# The Structure of Language - and More

**Helmut Satz**

**Universität Bielefeld, Germany**

**Karpacz, Poland** **May 20024**

# Critical Behavior and Power Law Distribution

Recall Ising model on 2-D lattice, $s_i = \pm 1$, Hamiltonian

$$H(s) = J \sum_{i,j=1}^{N} s_i s_j.$$

correlation function

$$\Gamma(r) = \sum_{s_1} \sum_{s_2} \ldots \sum_{s_N} |r_i - r_j| \exp -\beta H(s),$$

$r = r_i - r_j$; has Ornstein-Zernicke form

$$\Gamma(r) = \frac{\exp -(r/\lambda)}{r^p},$$

$\lambda(\beta)$: correlation length at temperature $T = 1/\beta$,

$p = 1 - \eta \simeq 1$, anomalous dimension parameter $\eta \simeq 0$.

For $T \to T_c$, correlation length $\lambda$ diverges: $\lambda \sim |T - T_c|^{-\nu}$

$$\Gamma(r, T_c) \simeq 1/r$$

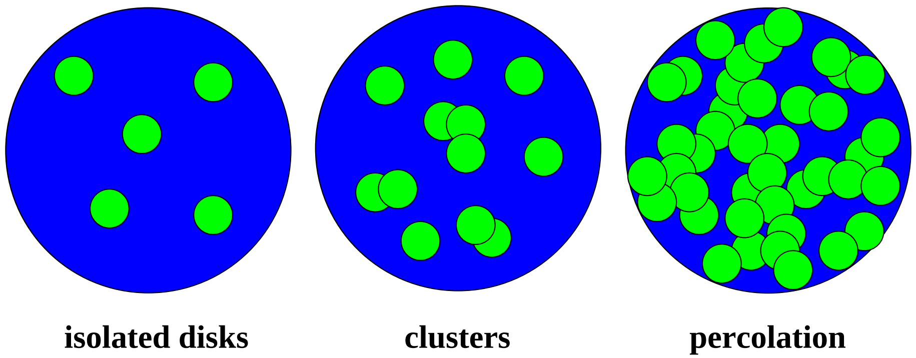
system becomes scale-invariant

$$\Gamma(kr)/\Gamma(r) = k^{-1} \sim r - independent.$$

critical behavior $\longrightarrow$ scale-invariance at the critical point,

$\Rightarrow$ power-law behavior $\sim$ indication of criticality.

repeat for percolation




cluster size $s$ distribution $n(s, \rho)$ at density $\rho$

$$n(s, \rho) \sim \frac{\exp[-s/\sigma(\rho)]}{s}$$

with $\sigma(\rho)$ average cluster size at given $\rho$.

**at critical point, largest cluster diverges, $\sigma(\rho \to \rho_c) \to \infty$**
**and**

$$n(s, \rho_c) \sim \frac{1}{s}$$

**power-law distribution of cluster size.**

**General conclusion:**

**criticaliy $\sim$ power-law distribution, scale-invariance**

# Language Structure

Conventional questions about literature texts:
- meaning and aim, style, details on author, etc.

New approach by George K. Zipf (Harvard):
- text is a many-body system, statistical treatment, words are the constituents

what are the most frequently occurring words?
- ranking order k=1,2,3,... from Wikipedia & more...; first ten:

in English: the of and to in a is was that for ....

Zipf's discovery: frequency pattern, with $f(the) = f(1)$

$f(of) = f(2) \simeq f(1)/2; f(and) = f(3) \simeq f(1)/3; f(to) = f(4) \simeq f(1)/4$

and so on; in general, power-law: Zipf's law: $f(k) \simeq \dfrac{f(1)}{k}$

## Language Structure

**Conventional questions about literature texts:**
**- meaning and aim, style, details on author, etc.**

**New approach by George K. Zipf (Harvard):**
**- text is a many-body system, statistical treatment, words are the constituents**

**what are the most frequently occurring words?**
**- ranking order k=1,2,3,... from Wikipedia & more...; first ten:**

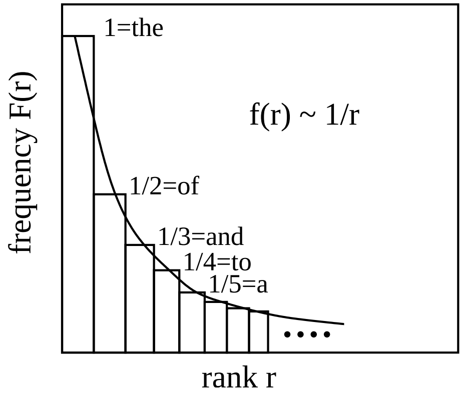
**in English: the of and to in a is was that for ....**

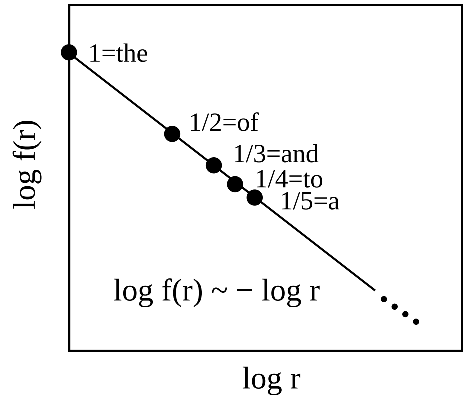
**Zipf's discovery: frequency pattern, with** $f(the) = f(1)$

$f(of) = f(2) \simeq f(1)/2; f(and) = f(3) \simeq f(1)/3; f(to) = f(4) \simeq f(1)/4$

**and so on; in general, power-law: Zipf's law:** $f(k) \simeq \dfrac{f(1)}{k}$

**in Polish: nie to sie na co ze jest do tak jak...**
**in German: der und die in von den mit zu ist für...**



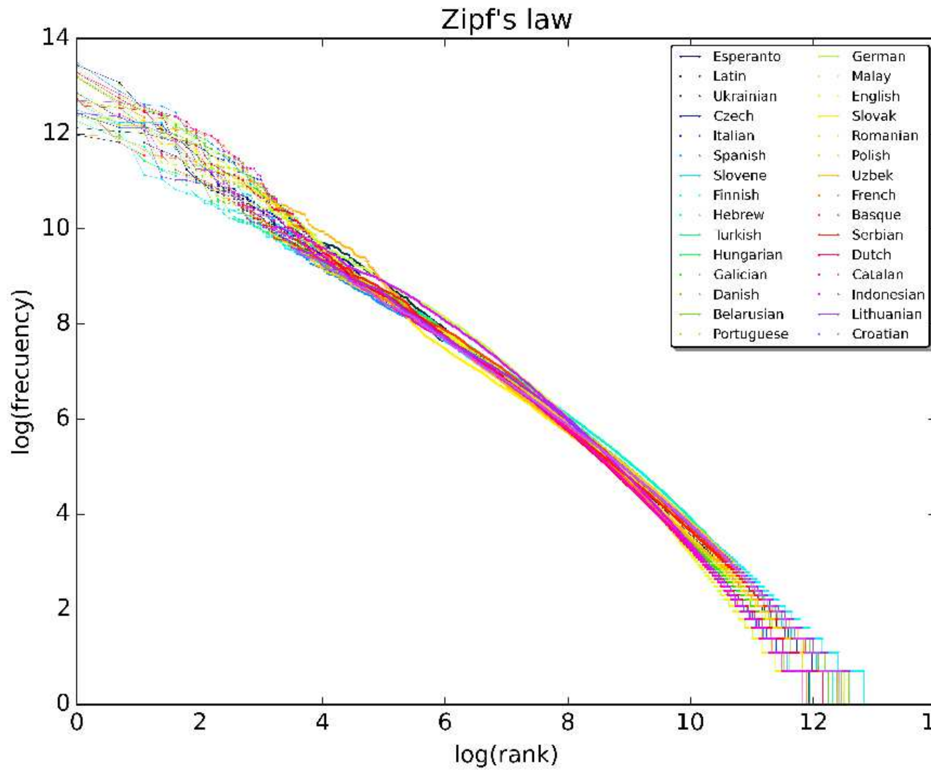




**Zipf used "Ulysses" by James Joyce; is the law general?**

**subsequent studies:**
**$\sim$ all texts in all languages, even**
- **Esperanto (artificial language)**
- **Meroitic (undeciphered ancient language)**

## Conclusion

**human languages contain intrinsic power-law distribution of word frequencies;**
**languages are somehow poised at a critical point**


Zipf's law
log(frecuency)
log(rank)
Esperanto
Latin
Ukrainian
Czech
Italian
Spanish
Slovene
Finnish
Hebrew
Turkish
Hungarian
Galician
Danish
Belarusian
Portuguese
German
Malay
English
Slovak
Romanian
Polish
Uzbek
French
Basque
Serbian
Dutch
Catalan
Indonesian
Lithuanian
Croatian

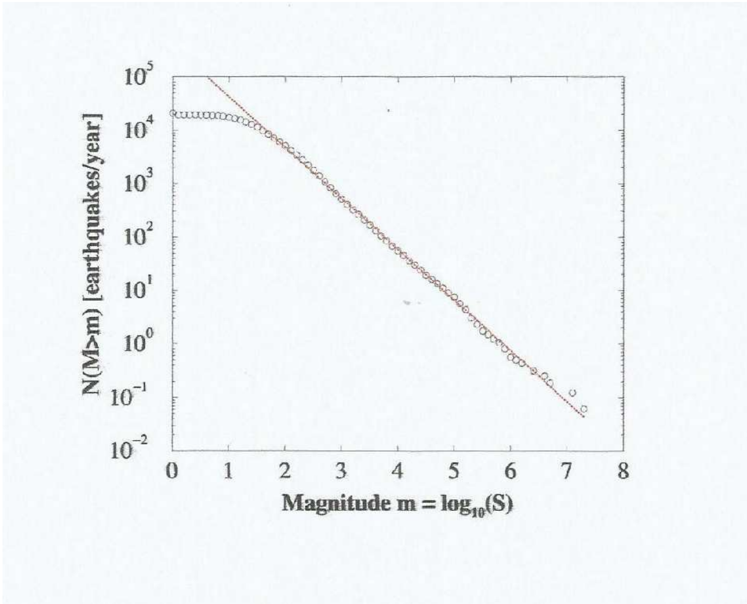
# Why?

Many attempts, linguistic, sociological, statistical (monkey typing)....

But:

there are many other instances of such behavior:

– earth quakes (Gutenberg-Richter law)



– **city sizes (Auerbach law)**

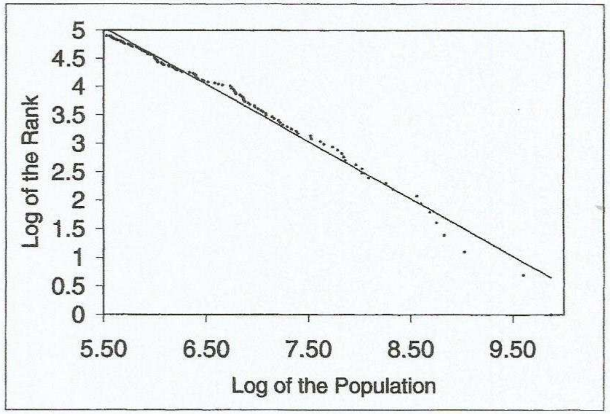



FIGURE I
Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].

**consider a case detached from observation, not relying on "data":**

**<u>Prime Number Components of Integers</u>**

**HS: arXiv:2403.12773**

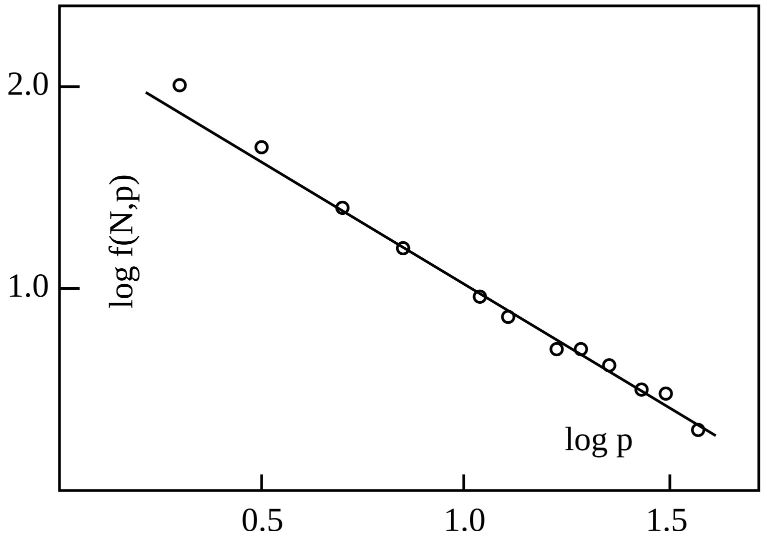
**N=10: 2, 3, 4=2x2, 5, 6=2x3, 7, 8=2x2x2, 9=3x3, 10=2x5**

**f(2)=8, f(3)=3, f(5)=1, f(7)=1**

**try larger sets: N=1-100 and N=900-1000**

| p | 2 | 3 | 5 | 7 | 11 | 13 | 17 | 19 | 23 | 29 | 31 | 37 | 41 | 43 | 47 | 53 | ... | 97 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| f(100) | 97 | 48 | 24 | 16 | 9 | 7 | 5 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | ... | 1 |
| f(1000) | 98 | 50 | 25 | 16 | 10 | 6 | 6 | 5 | 2 | 3 | 4 | 3 | 3 | 3 | 2 | 2 | ... | 1 |

**so far, still *empirical*: the data look like Zipf.**



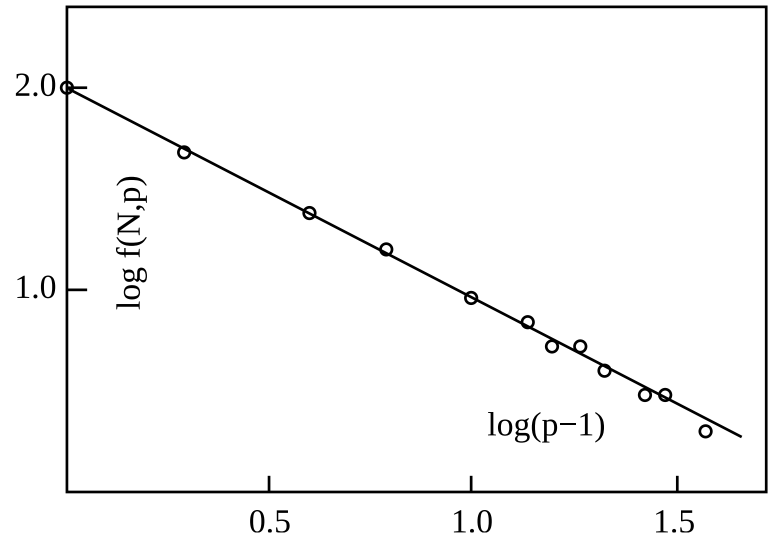
**But here we have**

$$\boldsymbol{f(N,p) = N(1/p + 1/p^2 + 1/p^3 + ...) \simeq \frac{N}{(p-1)}},$$

**so that**

$$\boldsymbol{\log\ f(N,p) \simeq \log N - \log(p-1)}$$

**for large N and large p we have** *analytically derived* **a Zipf form.**



# Conclusion

**Zipf's law is perhaps the most general known regularity in our world:**

**earthquakes, sandpiles, city sizes, prime numbers, economics, .... much more**

**but in general no derivation.**

**Pre-Galileo Stage: why do all objects fall at the same rate?**