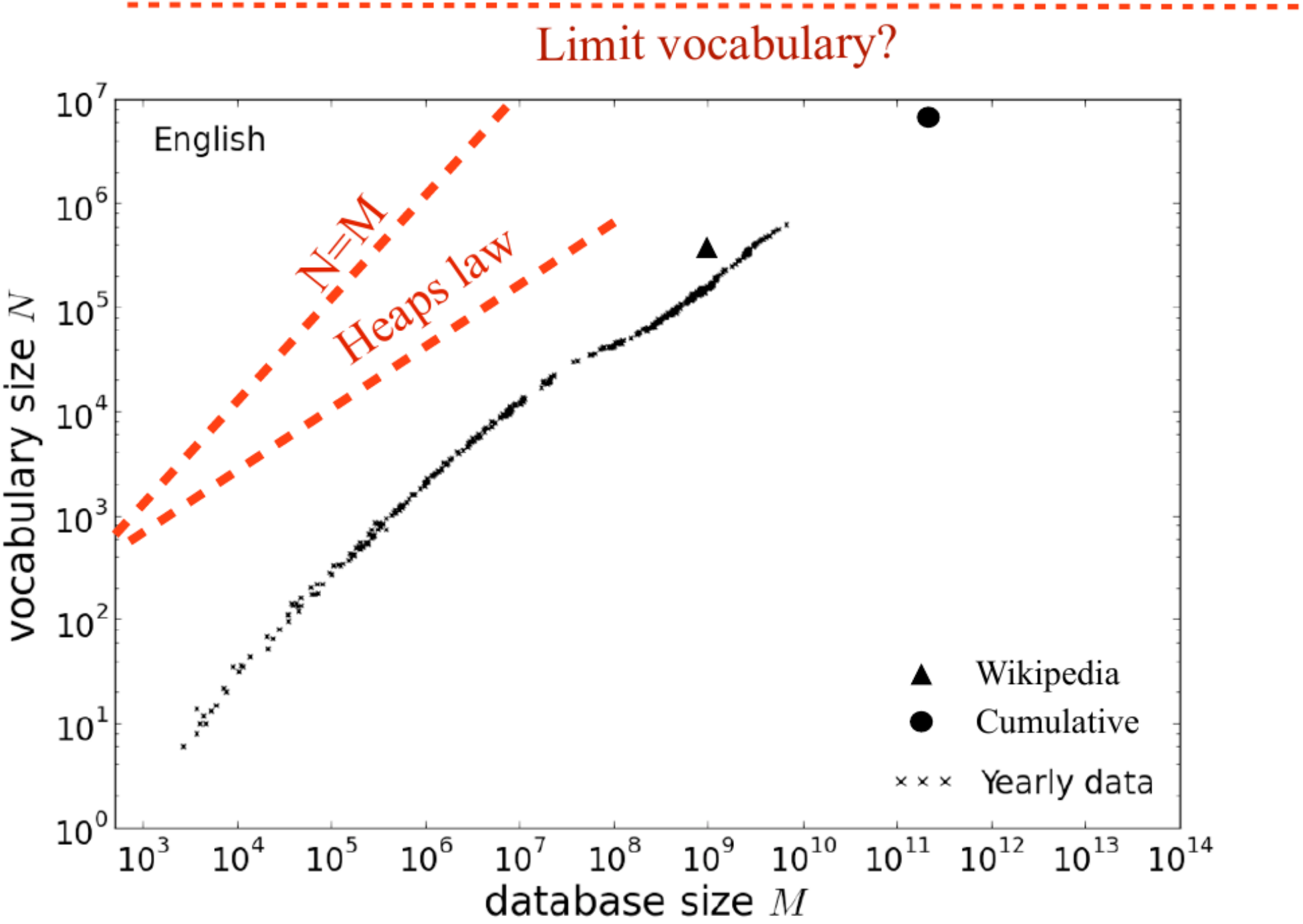
Vocabulary growth with database size



Vocabulary growth with database size

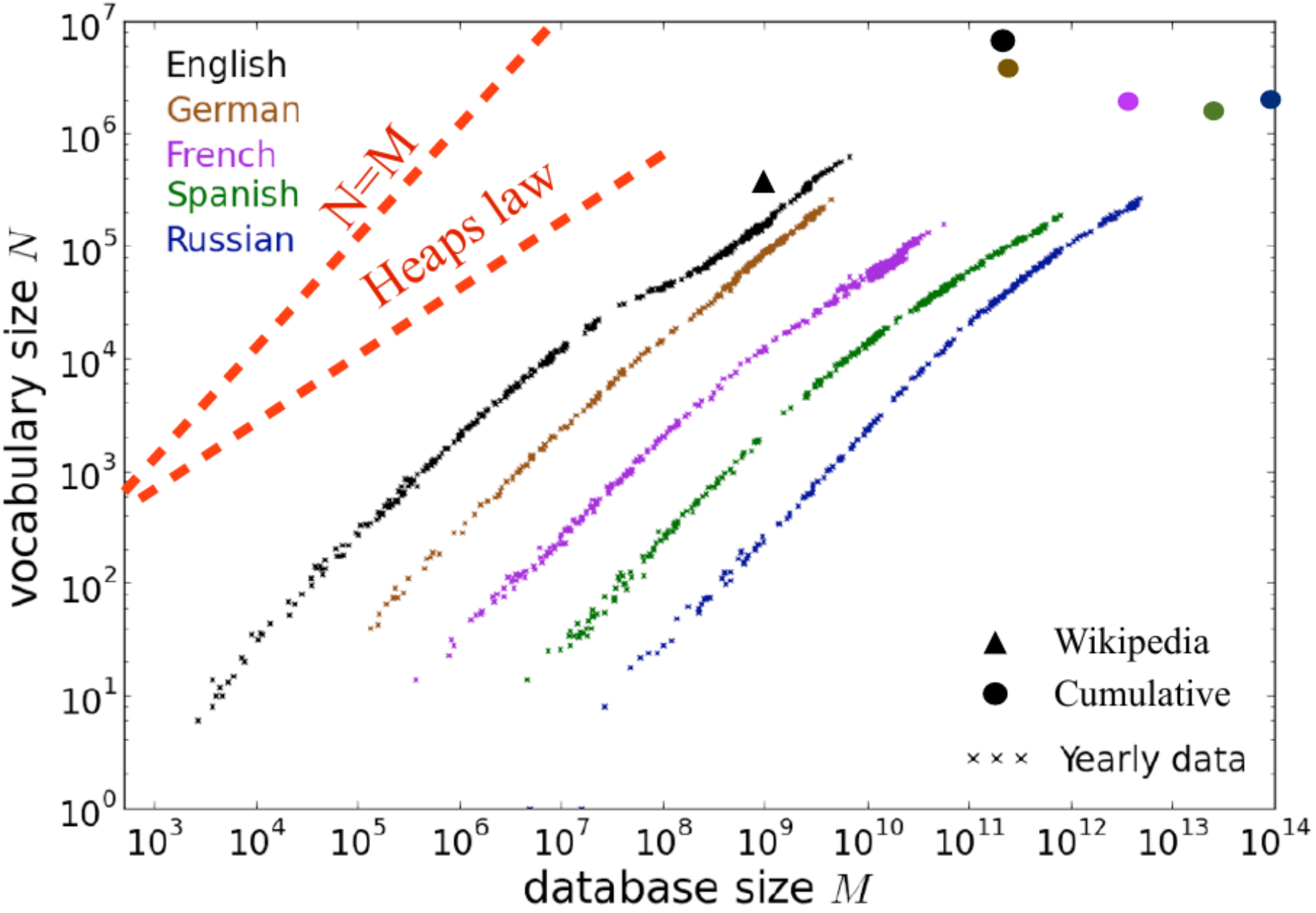


Vocabulary growth with database size

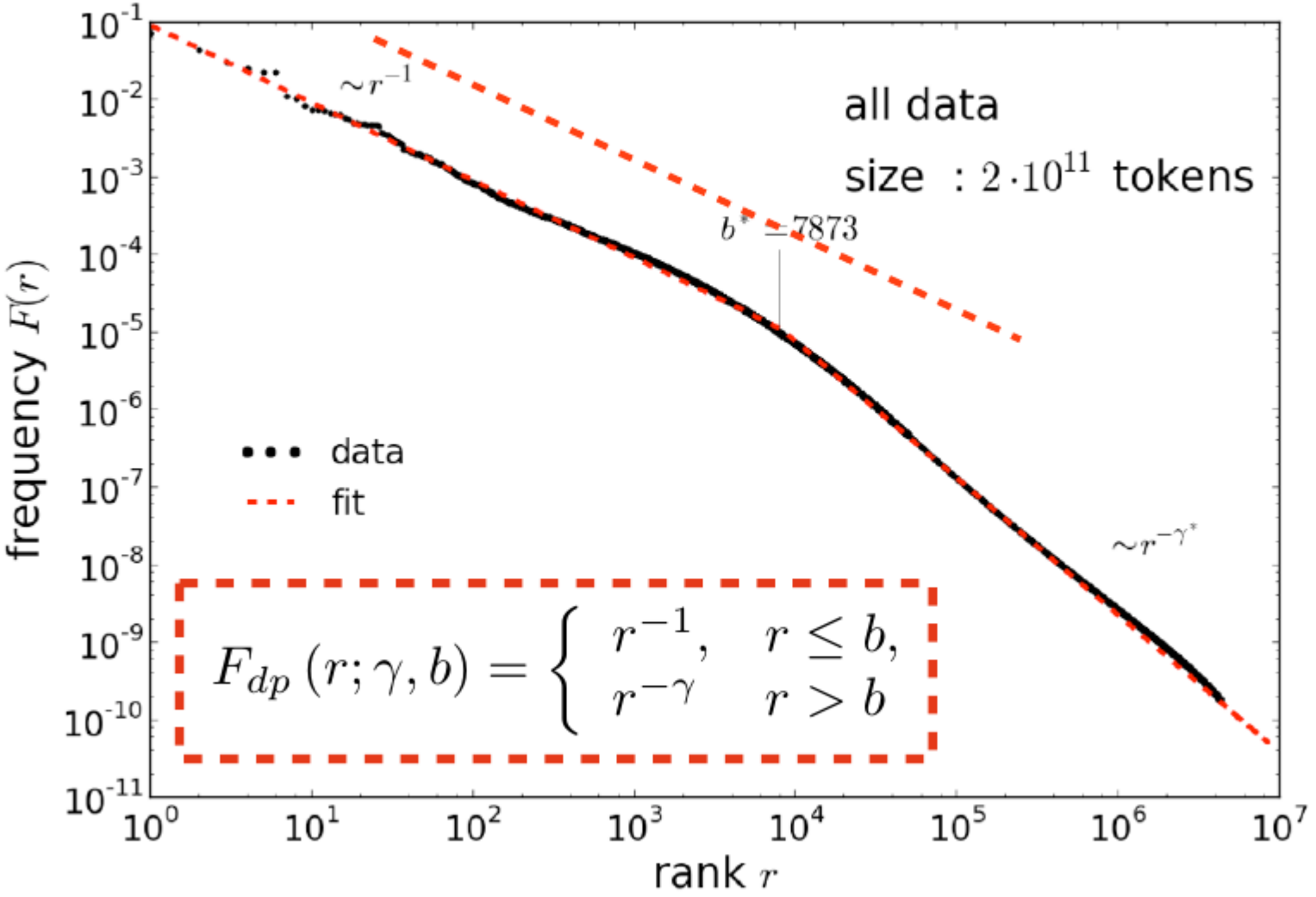


Vocabulary growth with database size

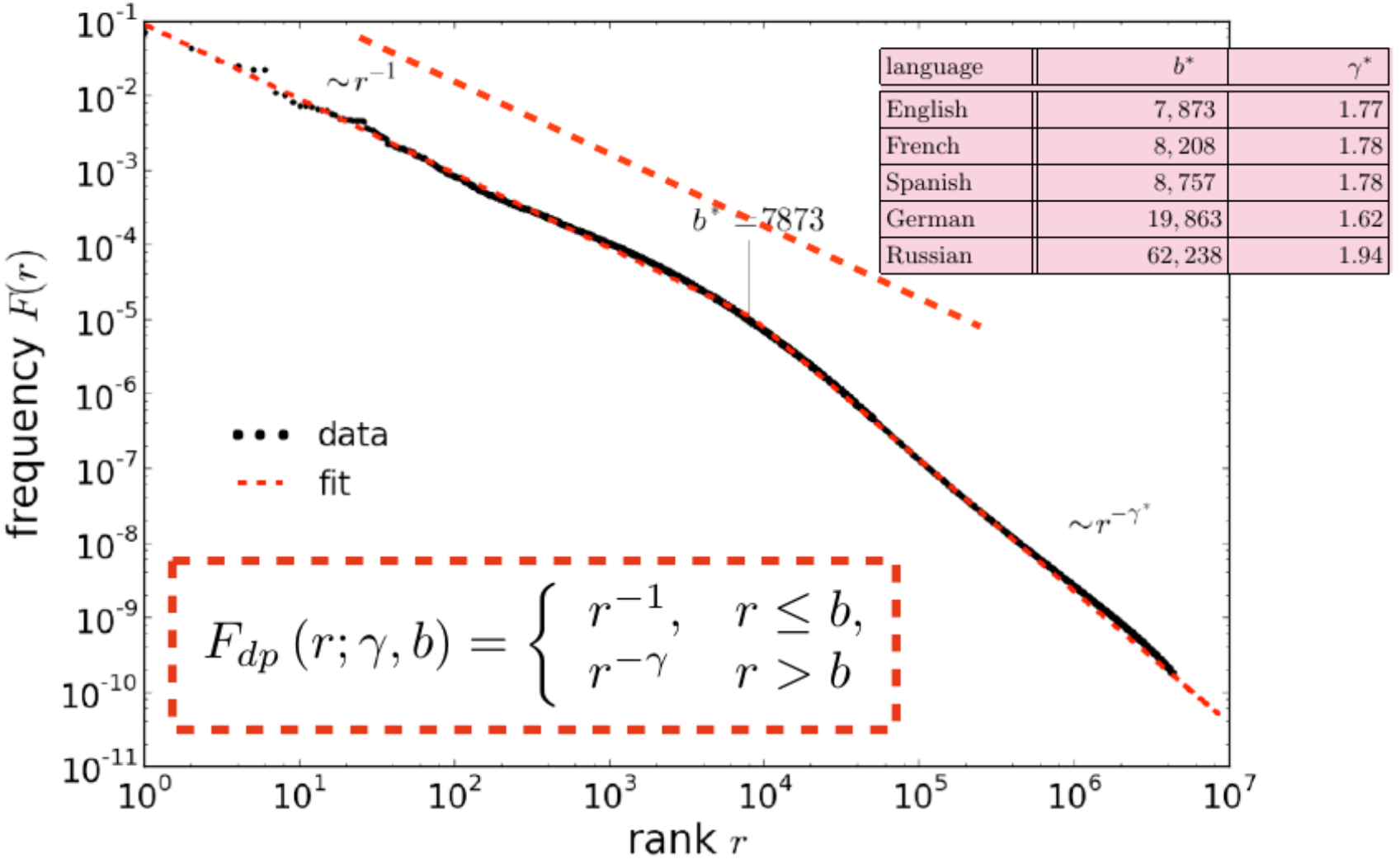
Limit vocabulary?



Zipf's law?


frequency $F(r)$
rank r
$\sim r^{-1}$
$\sim r^{-\gamma^*}$
$b^* = 7873$
all data
size : $2 \cdot 10^{11}$ tokens
data
fit
$F_{dp}(r;\gamma,b) = \begin{cases} r^{-1}, & r \le b, \\ r^{-\gamma} & r > b \end{cases}$

Zipf's law?



language	b^*	γ^*
English	7,873	1.77
French	8,208	1.78
Spanish	8,757	1.78
German	19,863	1.62
Russian	62,238	1.94

Vocabulary growth with database size

Simple mode: usage of each word follows a Poisson process with fixed frequency

$$\langle N(M) \rangle = \sum_r 1 - e^{-F(r)M}$$

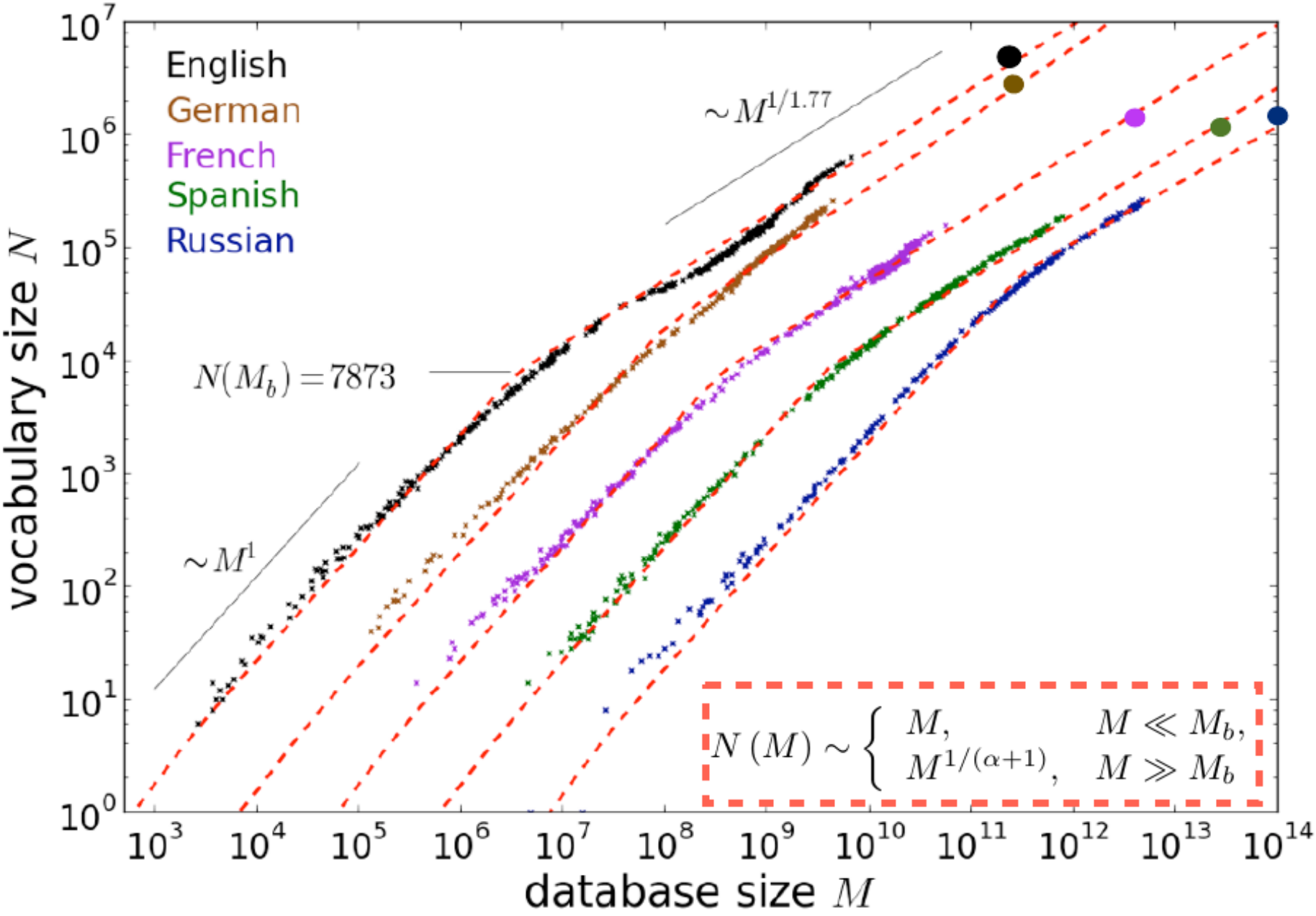
where F(r) is the frequency of the r-th most frequent word (r = rank).

$$F_{dp}(r;\gamma,b) = \begin{cases} r^{-1}, & r \leq b, \\ r^{-\gamma} & r > b \end{cases}$$

$$N_{dp}(N_c) = \begin{cases} M, & M \ll M_b, \\ M^{1/\gamma}, & M \gg M_b \end{cases}$$

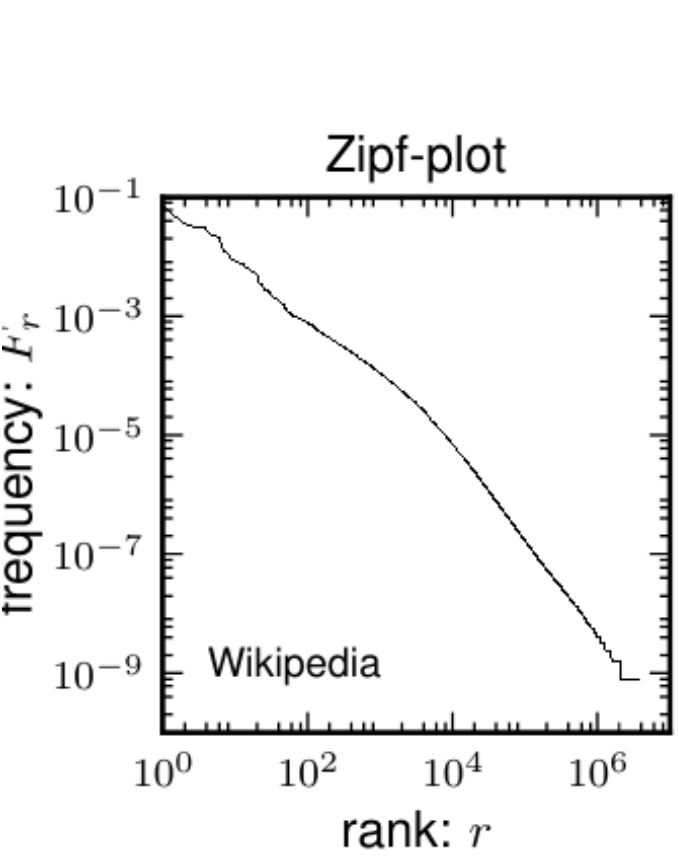
Extension of the Zipf-Heaps connection [<Mandelbrot 1950's]!

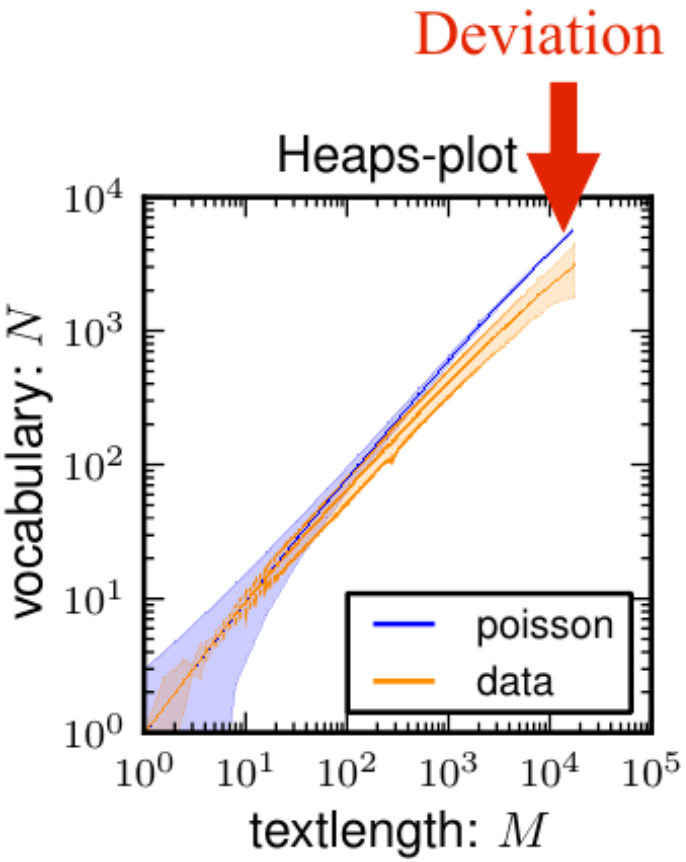
Vocabulary growth with database size

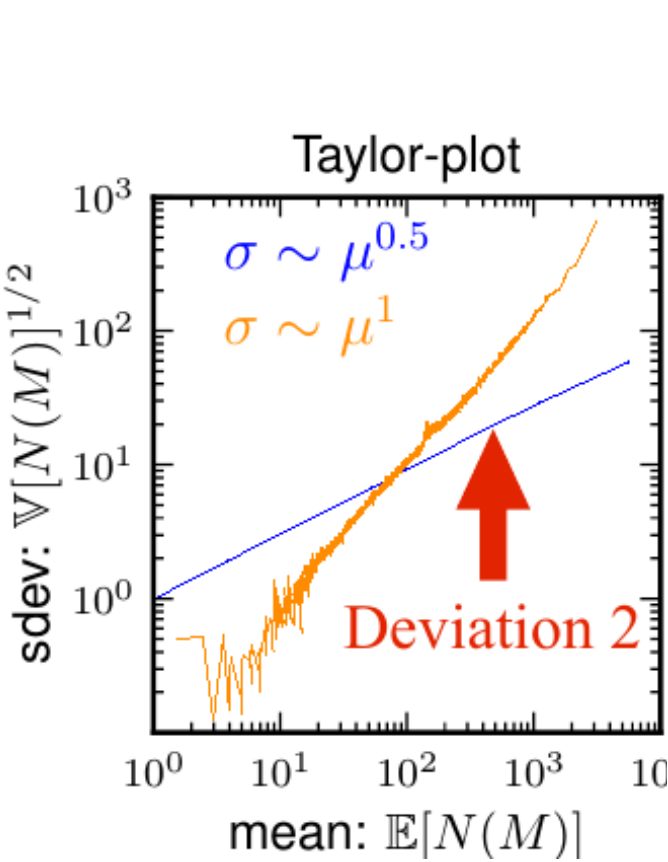


Gerlach and Altmann, PRX 2013

Fluctuations







Gerlach and Altmann, New J. Phys. 2014

Fluctuations

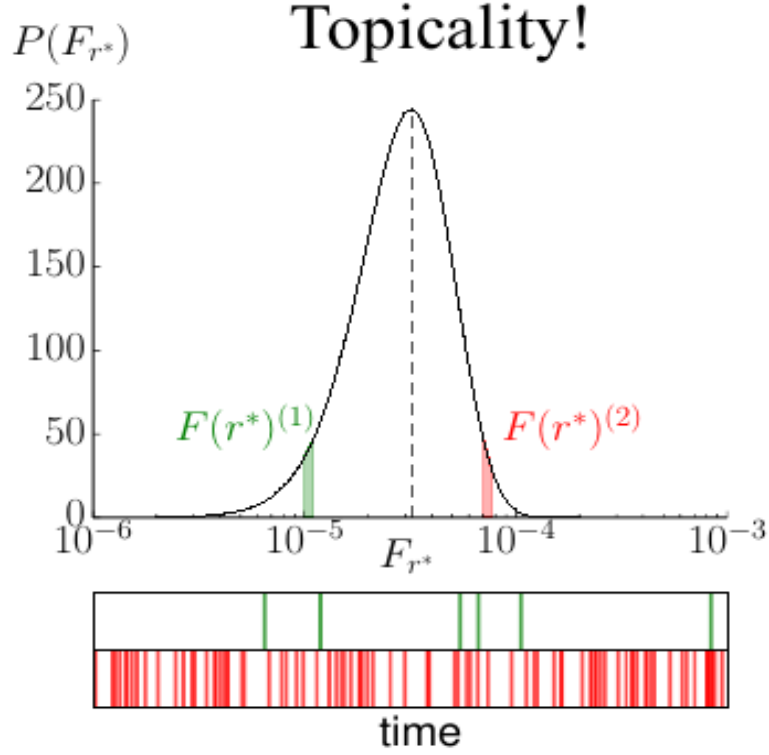
$$\mathbb{E}_q\left[N(M)\right] = \langle N(M)^{(i,j)}\rangle_{i,j} = \sum 1 - \langle e^{-MF_r}\rangle$$
$$\leq \sum_r 1 - e^{-M\langle F_r\rangle} \equiv \mathbb{E}_a\left[N(M)\right].$$

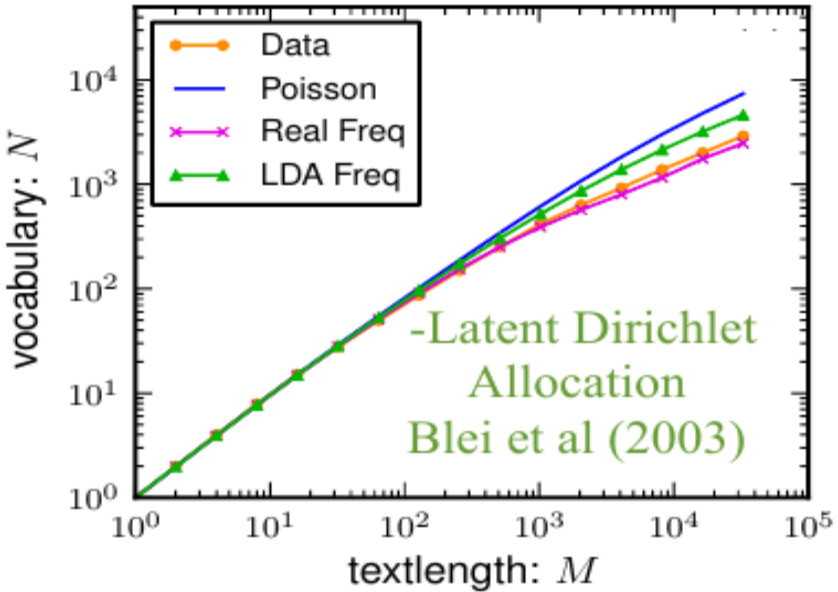
Explains deviation 1

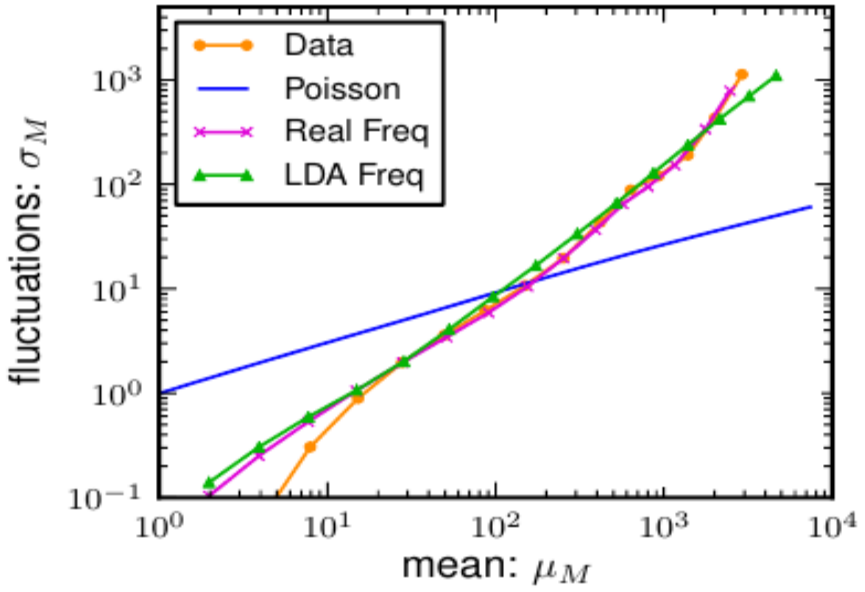
$$\mathbb{V}_q\left[N(M)\right] \equiv \mathbb{E}_q\left[N(M)^2\right] - \mathbb{E}_q\left[N(M)\right]^2$$
$$= \sum_r \langle e^{-MF_r}\rangle - \langle e^{-MF_r}\rangle^2$$
$$+ \sum_r \sum_{r' \neq r} \mathrm{Cov}[e^{-MF_r}, e^{-MF_{r'}}]$$

~ M^2, Explains deviation 2

Topicality!







Summary

Fat-tailed distributions...

1. … have peculiar statistical properties
   - no characteristic scale
   - moments diverge
   - statistics dominated either by the few most frequent or by the many rare items

2. … appear in different social and natural datasets.
   - extreme events: earthquakes and avalanches
   - social inequality: income, attention
   - social networks: www, citations to papers
   - economy: stock market, sales, popularity
   - city sizes, words, etc.

3 … are important in practice!
   - can lead to high risk/impact situations
   - prediction becomes harder!
   - estimation of vocabulary sizes show sub-linear scaling with sample size

Outlook: why so many fat-tailed distributions?

References:

Power laws, Pareto distributions and Zipf's law, *Newman* (2005)

A Brief History of Generative Models for Power Law and Lognormal Distributions, *Mitzenmacher* (2004)

1. Central limit theorem

2. Monkey Typewritter:

3. Proportional growth or preferential attachment