Zipf’s Law

George Kingsley Zipf
1902-1950

- Frequency of occurrence of words is inversely proportional to the rank in this frequency of occurrence.
- When both are plotted on a log scale, the graph is a straight line.

Zipf Distribution

- The Important Points:
  - a few elements occur *very frequently*
  - a medium number of elements have medium frequency
  - many elements occur *very infrequently*

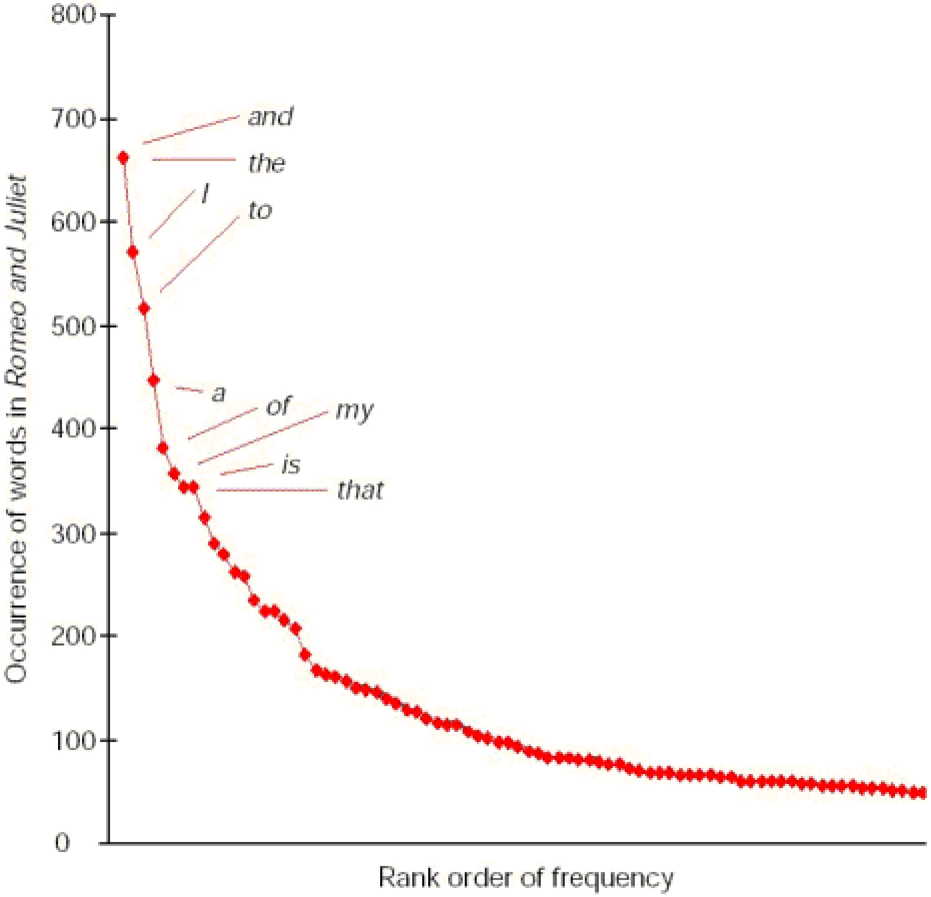
Zipf Distribution

The product of the frequency of words (f) and their rank (r) is approximately constant

Rank = order of words' frequency of occurrence

$$f = C * 1/r$$

$$C \cong N/10$$



Zipf Distribution (Same curve on linear and log scale)

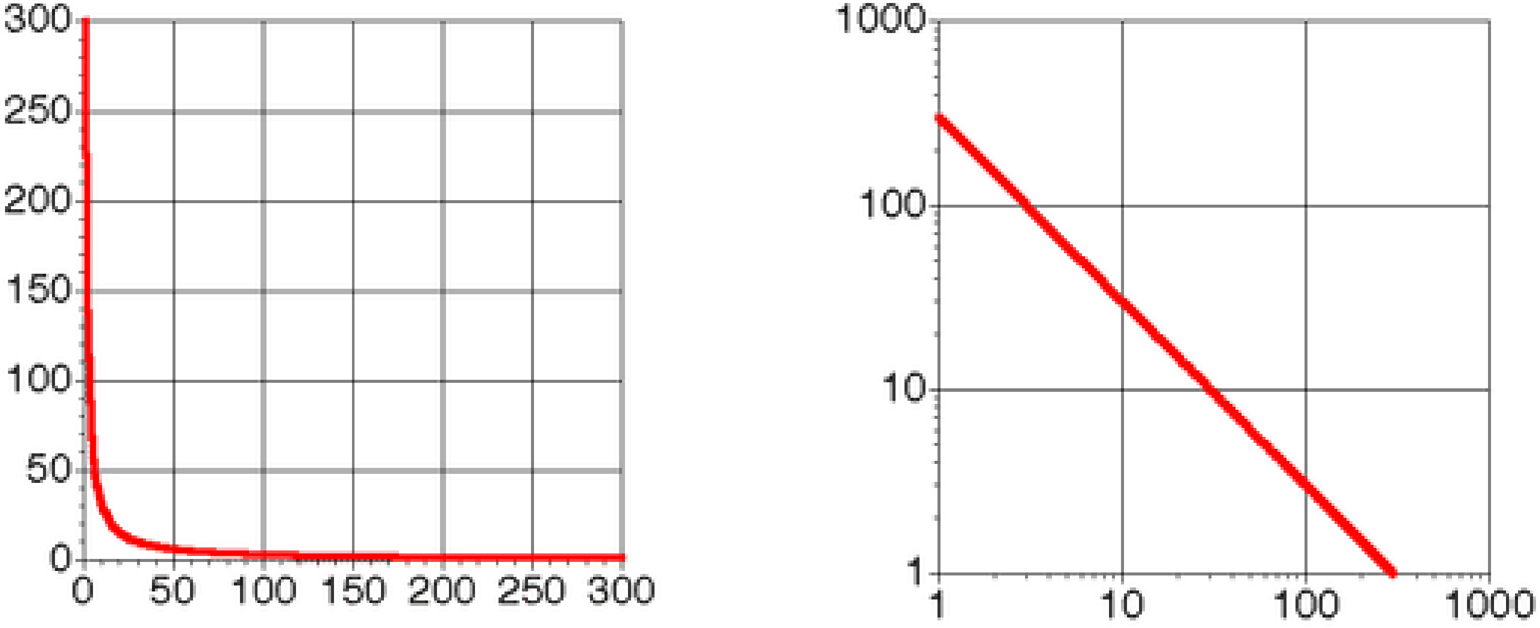


Illustration by Jacob Nielsen

What Kinds of Data Exhibit a Zipf Distribution?

- Words in a text collection
  - Virtually any language usage
- Library book checkout patterns
- Incoming Web Page Requests (Nielsen)
- Outgoing Web Page Requests (Cunha & Crovella)
- Document Size on Web (Cunha & Crovella)

Characteristics of WWW Client-based Traces

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Zipf's Law Applied To WWW Documents

Distribution of users among web sites

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Binned distribution of users to sites

Exponentially increasing bins

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Cumulative distribution of users to sites

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Communautés virtuelles: une analyse expérimentale du réseau Peer-to-Peer Gnutella

Jean Vaucher

Informatique, Université de Montréal

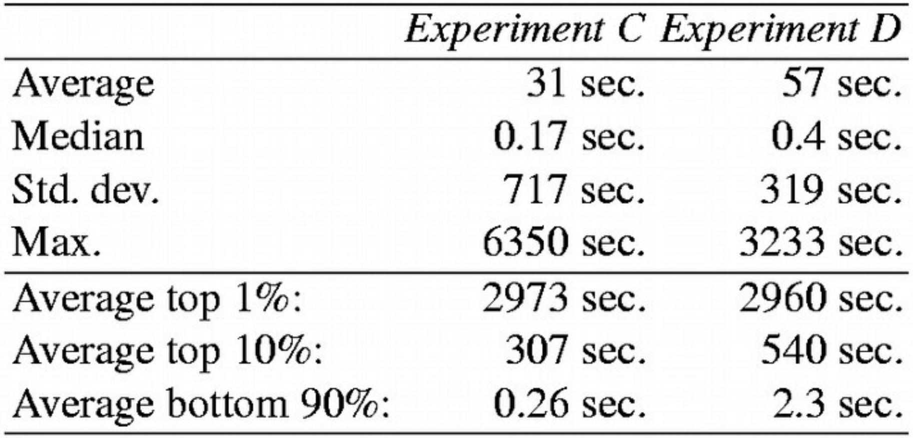
Ref.: J. Vaucher, G. Babin, P. Kropf, Th. Jouve: *Experimenting with Gnutella Communities*, Distributed Communities on the Web, Sydney, April 2002, LNCS 2468, Springer Berlin, pp. 85-99

Durée des connexions

(power law)

Exp C : Nov. 18th, 2001, pendant 24 heures :

- 20954 connexions valides (17735 IN, 3218 OUT)
- session la plus longue : 11 heures ; 5 sessions ont durées plus que 8 heures



	Experiment C	*Experiment D*
Average	31 sec.	57 sec.
Median	0.17 sec.	0.4 sec.
Std. dev.	717 sec.	319 sec.
Max.	6350 sec.	3233 sec.
Average top 1%:	2973 sec.	2960 sec.
Average top 10%:	307 sec.	540 sec.
Average bottom 90%:	0.26 sec.	2.3 sec.

La durée moyenne des connexions est courte (entre 30 et 60 sec), mais il existe des connexions très durable