INFO 2950
Intro to Data Science

Lecture 17: Power Laws and Big Data

Paul Ginsparg

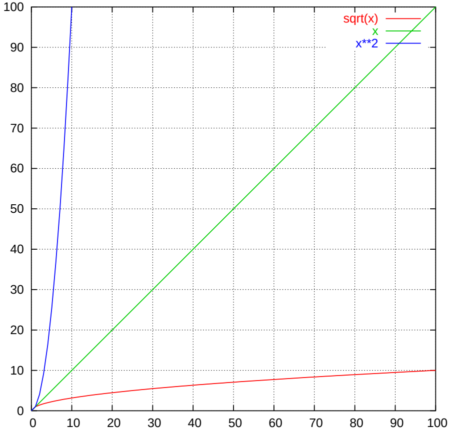
Cornell University, Ithaca, NY

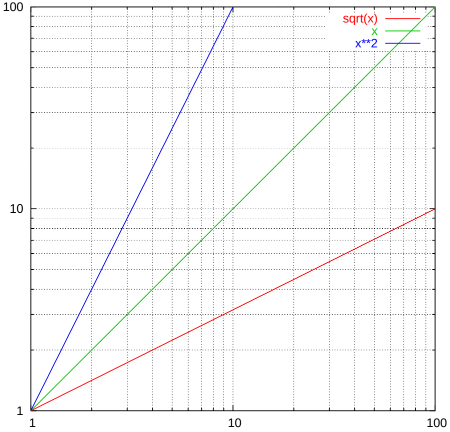
29 Oct 2013

Power Laws in log-log space

$y = cx^k$ (k=1/2,1,2)

$\log_{10} y = k * \log_{10} x + \log_{10} c$





Zipf's law

- Now we have characterized the growth of the vocabulary in collections.
- We also want to know how many frequent vs. infrequent terms we should expect in a collection.
- In natural language, there are a few very frequent terms and very many very rare terms.
- Zipf's law (linguist/philologist George Zipf, 1935):
  The i^{th} most frequent term has frequency proportional to $1/i$.
- $\mathrm{cf}_i \propto \frac{1}{i}$
- cf_i is collection frequency: the number of occurrences of the term t_i in the collection.

http://en.wikipedia.org/wiki/Zipf's_law

Zipf's law: the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. Brown Corpus:

- "the": 7% of all word occurrences (69,971 of $\gtrsim$1M).
- "of": $\sim$3.5% of words (36,411)
- "and": 2.9% (28,852)

Only 135 vocabulary items account for half the Brown Corpus.

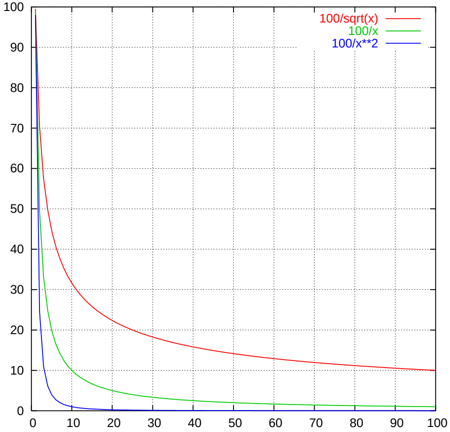
The Brown University Standard Corpus of Present-Day American English is a carefully compiled selection of current American English, totaling about a million words drawn from a wide variety of sources . . . for many years among the most-cited resources in the field.

Zipf's law

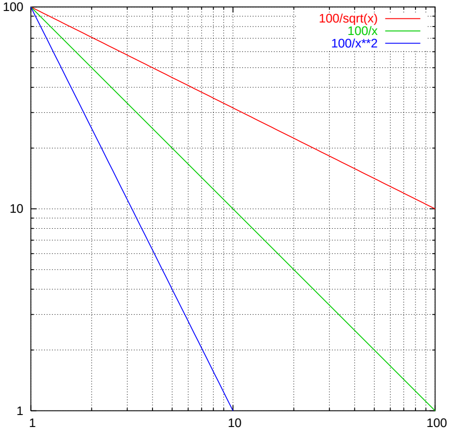
- Zipf's law: The i^{th} most frequent term has frequency proportional to $1/i$.
- $\mathrm{cf}_i \propto \frac{1}{i}$
- cf is collection frequency: the number of occurrences of the term in the collection.
- So if the most frequent term (*the*) occurs cf_1 times, then the second most frequent term (*of*) has half as many occurrences $\mathrm{cf}_2 = \frac{1}{2}\mathrm{cf}_1$ ...
- ...and the third most frequent term (*and*) has a third as many occurrences $\mathrm{cf}_3 = \frac{1}{3}\mathrm{cf}_1$ etc.
- Equivalent: $\mathrm{cf}_i = ci^k$ and $\log \mathrm{cf}_i = \log c + k \log i$ (for $k = -1$)
- Example of a power law

Power Laws in log-log space

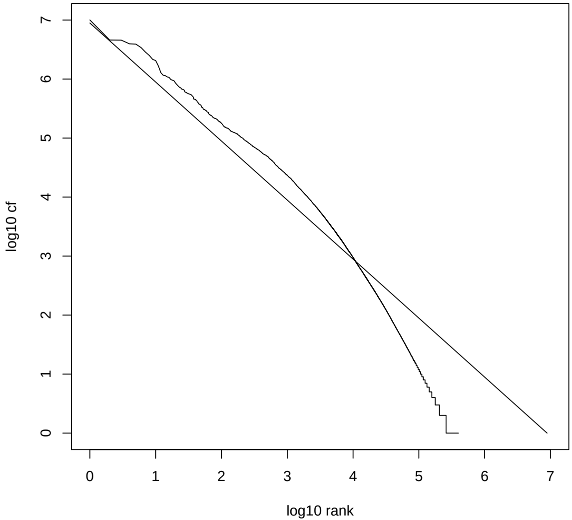
$y = cx^{-k}$ (k=1/2,1,2)



$\log_{10} y = -k * \log_{10} x + \log_{10} c$



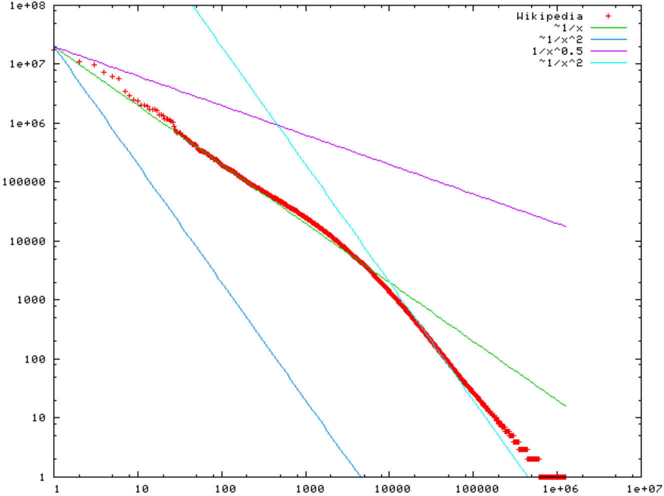
Zipf's law for Reuters



Fit far from perfect, but nonetheless key insight:

Few frequent terms, many rare terms.

more from http://en.wikipedia.org/wiki/Zipf's_law



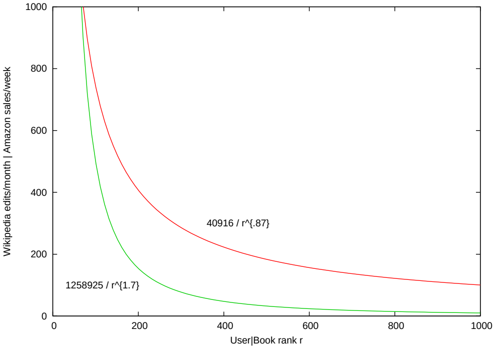
"A plot of word frequency in Wikipedia (27 Nov 2006). The plot is in log-log coordinates. x is rank of a word in the frequency table; y is the total number of the words occurrences. Most popular words are "the", "of" and "and", as expected. Zipf's law corresponds to the upper linear portion of the curve, roughly following the green (1/x) line."

Power laws more generally

E.g., consider power law distributions of the form $c\,r^{-k}$, describing the number of book sales versus sales-rank r of a book, or the number of Wikipedia edits made by the r^{th} most frequent contributor to Wikipedia.

- Amazon book sales: $c\,r^{-k}$, $k \approx .87$
- number of Wikipedia edits: $c\,r^{-k}$, $k \approx 1.7$

(More on power laws and the long tail here:
Networks, Crowds, and Markets:
Reasoning About a Highly Connected World
by David Easley and Jon Kleinberg
Chpt 18: `http://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch18.pdf`)

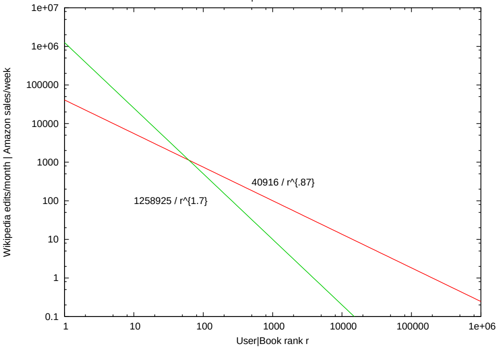


Normalization given by the roughly 1 sale/week for the 200,000th ranked Amazon title:

$$40916r^{-.87}$$

and by the 10 edits/month for the 1000th ranked Wikipedia editor:

$$1258925r^{-1.7}$$



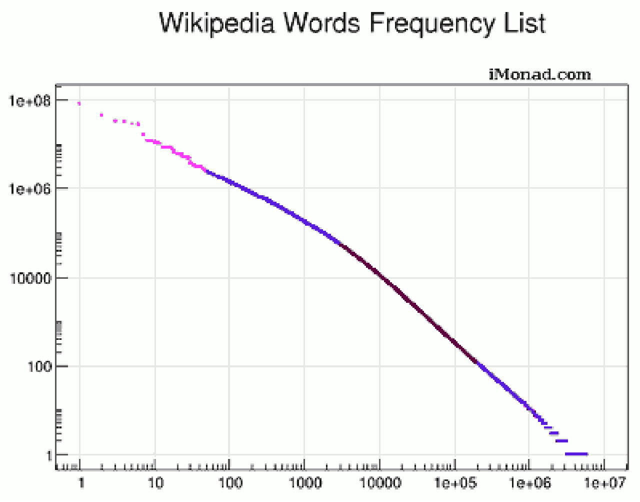
Long tail: about a quarter of Amazon book sales estimated to come from the long tail, i.e., those outside the top 100,000 bestselling titles

Another Wikipedia count (15 May 2010)

`http://imonad.com/seo/wikipedia-word-frequency-list/`

All articles in the English version of Wikipedia, 21GB in XML format (five hours to parse entire file, extract data from markup language, filter numbers, special characters, extract statistics):

- Total tokens (words, no numbers): $T = 1{,}570{,}455{,}731$
- Unique tokens (words, no numbers): $M = 5{,}800{,}280$



"Word frequency distribution follows Zipf's law"

- rank 1–50 (86M-3M), stop words (the, of, and, in, to, a, is, ...)
- rank 51–3K (2.4M-56K), frequent words (university, January, tea, sharp, ...)
- rank 3K–200K (56K-118), words from large comprehensive dictionaries (officiates, polytonality, neologism, ...)
  above rank 50K mostly Long Tail words
- rank 200K–5.8M (117-1), terms from obscure niches, misspelled words, transliterated words from other languages, new words and non-words (euprosthenops, eurotrochilus, lokottaravada, ...)

Some selected words and associated counts

- Google 197920
- Twitter 894
- domain 111850
- domainer 22
- Wikipedia 3226237
- Wiki 176827
- Obama 22941
- Oprah 3885
- Moniker 4974
- GoDaddy 228

Project Gutenberg (per billion)

http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#Project_Gutenberg

Over 36,000 items (Jun 2011), average of $>$ 50 new e-books / week

http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/PG/2006/04/1-10000

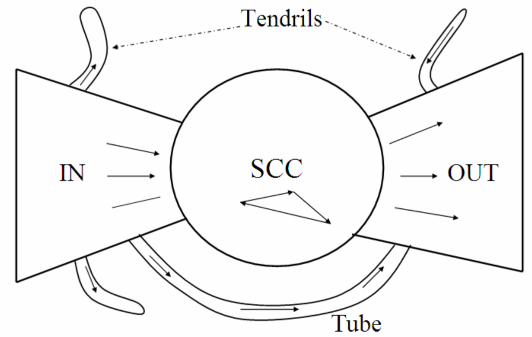
- the 56271872
- of 33950064
- and 29944184
- to 25956096
- in 17420636
- I 11764797
- that 11073318
- was 10078245
- his 8799755
- he 8397205

- it 8058110
- with 7725512
- is 7557477
- for 7097981
- as 7037543
- had 6139336
- you 6048903
- not 5741803
- be 5662527
- her 5202501

$\ldots 100,000^{\text{th}}$

Bowtie structure of the web

A.Broder,R.Kumar,F.Maghoul,P.Raghavan,S.Rajagopalan,S. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. Computer Networks, 33:309–320, 2000.



- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don't link (OUT)
- Lots of pages that link to other pages, but don't get linked to (IN)
- Tendrils, tubes, islands

# of in-links (in-degree) averages 8–15, not randomly distributed (Poissonian), instead a power law:

pages with in-degree i is $\propto 1/i^{\alpha}$, $\alpha \approx 2.1$

Poisson Distribution

Bernoulli process with N trials, each probability p of success:

$$p(m) = \binom{N}{m} p^m (1-p)^{N-m} \ .$$

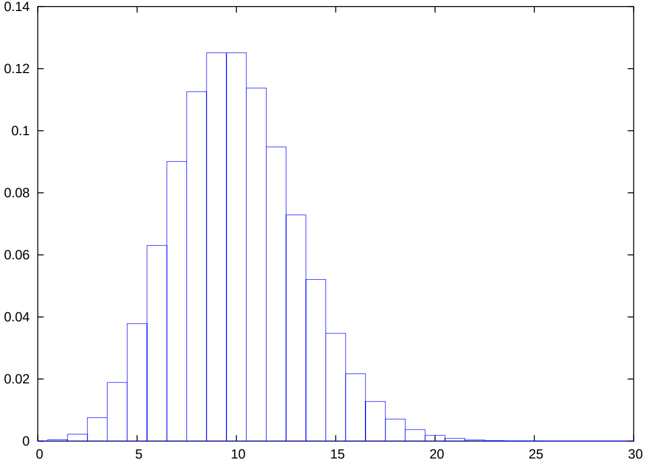
Probability $p(m)$ of m successes, in limit N very large and p small, parametrized by just $\mu = Np$ ($\mu =$ mean number of successes).
For $N \gg m$, we have $\frac{N!}{(N-m)!} = N(N-1)\cdots(N-m+1) \approx N^m$, so $\binom{N}{m} \equiv \frac{N!}{m!(N-m)!} \approx \frac{N^m}{m!}$, and

$$p(m) \approx \frac{1}{m!} N^m \left(\frac{\mu}{N}\right)^m \left(1 - \frac{\mu}{N}\right)^{N-m} \approx \frac{\mu^m}{m!} \lim_{N\to\infty} \left(1 - \frac{\mu}{N}\right)^N = \mathrm{e}^{-\mu} \frac{\mu^m}{m!}$$

(ignore $(1-\mu/N)^{-m}$ since by assumption $N \gg \mu m$).
N dependence drops out for $N \to \infty$, with average μ fixed ($p \to 0$).
The form $p(m) = \mathrm{e}^{-\mu} \frac{\mu^m}{m!}$ is known as a Poisson distribution
(properly normalized: $\sum_{m=0}^{\infty} p(m) = \mathrm{e}^{-\mu} \sum_{m=0}^{\infty} \frac{\mu^m}{m!} = \mathrm{e}^{-\mu} \cdot \mathrm{e}^{\mu} = 1$).

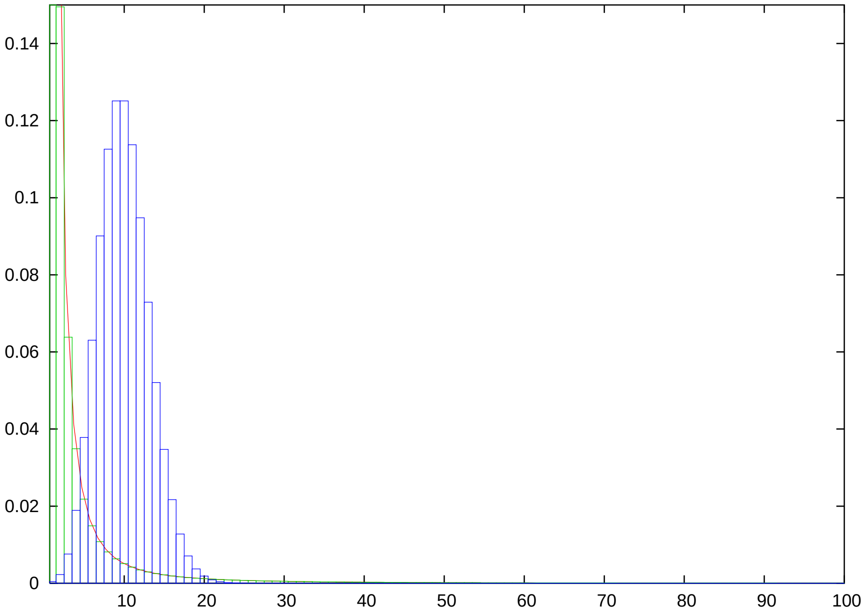
Poisson Distribution for $\mu = 10$

$p(m) = \mathrm{e}^{-10} \frac{10^m}{m!}$

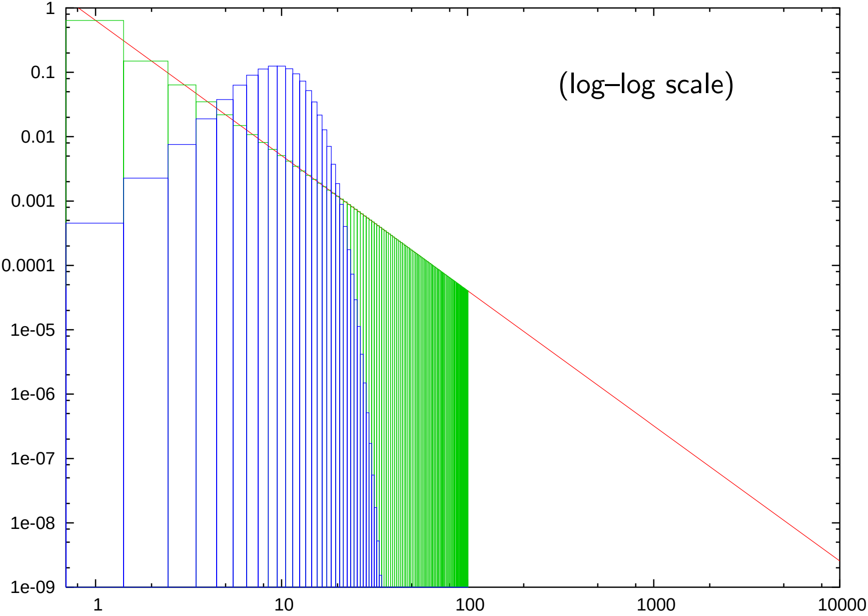


Compare to power law $p(m) \propto 1/m^{2.1}$

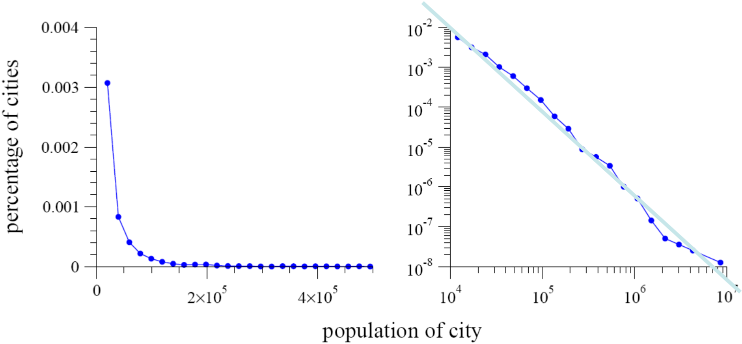
Power Law $p(m) \propto 1/m^{2.1}$ and Poisson $p(m) = \mathrm{e}^{-10}\frac{10^m}{m!}$



Power Law $p(m) \propto 1/m^{2.1}$ and Poisson $p(m) = e^{-10}\frac{10^m}{m!}$



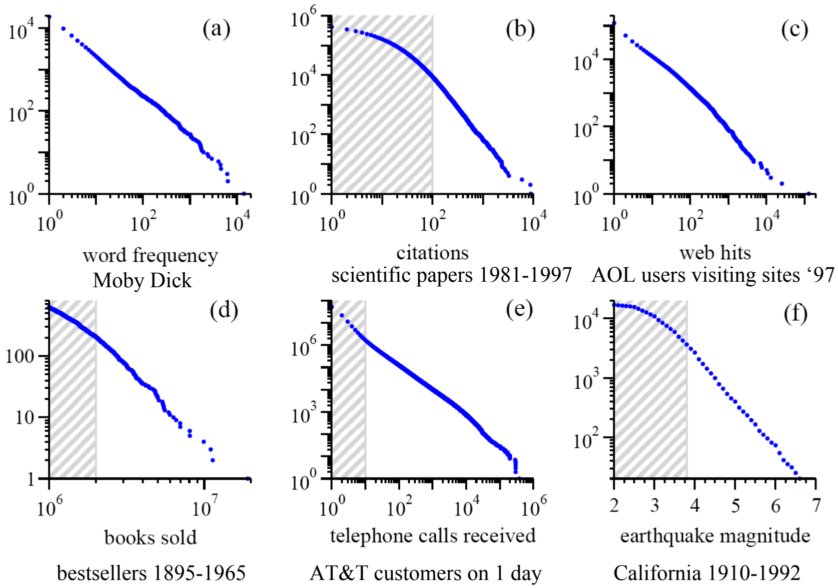
Power law distributions

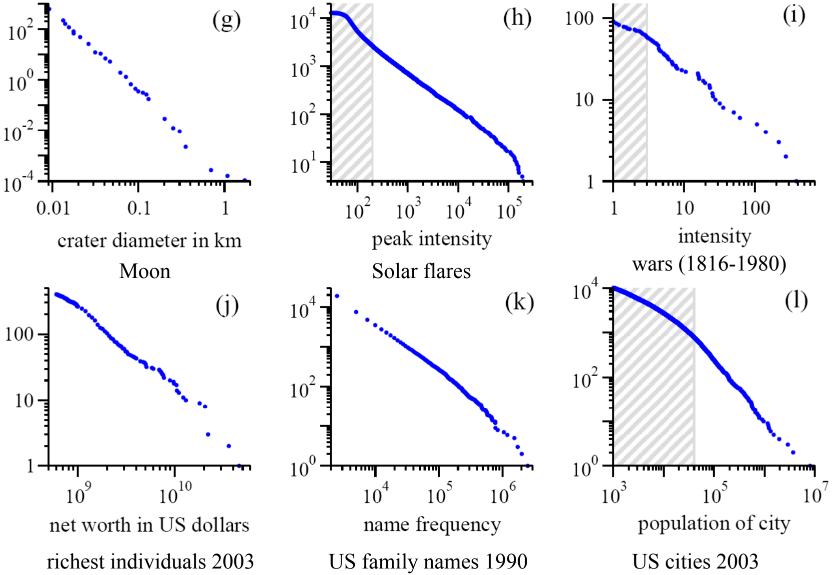


$$f(x) = ax^k + o(x^k),$$

$$\log(f(x)) = k \log x + \log a.$$

Slide credit: Dragomir Radev


(a)
(b)
(c)
(d)
(e)
(f)
word frequency
Moby Dick
citations
scientific papers 1981-1997
web hits
AOL users visiting sites ‘97
books sold
bestsellers 1895-1965
telephone calls received
AT&T customers on 1 day
earthquake magnitude
California 1910-1992


(g)
crater diameter in km
Moon
(h)
peak intensity
Solar flares
(i)
intensity
wars (1816-1980)
(j)
net worth in US dollars
richest individuals 2003
(k)
name frequency
US family names 1990
(l)
population of city
US cities 2003

Power law in networks

- For many interesting graphs, the distribution over node degree follows a 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

Slide credit: Dragomir Radev

Next Time: More Statistical Methods

Peter Norvig, “How to Write a Spelling Corrector”
`http://norvig.com/spell-correct.html`

(See video:
`http://www.youtube.com/watch?v=yvDCzhbjYWs`
“The Unreasonable Effectiveness of Data”, given 23 Sep 2010.)

Additional related references:
`http://doi.ieeecomputersociety.org/10.1109/MIS.2009.36`
A. Halevy, P. Norvig, F. Pereira,
The Unreasonable Effectiveness of Data,
Intelligent Systems Mar/Apr 2009 (copy at resources/unrealdata.pdf)

`http://norvig.com/ngrams/ch14.pdf`
P. Norvig, “Natural Language Corpus Data”