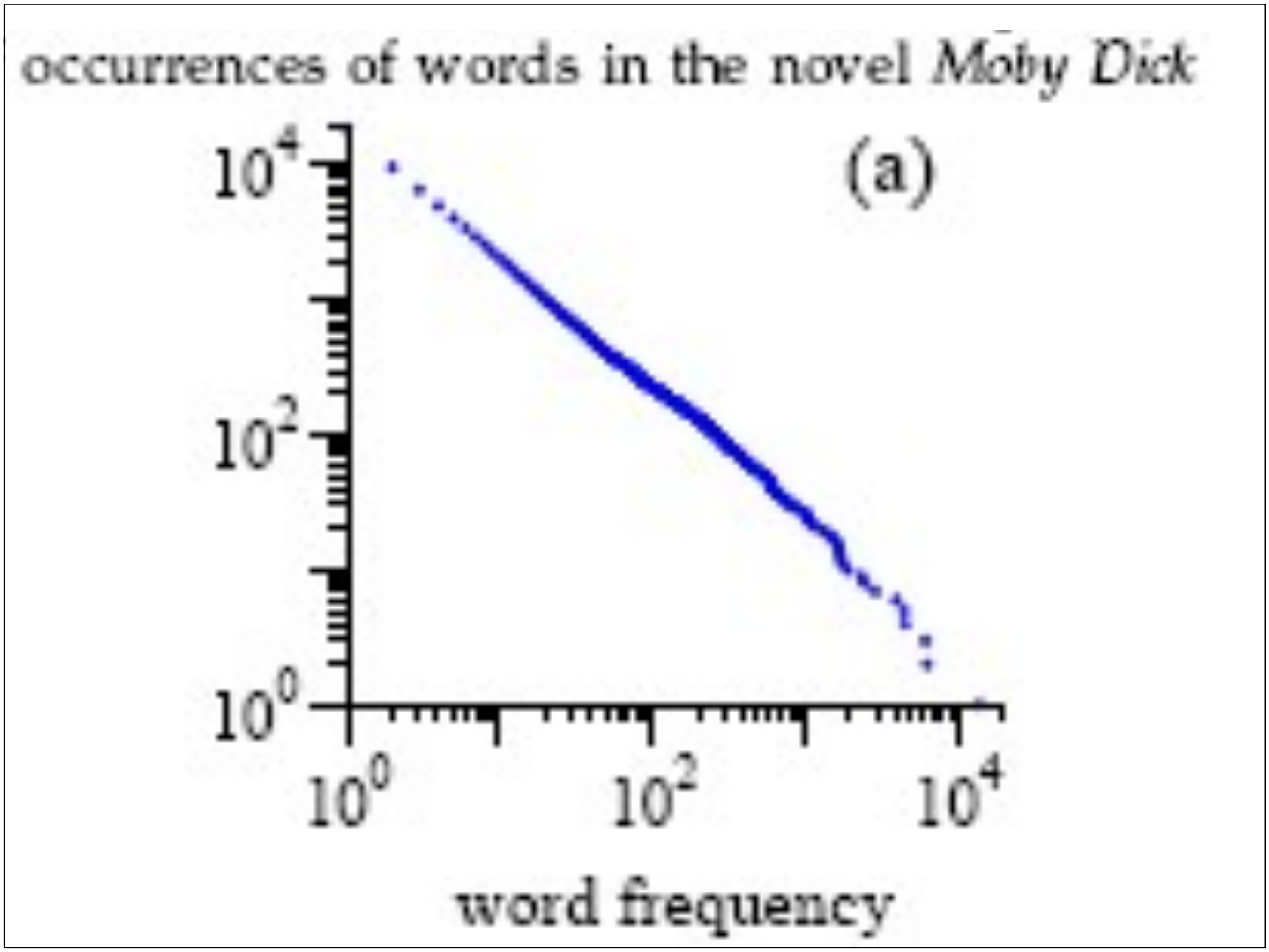
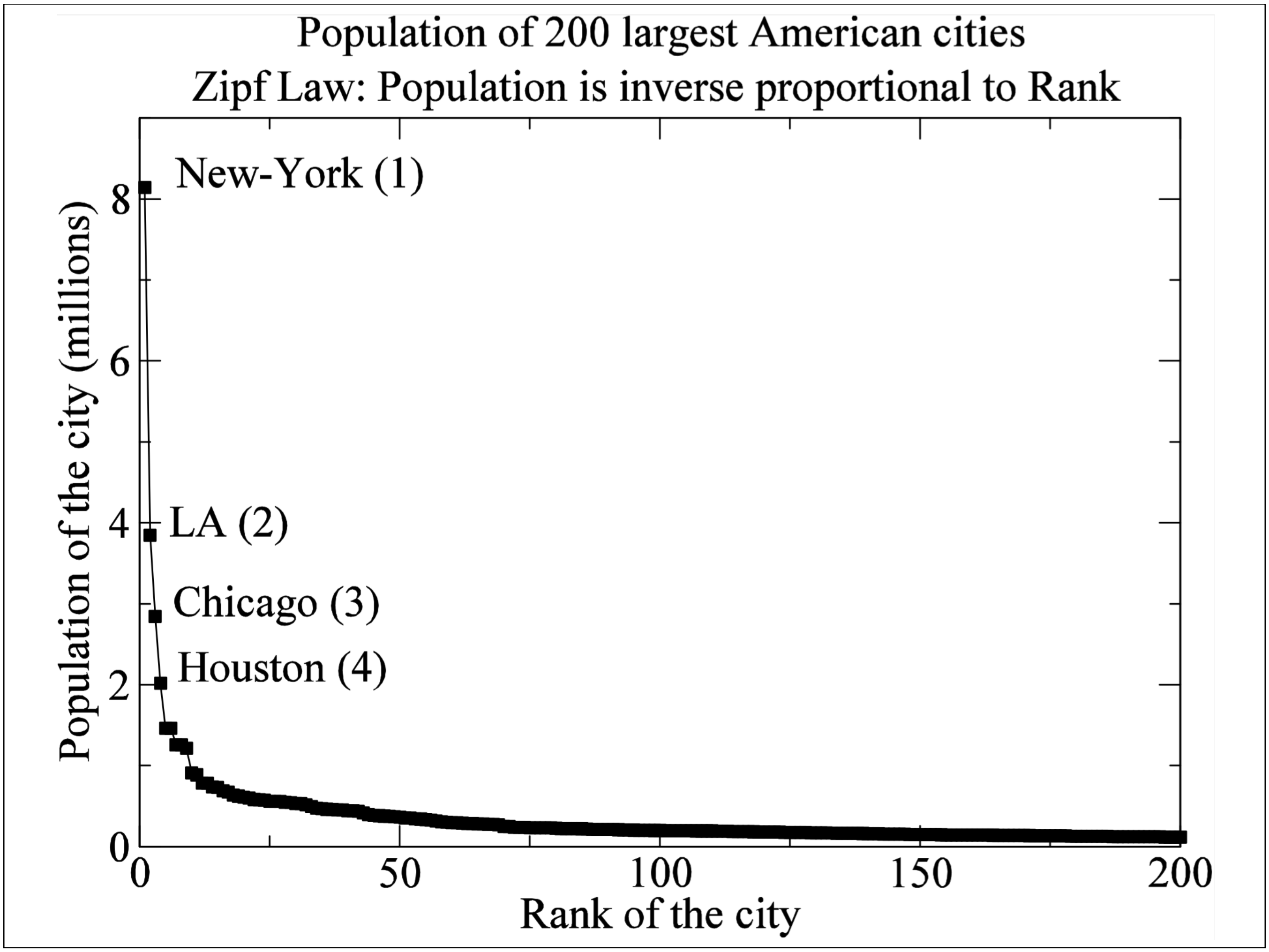
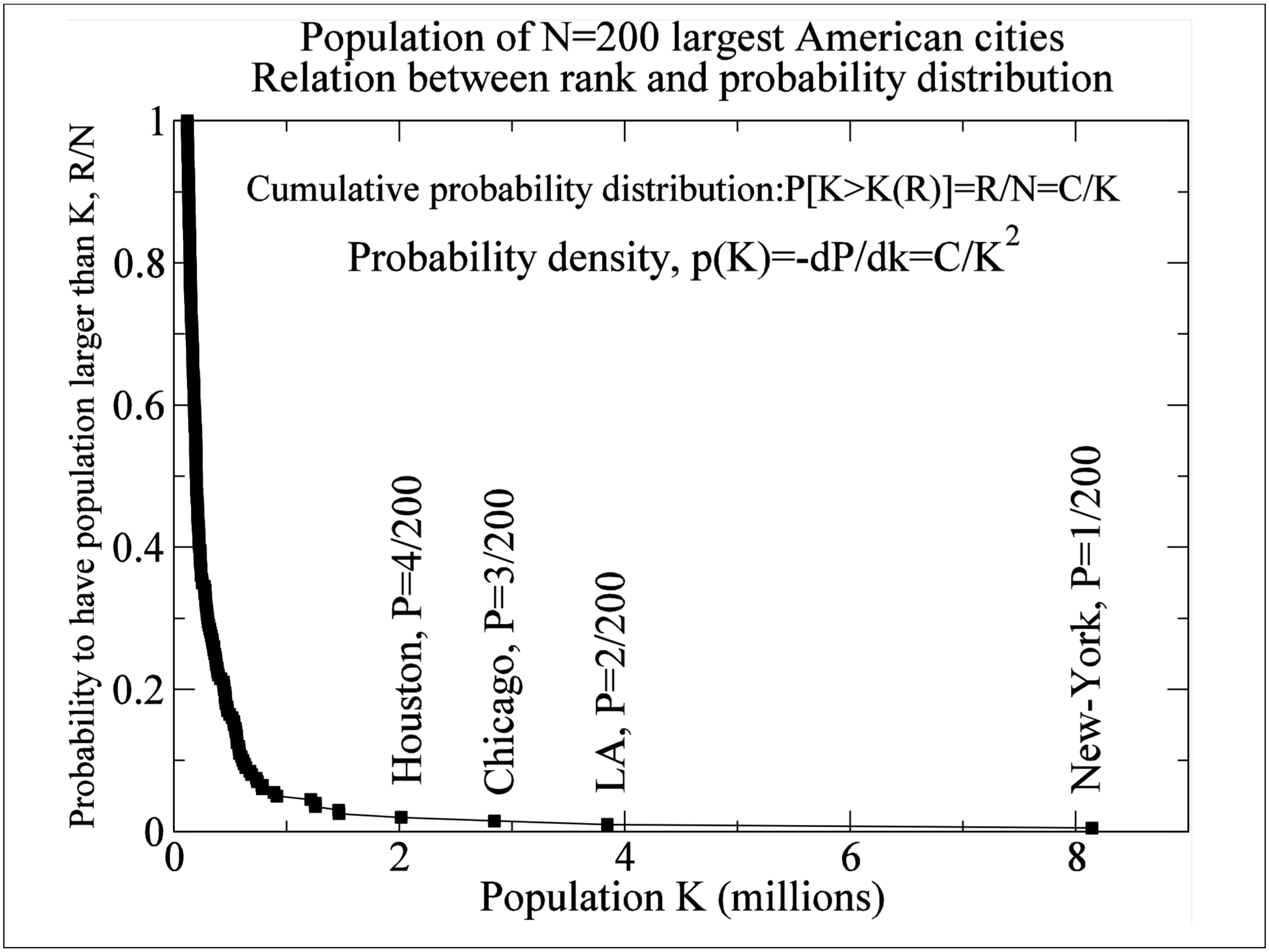
George Kingsley Zipf (1902-1950)

(1949): *Human behavior and the principle of least effort*

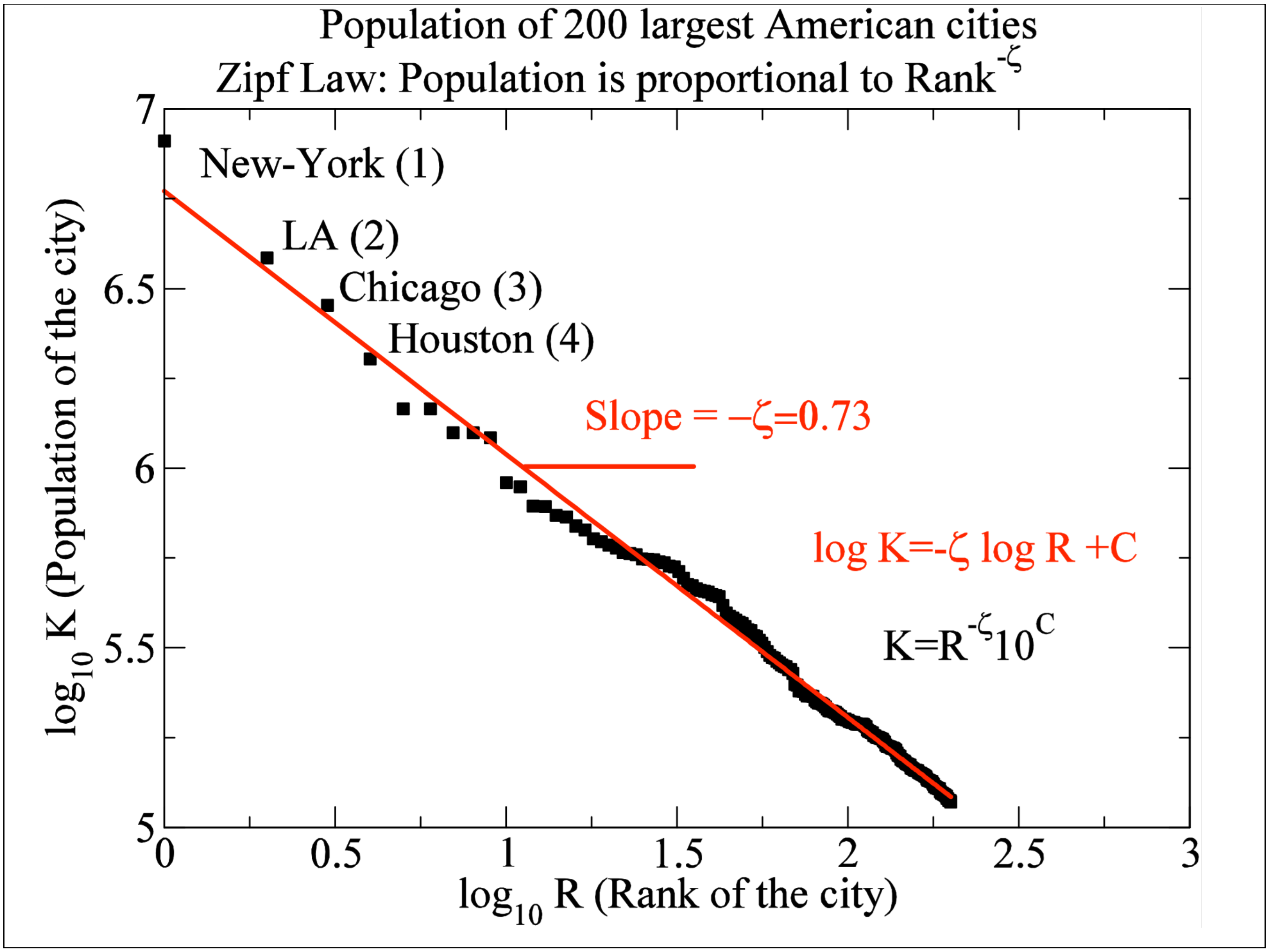

occurrences of words in the novel *Moby Dick*
(a)
10^4
10^2
10^0
10^0
10^2
10^4
word frequency

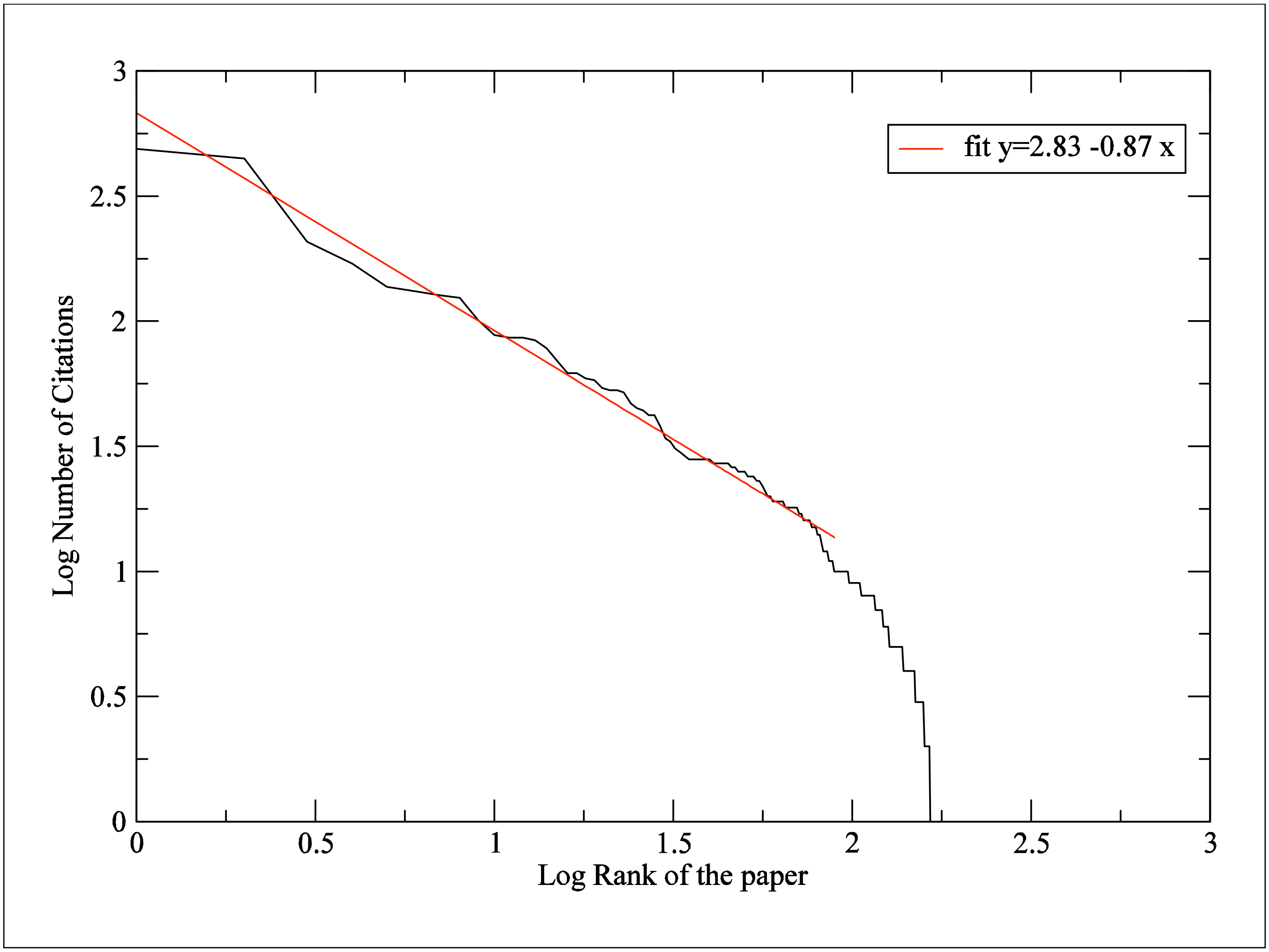
Rank	Name	Population
1	New York City, New York	8,143,197
2	Los Angeles, California	3,844,829
3	Chicago, Illinois	2,842,518
4	Houston, Texas	2,016,582
5	Philadelphia, Pennsylvania	1,463,281
6	Phoenix, Arizona	1,461,575
7	San Antonio, Texas	1,256,509
8	San Diego, California	1,255,540
9	Dallas, Texas	1,213,825
10	San Jose, California	912,332
11	Detroit, Michigan	886,671
12	Indianapolis, Indiana	784,118
13	Jacksonville, Florida	782,623
14	San Francisco, California	739,426
15	Columbus, Ohio	730,657
16	Austin, Texas	690,252
17	Memphis, Tennessee	672,277
18	Baltimore, Maryland	635,815
19	Fort Worth, Texas	624,067
20	Charlotte, North Carolina	610,949


Population of 200 largest American cities
Zipf Law: Population is inverse proportional to Rank
New-York (1)
LA (2)
Chicago (3)
Houston (4)
Population of the city (millions)
Rank of the city
0
2
4
6
8
0
50
100
150
200



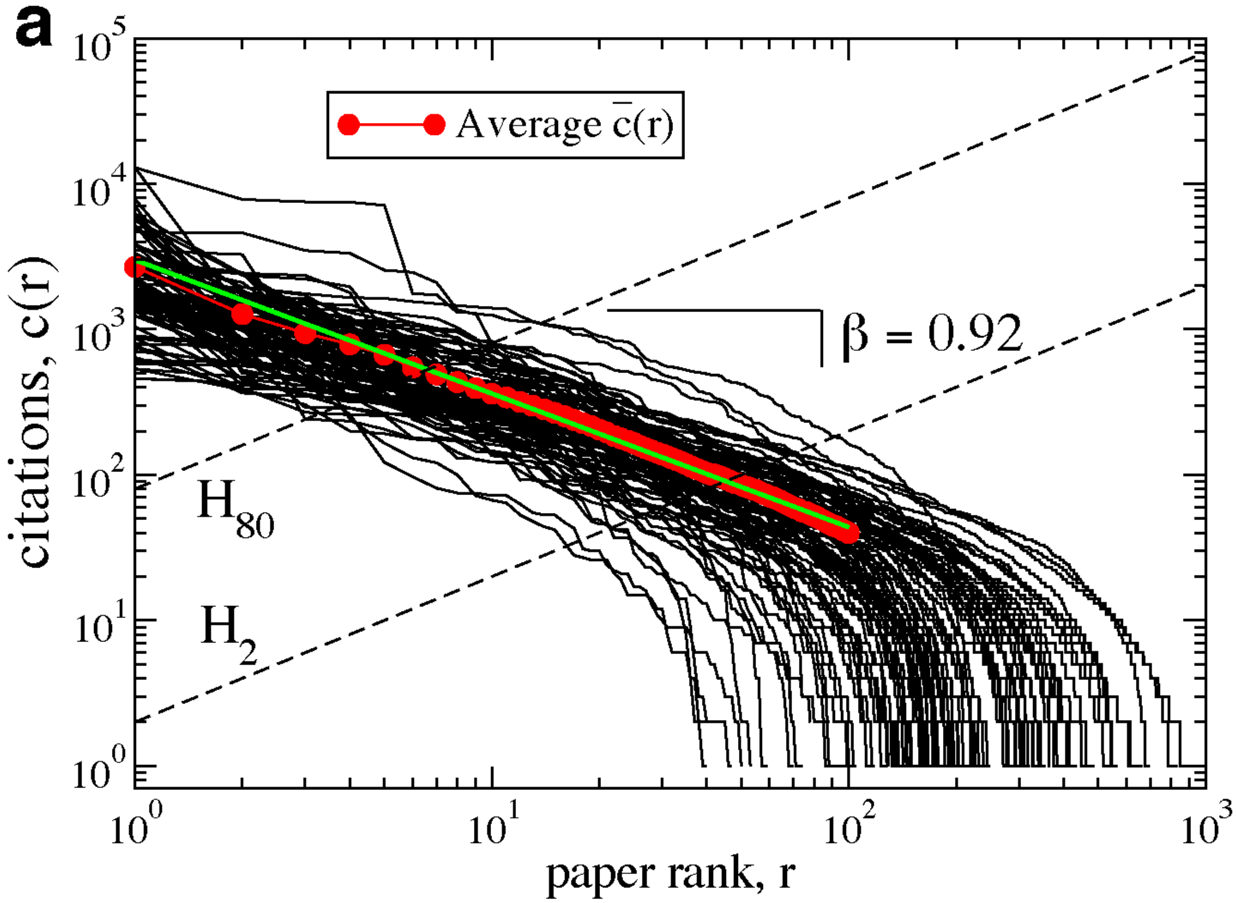
Population of N=200 largest American cities
Relation between rank and probability distribution
Cumulative probability distribution:P[K>K(R)]=R/N=C/K
Probability density, p(K)=-dP/dk=C/K^2
Houston, P=4/200
Chicago, P=3/200
LA, P=2/200
New-York, P=1/200
Probability to have population larger than K, R/N
Population K (millions)
0
0.2
0.4
0.6
0.8
1
2
4
6
8

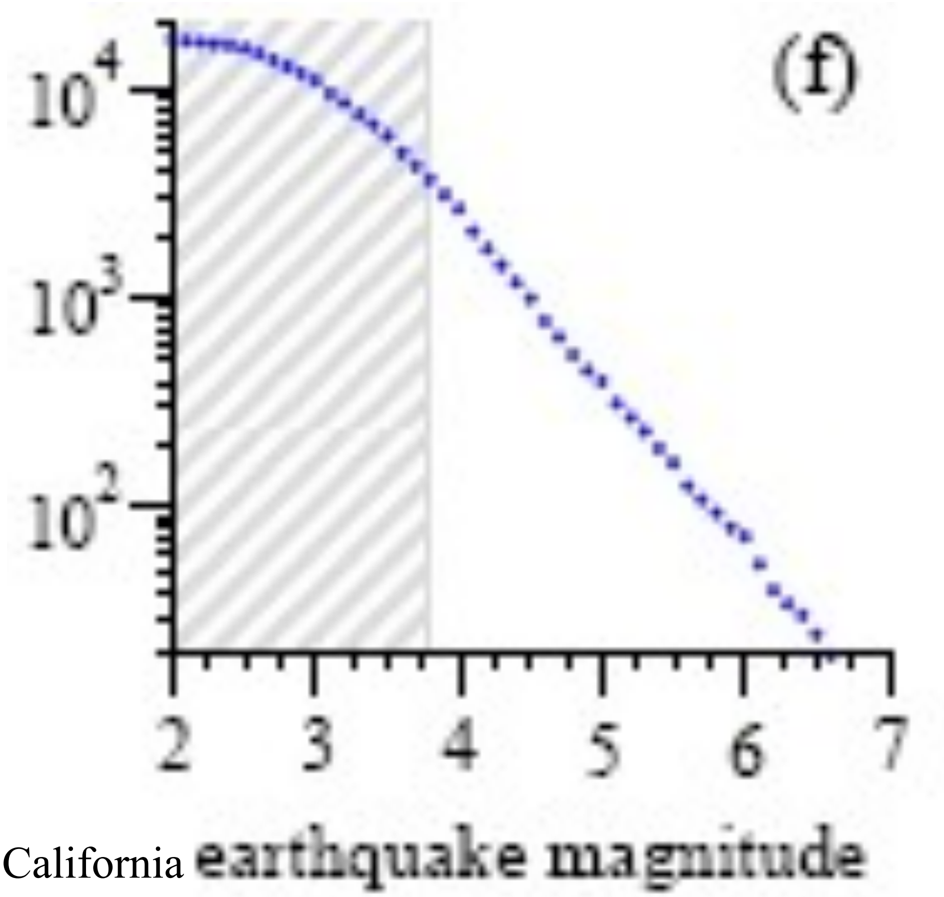

Population of 200 largest American cities
Zipf Law: Population is proportional to Rank$^{-\zeta}$
New-York (1)
LA (2)
Chicago (3)
Houston (4)
Slope = $-\zeta$=0.73
log K=-ζ log R +C
K=R$^{-\zeta}$10^{C}
$\log_{10}$ K (Population of the city)
$\log_{10}$ R (Rank of the city)
7
6.5
6
5.5
5
0
0.5
1
1.5
2
2.5
3



fit y=2.83 -0.87 x
Log Number of Citations
Log Rank of the paper
0
0.5
1
1.5
2
2.5
3

Rank-ordered citations of 100 physicists

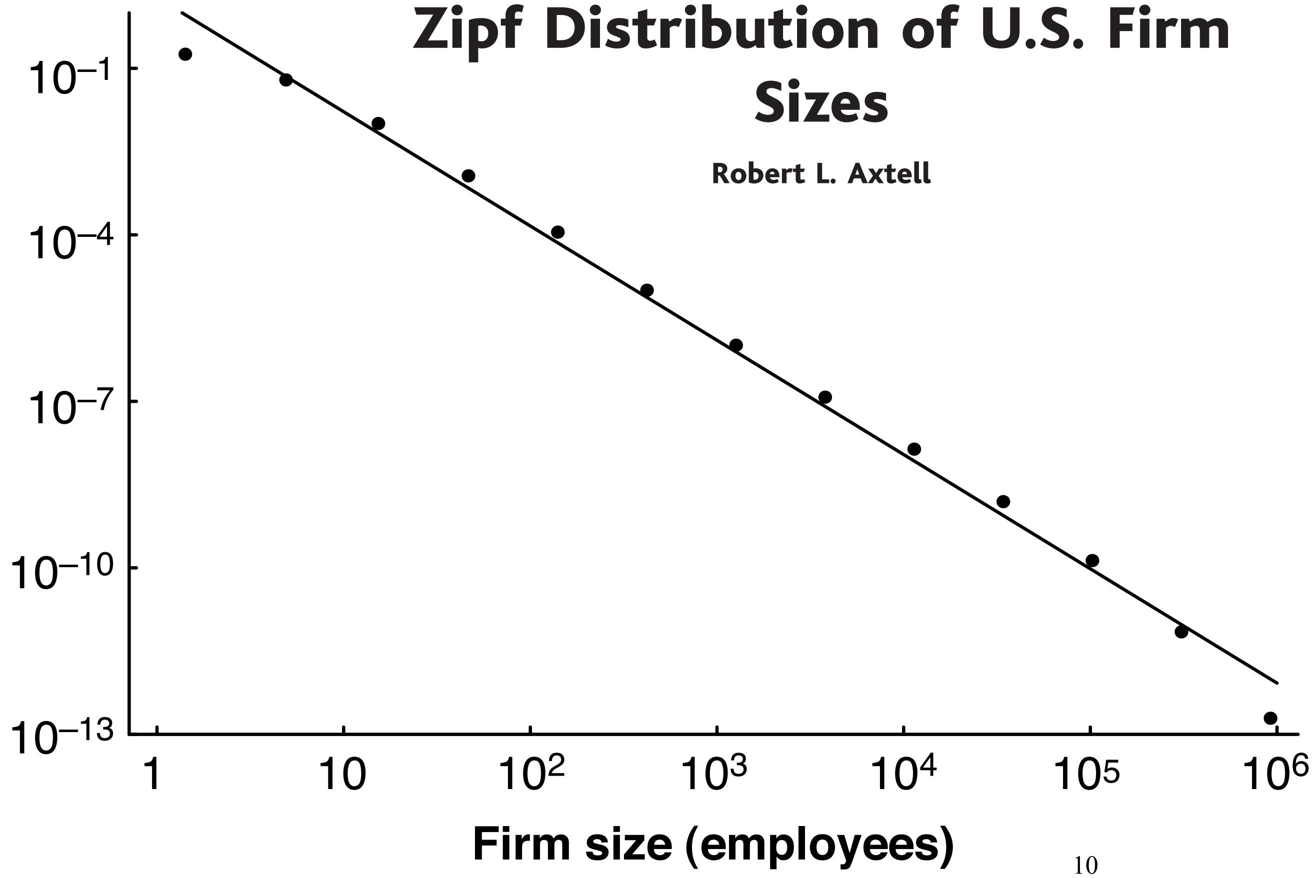




California

Zipf Distribution of U.S. Firm Sizes

Robert L. Axtell



Elementary derivation of the Zipf law

Rules of the model:

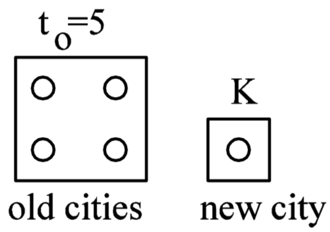
- At each time step a person is born in a city.
- All cities have approximately the same birth rate.
- With very small probability a person creates a new city.

Properties:

- The total population, n_0, of the cities existing at time t_0 is proportional to t_0 : $n_0 \sim t_0$
- The rank of the city created at time t is proportional to t: $R \sim t_0$
- The ratio of the size of this city to the total population remains the same $K/n \sim 1/n_0 \Rightarrow K \sim 1/n_0 \sim 1/t_0$
- Finally: $K \sim 1/t_0 \sim 1/R \Rightarrow K \sim 1/R$

Conclusion:

- Size is inversely proportional to its rank.

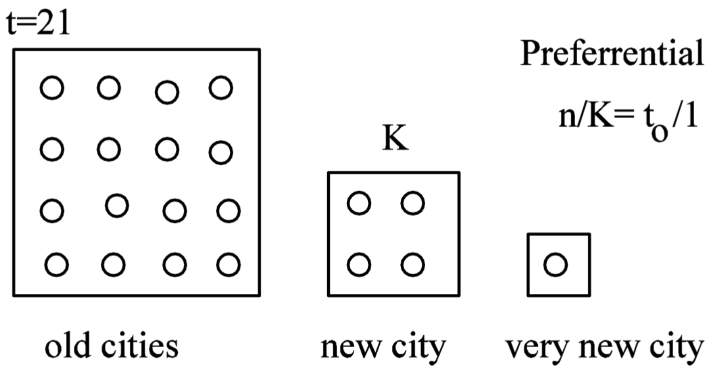


t - time

n - total population, n=t

N - number of cities: N=t b

In this example b=1/5, but we assume that b is very very small



Preferrential attachment:

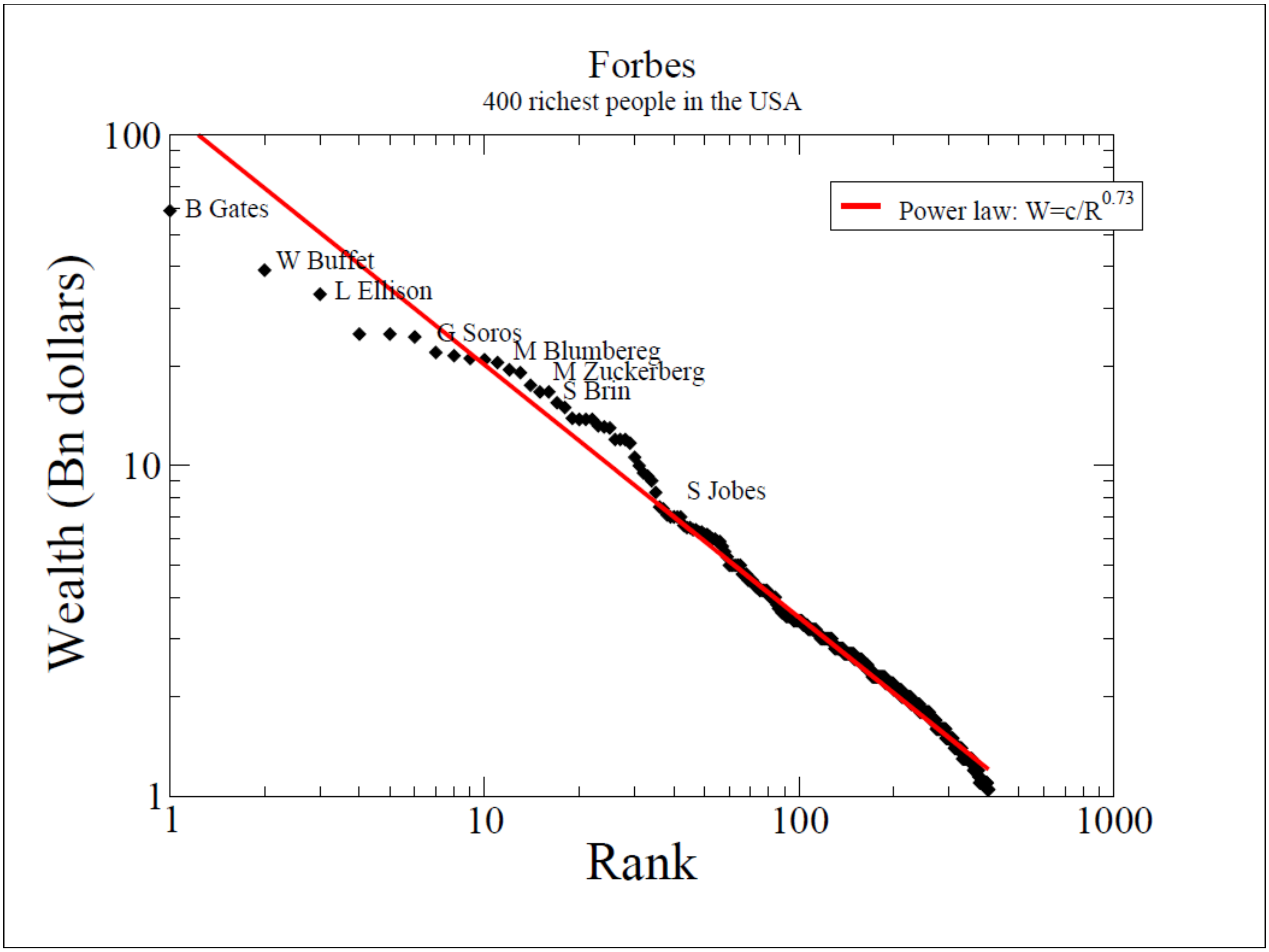
$n/K = t_o/1$

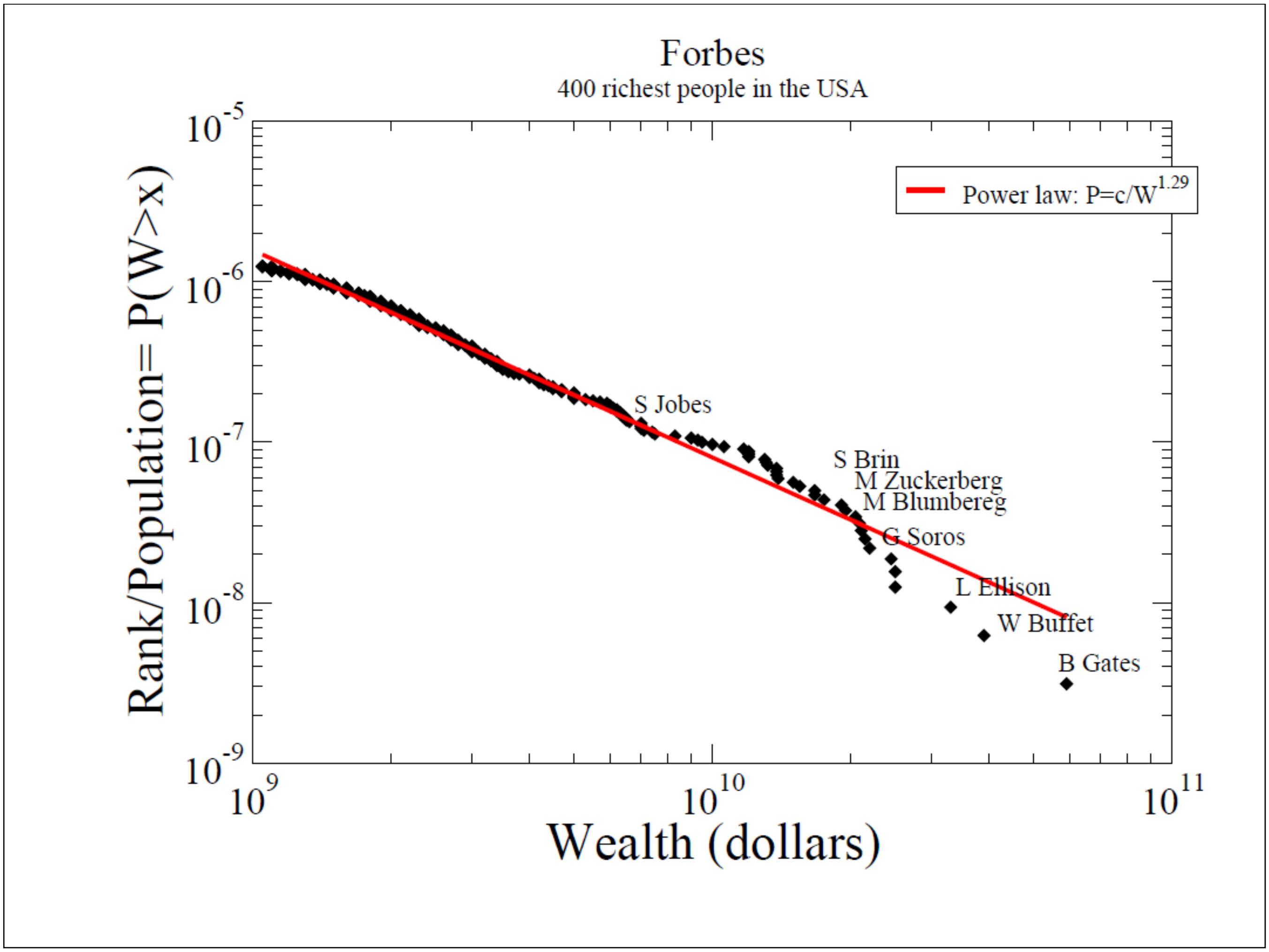
A city that born at time t_o has rank $R=t_o b$, it remains the same for the entire history
Due to preferntial attachment its population remains $K=n/t_o$ (if we neglect newer cities)

Thus $t_o=R/b$ and $K=nb/R$ $\Rightarrow$

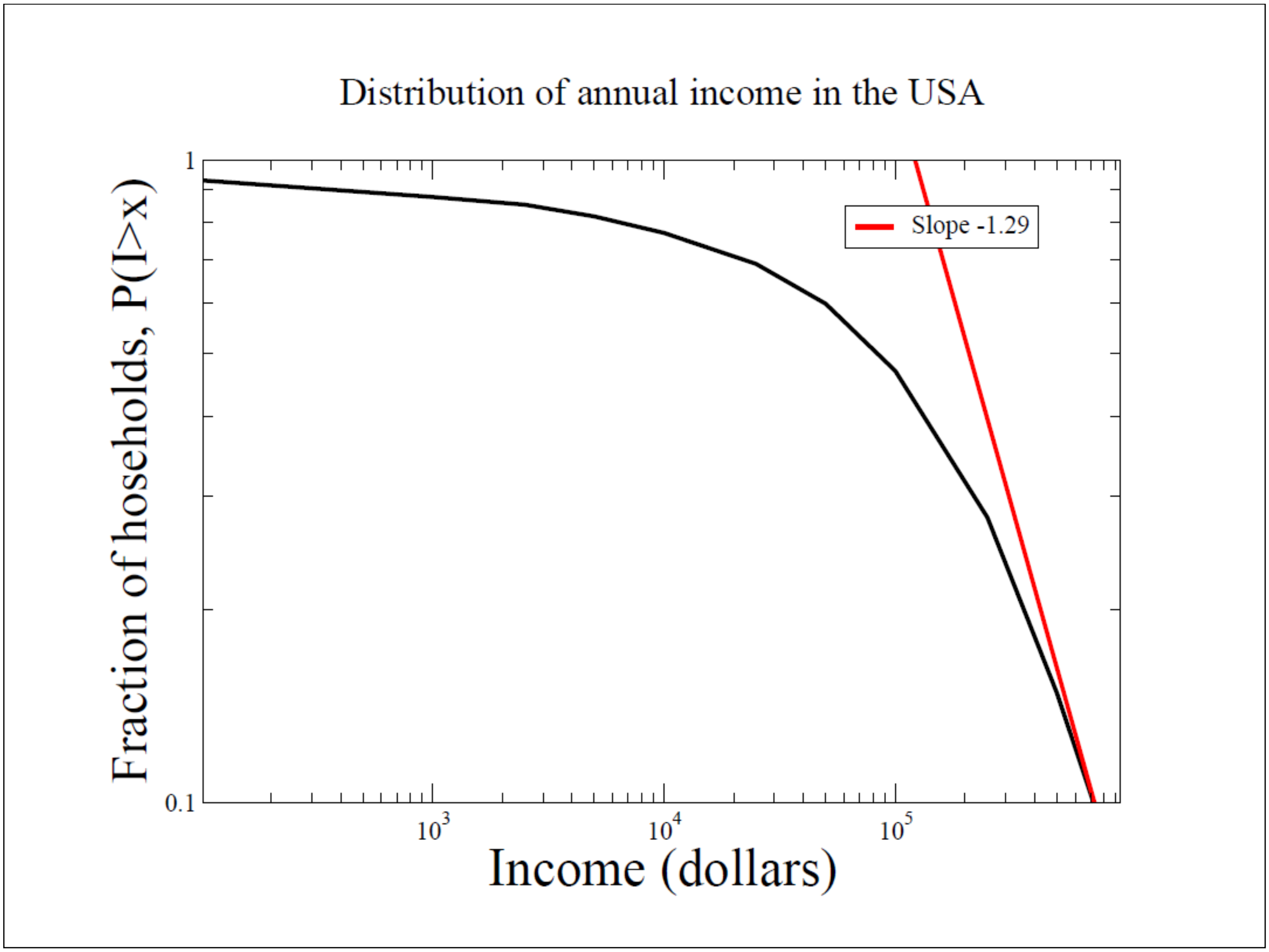
Population is inverse proportional to Rank

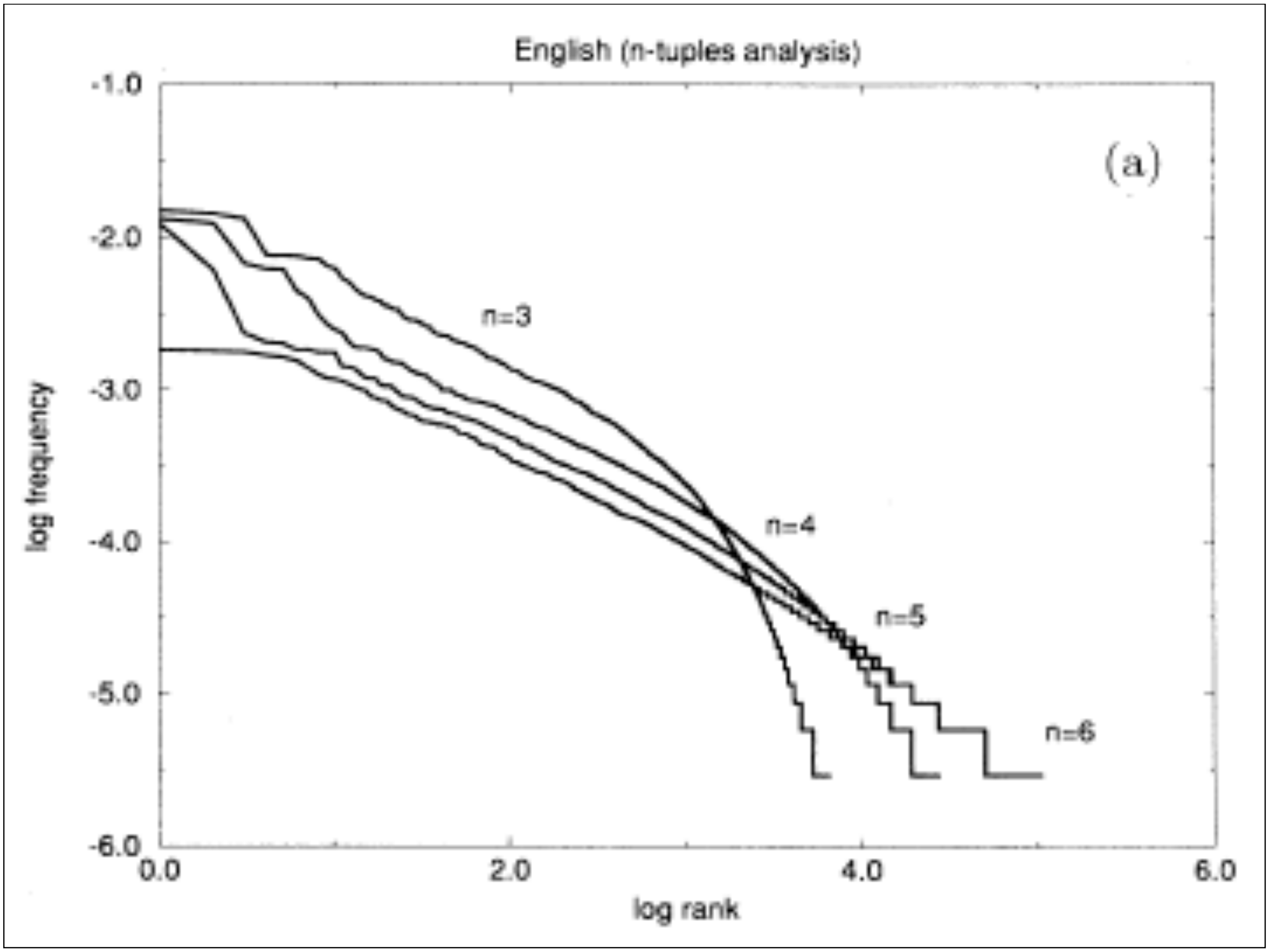
1		**Bill Gates**	$59 B	55	Medina, Washington	Microsoft
2		**Warren Buffett**	$39 B	81	Omaha, Nebraska	Berkshire Hathaway
3		**Larry Ellison**	$33 B	67	Woodside, California	Oracle
4		**Charles Koch**	$25 B	75	Wichita, Kansas	diversified
4		**David Koch**	$25 B	71	New York, New York	diversified
6		**Christy Walton**	$24.5 B	56	Jackson, Wyoming	Wal-Mart
7		**George Soros**	$22 B	81	Katonah, New York	hedge funds
8		**Sheldon Adelson**	$21.5 B	78	Las Vegas, Nevada	casinos
9		**Jim Walton**	$21.1 B	63	Bentonville, Arkansas	Wal-Mart
10		**Alice Walton**	$20.9 B	61	Fort Worth, Texas	Wal-Mart

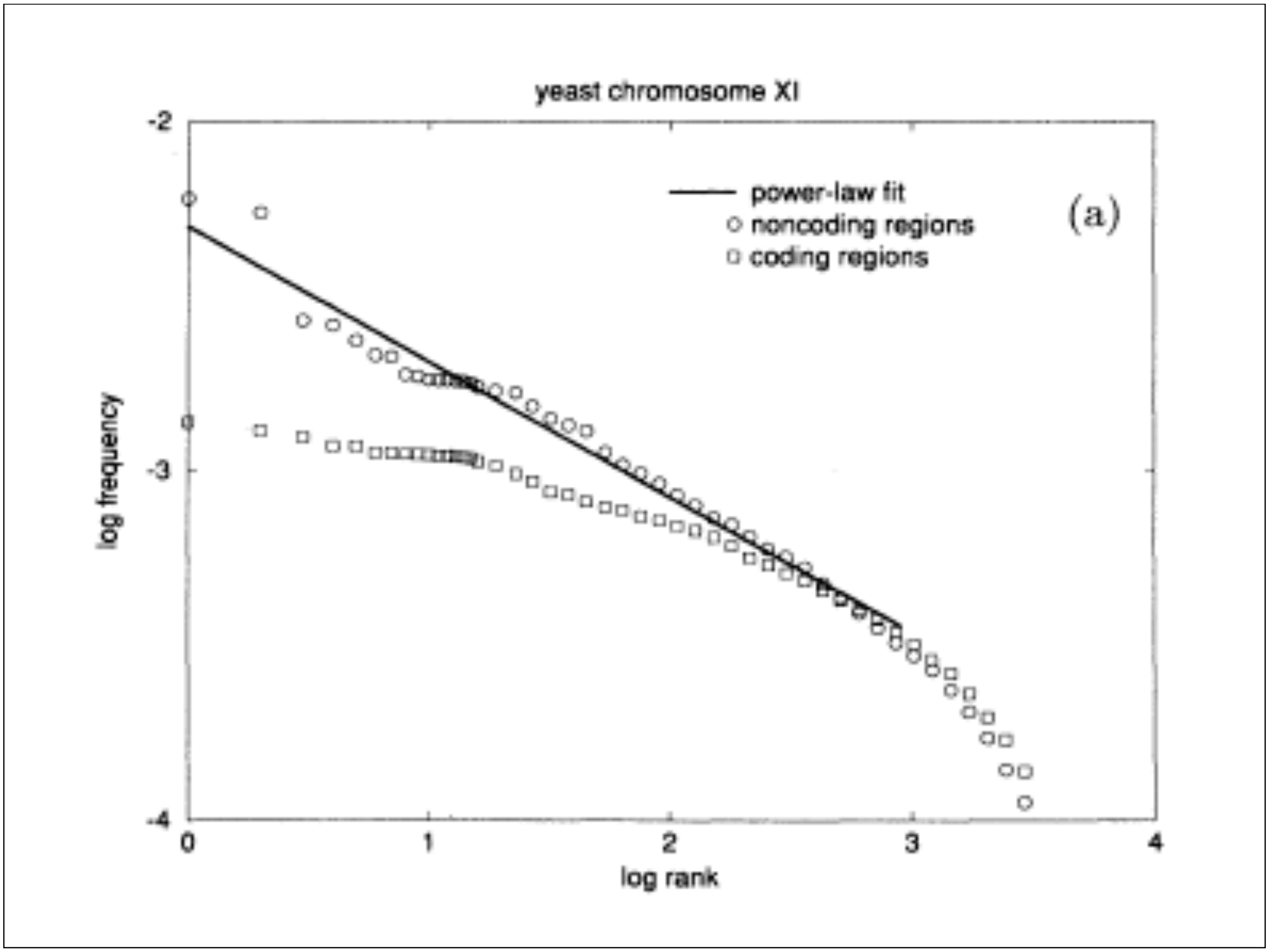

Forbes
400 richest people in the USA
Power law: W=c/R^0.73
B Gates
W Buffet
L Ellison
G Soros
M Blumbereg
M Zuckerberg
S Brin
S Jobes
Wealth (Bn dollars)
Rank
1
10
100
1000

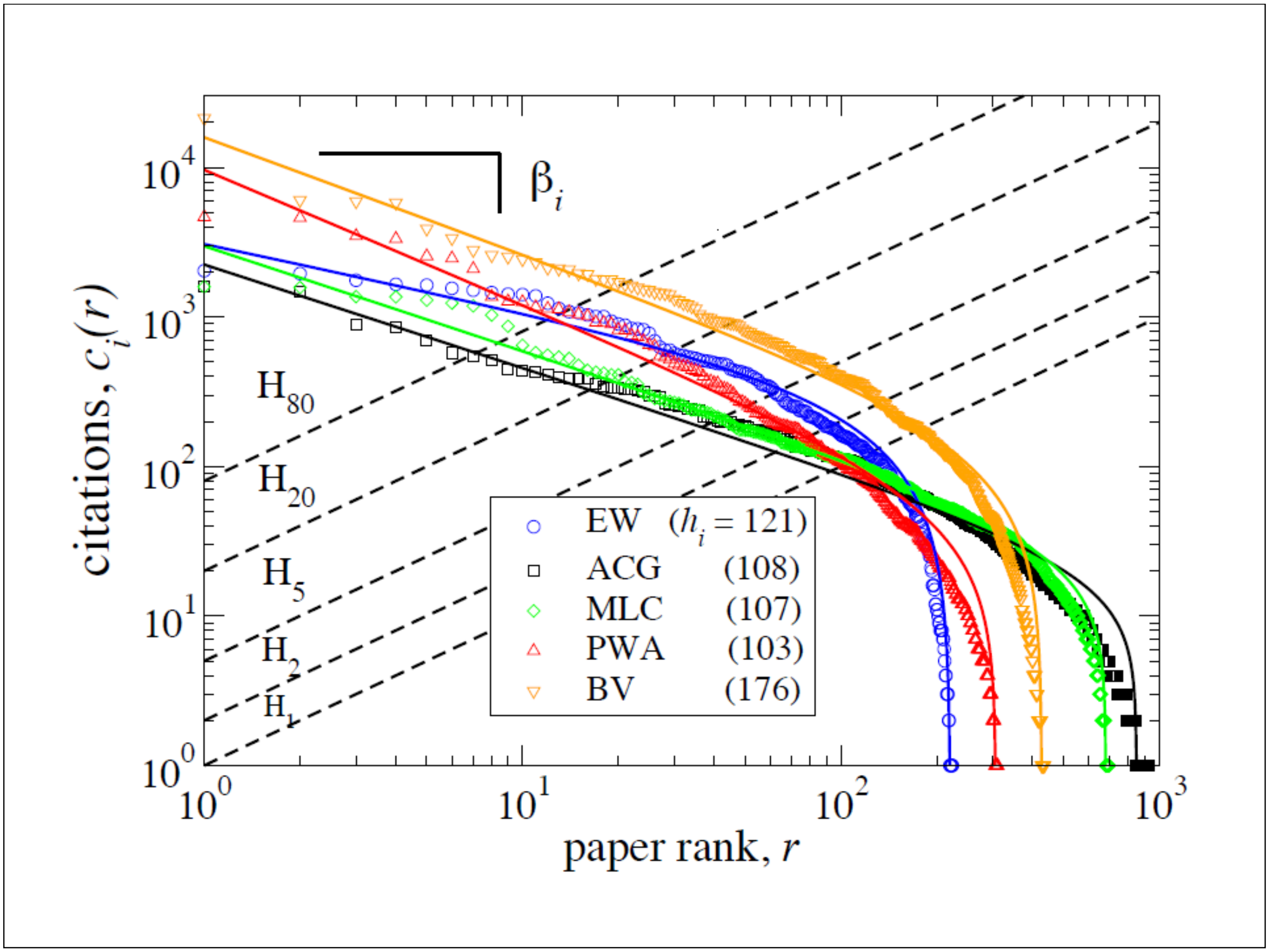


Forbes
400 richest people in the USA
Power law: P=c/W^1.29
Rank/Population= P(W>x)
Wealth (dollars)
S Jobes
S Brin
M Zuckerberg
M Blumbereg
G Soros
L Ellison
W Buffet
B Gates


Distribution of annual income in the USA
Fraction of hoseholds, P(I>x)
Income (dollars)
Slope -1.29
1
0.1
10^3
10^4
10^5

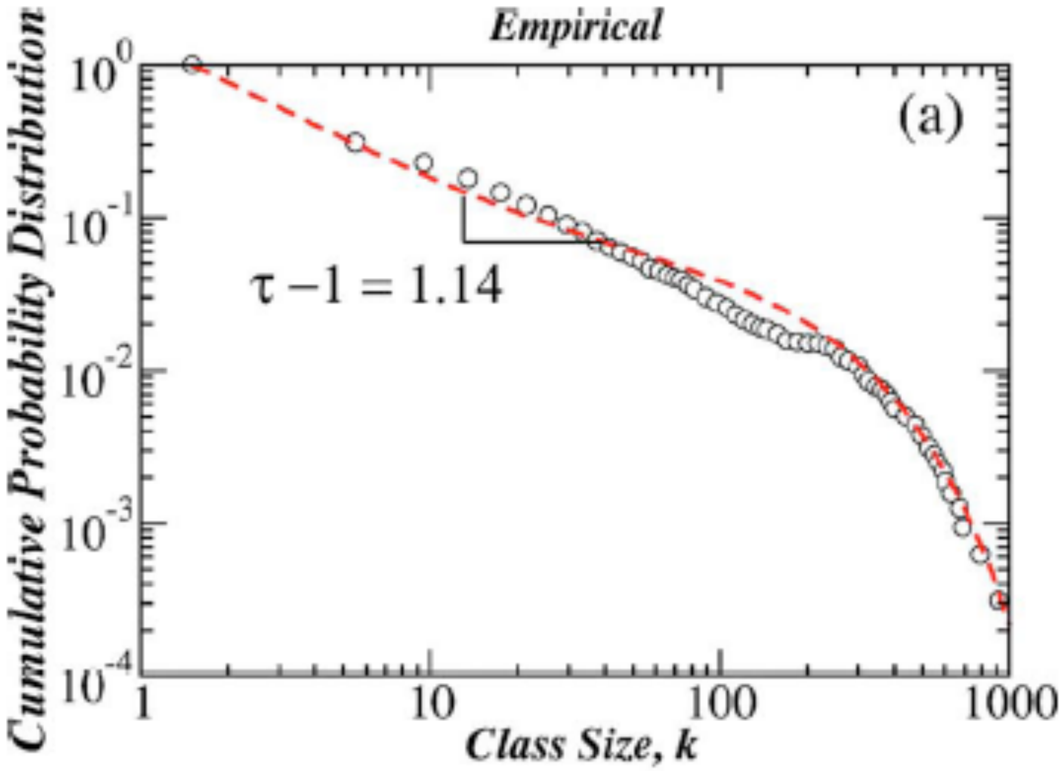

English (n-tuples analysis)
(a)
log frequency
log rank
n=3
n=4
n=5
n=6
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
0.0
2.0
4.0
6.0


yeast chromosome XI
power-law fit
noncoding regions
coding regions
(a)
-2
-3
-4
log frequency
0
1
2
3
4
log rank



citations, $c_i(r)$
paper rank, r
β_i
H_{80}
H_{20}
H_5
H_2
H_1
EW (h_i = 121)
ACG (108)
MLC (107)
PWA (103)
BV (176)
10^0
10^1
10^2
10^3
10^4

Our goal is to build a simple model which would explain this graph



Preferential Attachment Model

Let us take a look on the distribution of citations of papers of a given author or population of cities. In over model we will call cities or papers classes and we will call people and citations units.

Let us assume that in a unit of time

(1) existing classes get λ new units,
which are distributed to the existing classes
in proportion to their existing size measured in
number of units.
and

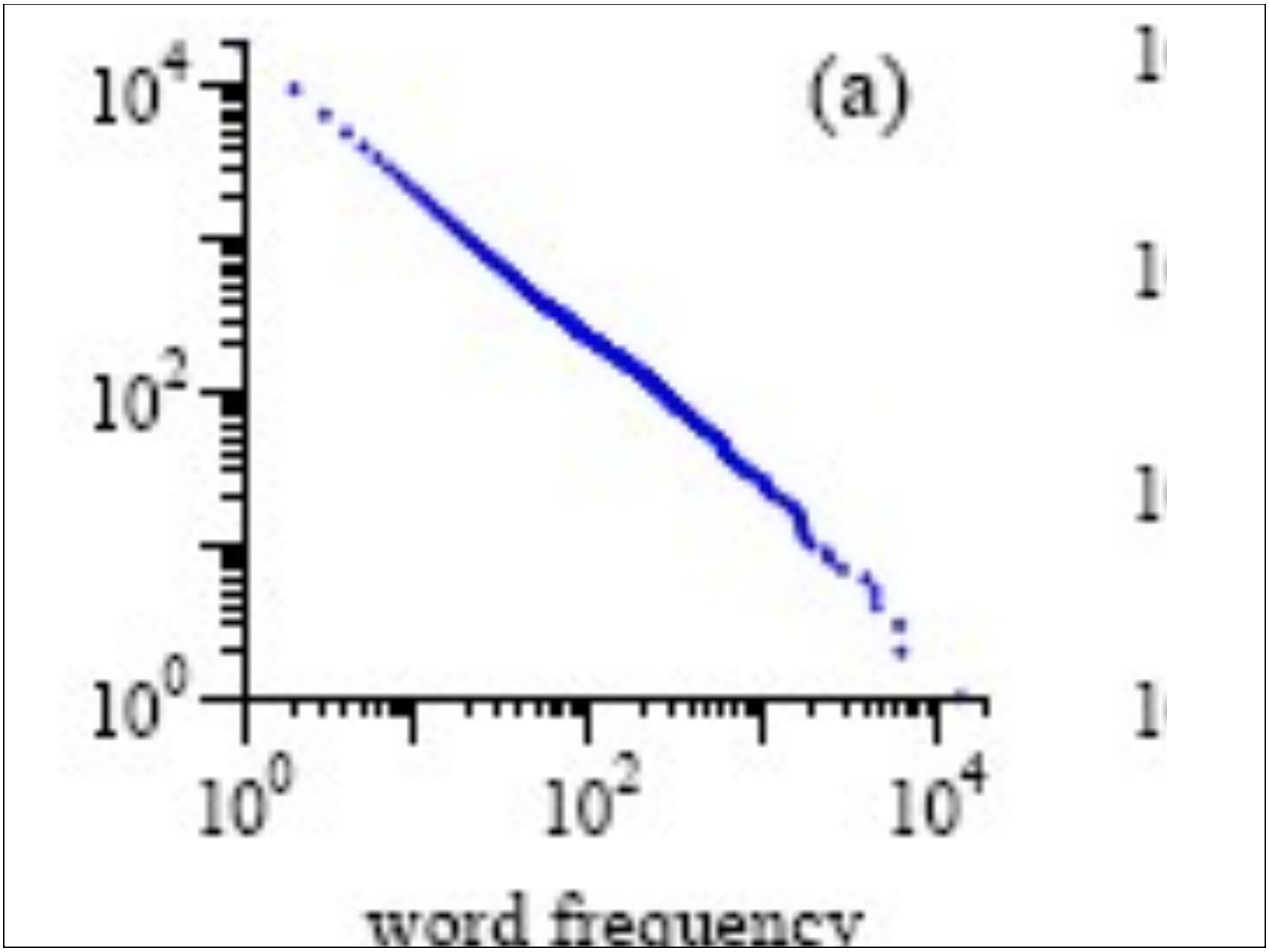
(2) β new classes, each of unit size are created. We introduce

$N(t)$ - number of classes as function of time;
$n(t)$ - number of units as function of time;
$N_0 = N(0)$ - initial number of classes at $t = 0$.
$n_0 = n(0)$ - initial number of units at $t = 0$.
Then

$$\frac{dN}{dt} = \beta \quad \text{and} \quad \frac{dn}{dt} = \beta + \lambda. \tag{1}$$

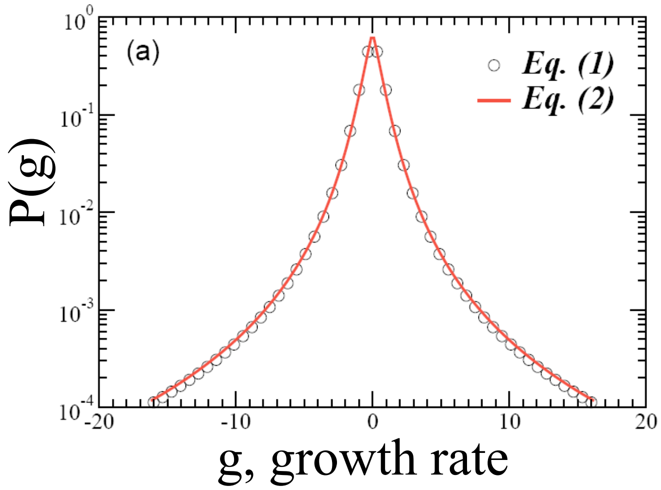
Integration gives:

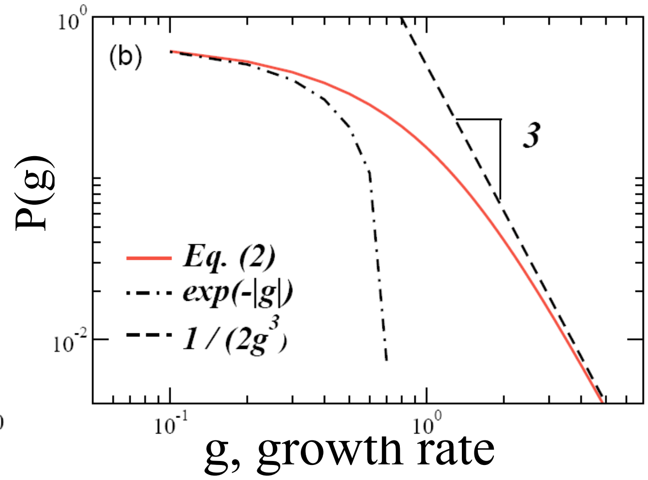
$$N(t) = \beta t + N_0 \quad \text{and} \quad n(t) = (\beta + \lambda)t + n_0. \tag{2}$$



(a)
10^4
10^2
10^0
10^0
10^2
10^4
word frequency

Crossover in P(g) from Exp. to Power Law



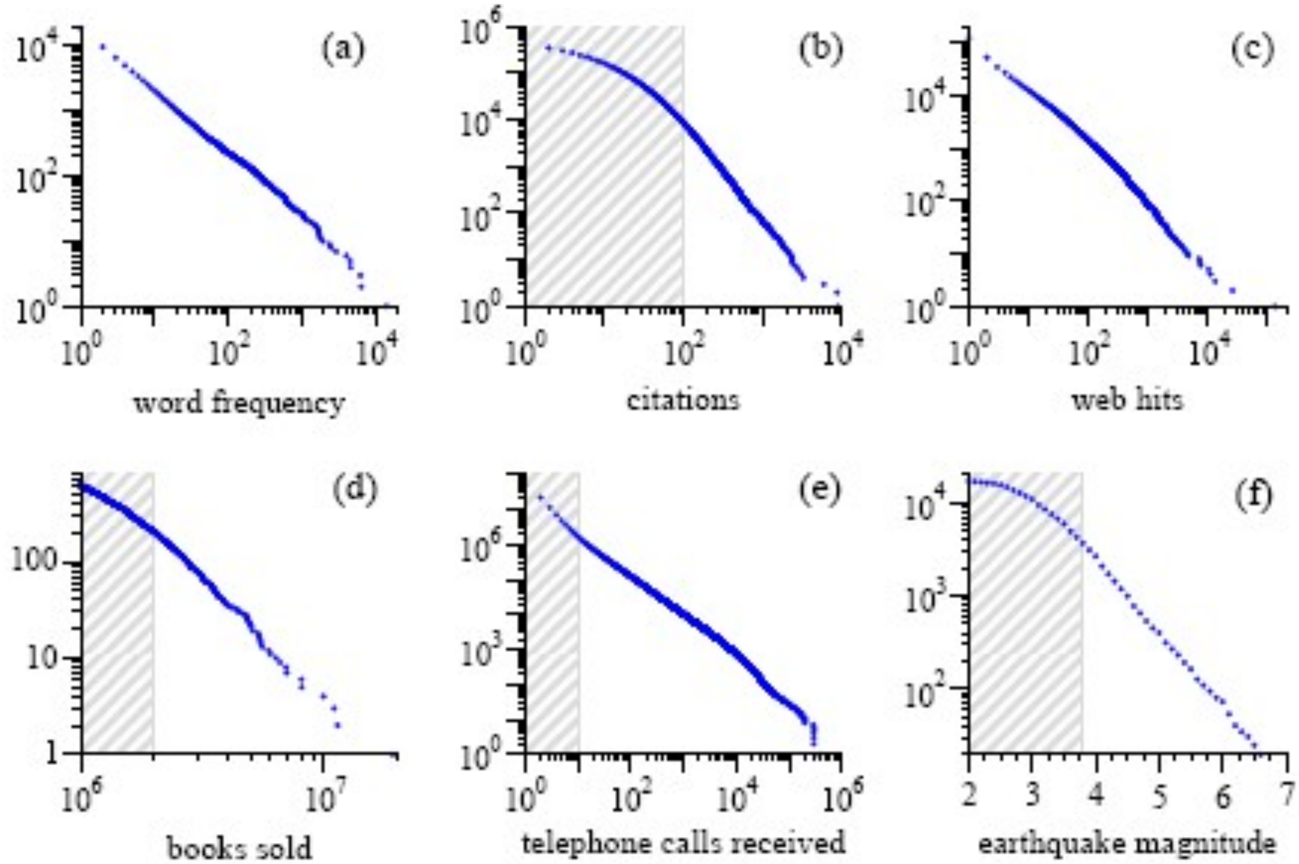


P(g) same as $P_{old}(n)$ and $P_{new}(n)$.

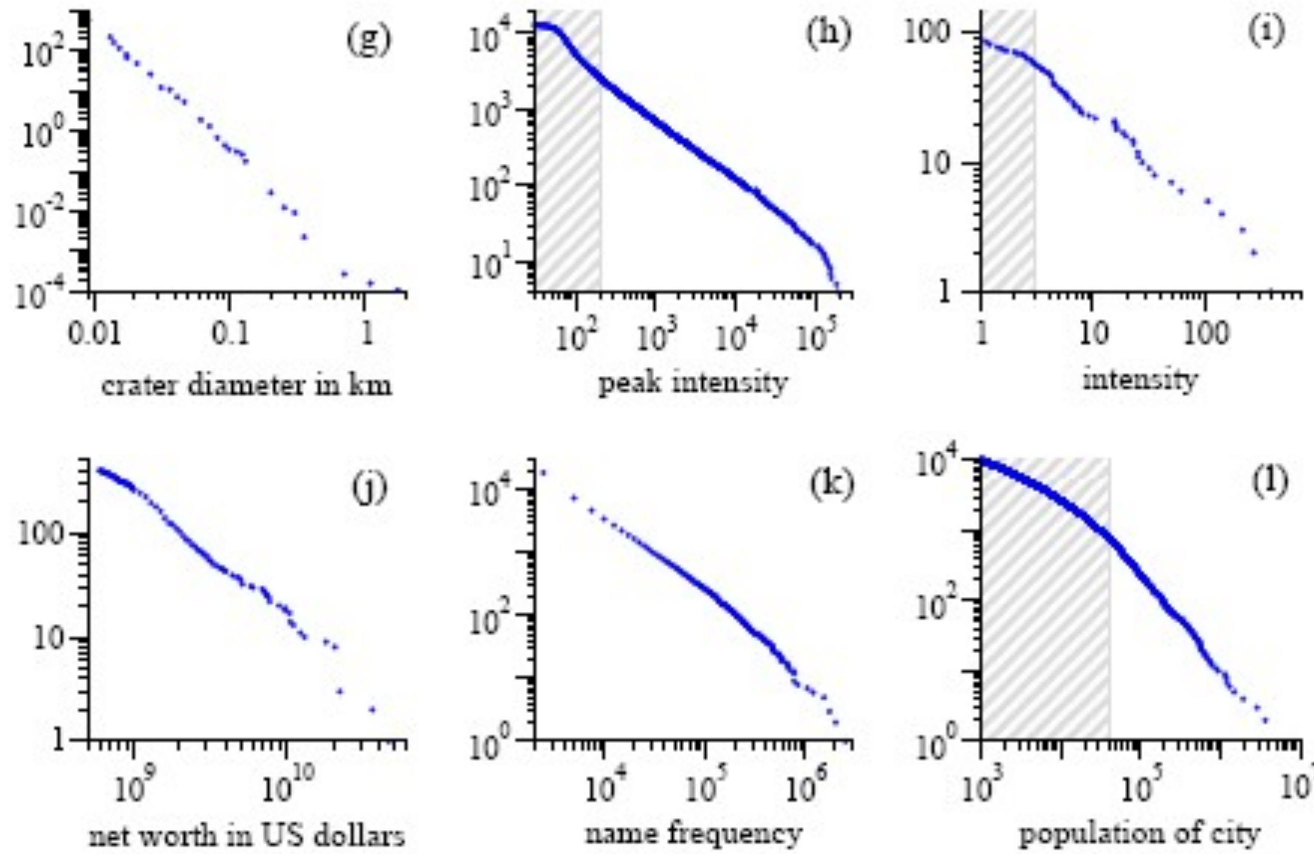
1. for small g,
   $P(g) \approx \exp[- |g| (2 / V_g)^{1/2}]$.

2. for large g, $P(g) \sim g^{-3}$.

Universality---6 examples



Universality---6 more examples



Behavior of the distribution for large time intervals

- Since

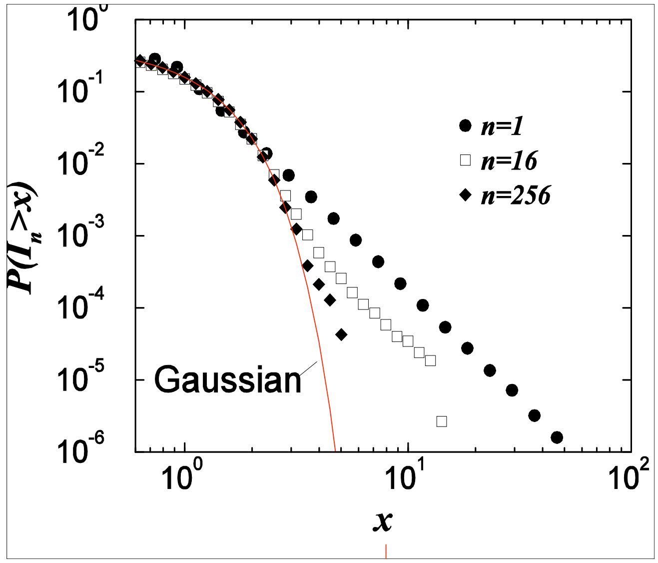
$$r_{N\Delta t} = \sum_{i=1}^{N} r_{\Delta t}^{i}$$

- We expect by Central Limit Theorem that P(R) for larger times to converge to a Gaussian

- Indeed a generated power law distribution with the same exponent ζ_R converges quickly to Gaussian under aggregation. Consider

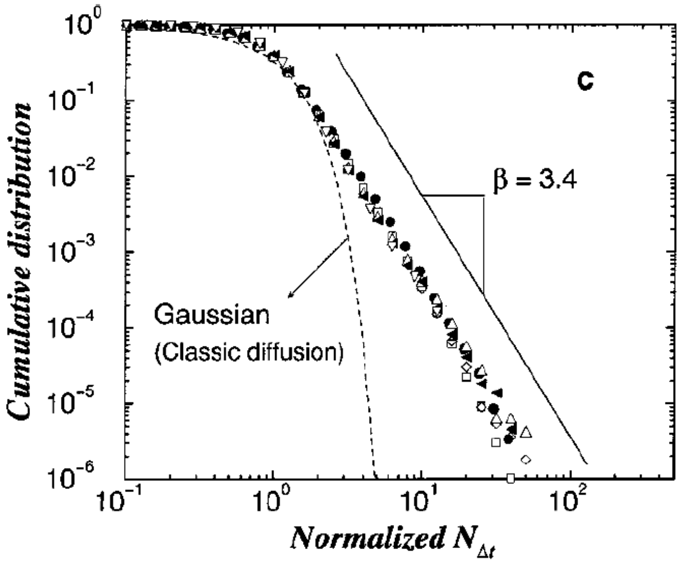
$$I_n = \sum_{i=1}^{n} x_i$$

Convergence of generated iid variables



Do large returns arise from large market activity ?

For this to be possible, since $R \sim \varepsilon\,\sigma\sqrt{N}$ we expect $\zeta_N = \dfrac{\zeta_R}{2}$



In sharp contrast, we find:

$$\zeta_N \approx 3$$

too large to explain

$$\zeta_R = 3$$

Fluctuations in market activity too mild to explain fat tails of returns.

Take home message

- P(growth rate) Laplace in Center: universal
- Width decreases as -1/6 power of size bin
- P(growth rate) crosses over to power law in wings
- No theory for -1/6 power law for width
- Theory (Buldyrev et al) for growth rate power law

http://polymer.bu.edu/hes (PDF of published papers)

Data analyzed (Gopikrisnan/Plerou/Liu/...)

Trades and Quotes (TAQ) database

- 2 years 1994-95
- 1000 stocks largest by market cap on Jan 1, '94 (200 million records)

To test "universality", also analyze other databases, including:

Center for Research in Security Prices (CRSP) database

- 35 years 1962-96
- approximately 6000 stocks

Tick data for the London Stock Exchange

- 2 yrs 2000-01
- 250 stocks.

Transactions data from the Paris Bourse

- 30 stocks; 1994-95

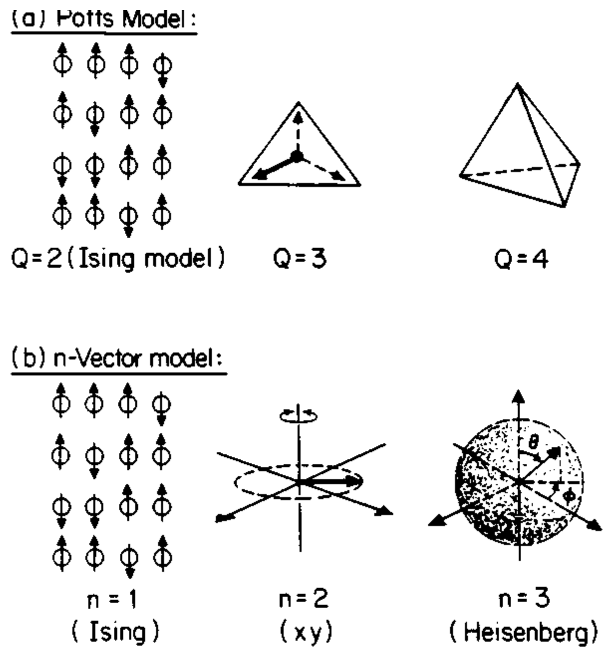
After-Dinner Drink: theory/model?

Each stock is a unit, interacting with other stocks (units). This type of model studied in statistical physics.

Typical models:

1. set of units,each of which can be in Q different states (POTTS MODEL).

2. set of n-dimensional units, each of which can be in a continuum of states (n-VECTOR MODEL)



(c)

Pillar 2, Universality (Universality classes):

Experimental fact:
A wide range of magnetic materials belong to one of two families of "Universality classes": the Q-state Potts model (Potts 1952) and the n-vector model (HES 1968). The purely geometric phase transition "percolation" corresponds to the limit Q=1, while the self-avoiding random walk corresponds to n=0.

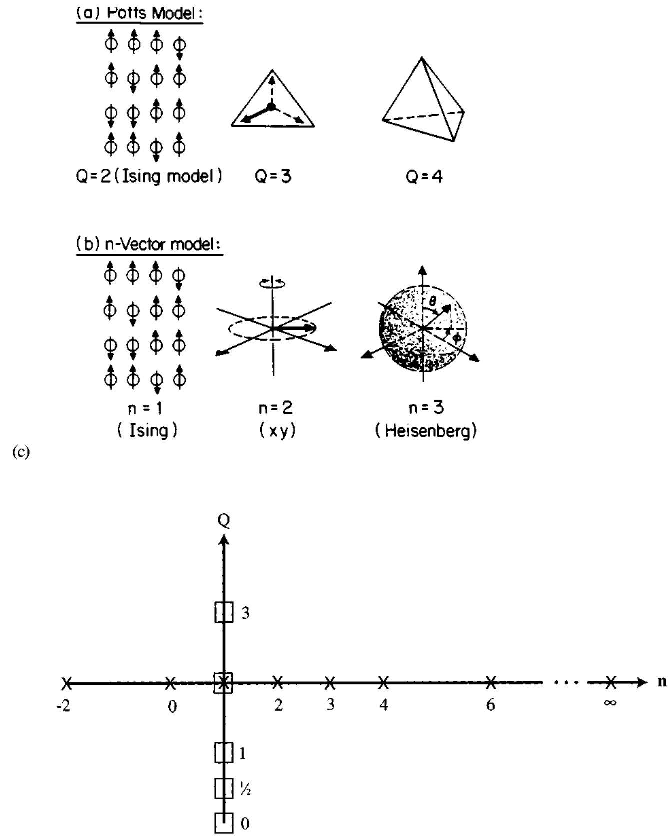


FIG. 2. Schematic illustrations of the possible orientations of the spins in (a) the *s*-state Potts model, and (b) the *n*-vector model. Note that the two models coincide when $Q=2$ and $n=1$. (c) North-south and east-west "Metro lines."

Pillar 1 (continued):

Experimental test of data collapse (Pillar 1): Equation of State for 5 different magnets near their respective critical points.

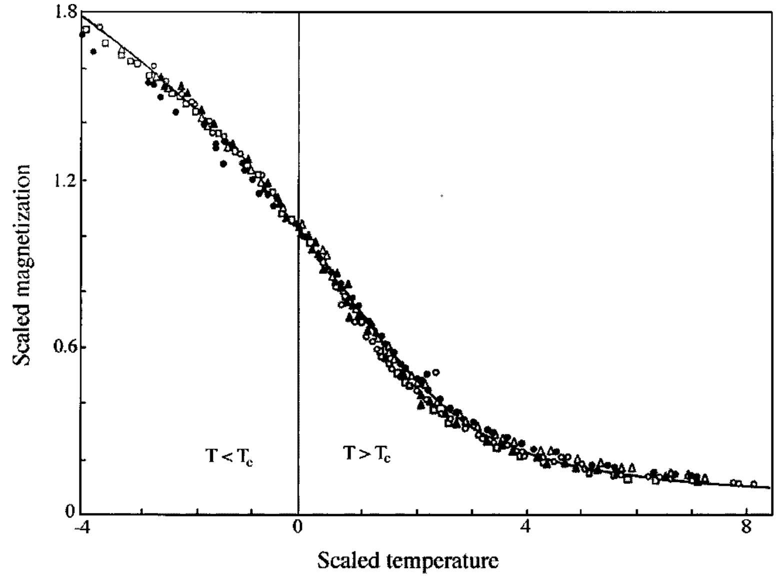
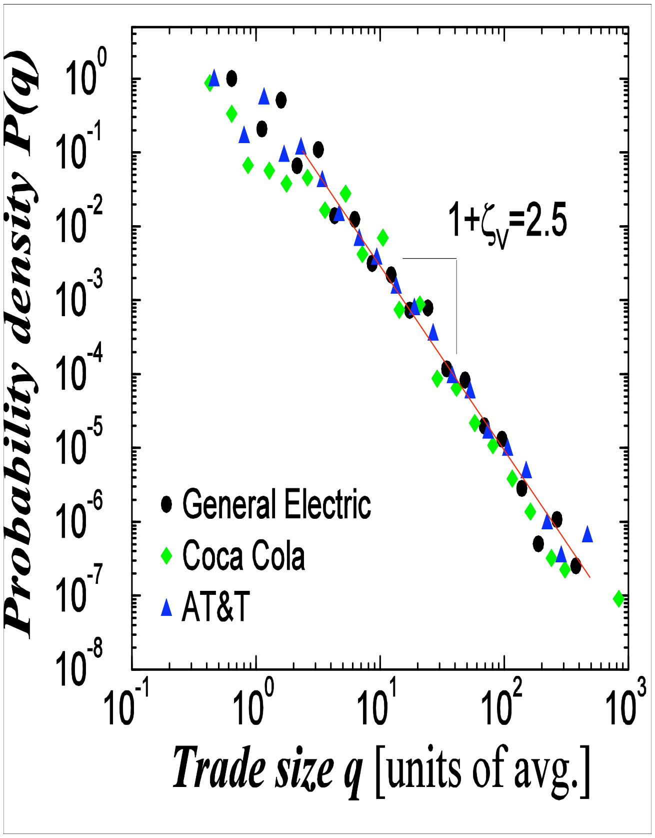


FIG. 1. Experimental MHT data on five different magnetic materials plotted in scaled form. The five materials are $CrBr_3$, EuO, Ni, YIG, and Pd_3Fe. None of these materials is an idealized ferromagnet: $CrBr_3$ has considerable lattice anisotropy, EuO has significant second-neighbor interactions. Ni is an itinerant-electron ferromagnet, YIG is a ferrimagnet, and Pd_3Fe is a ferromagnetic alloy. Nonetheless, the data for all materials collapse onto a single scaling function, which is that calculated for the $d=3$ Heisenberg model [after Milošević and Stanley (1976)].

Pillar 2: Universality

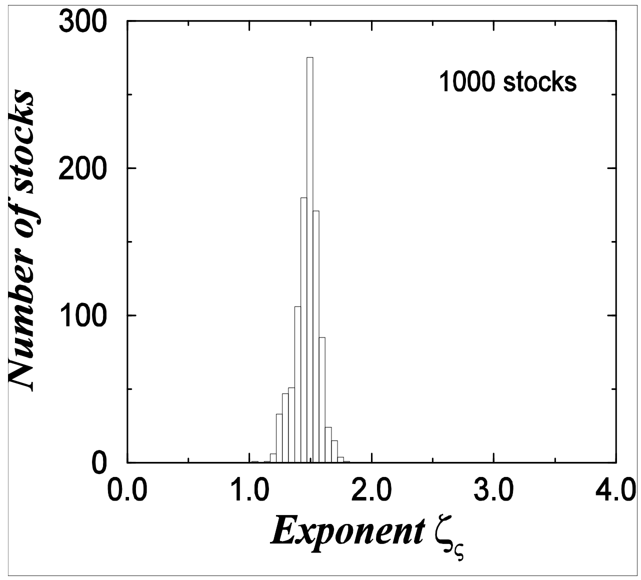
First hint: all 5 magnets have same scaled equation of state.

Statistics of Volume Traded

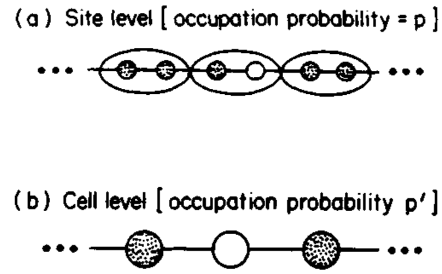


$$P(V > x) \sim x^{-\zeta_V}$$

$$\zeta_V \approx \frac{3}{2}$$

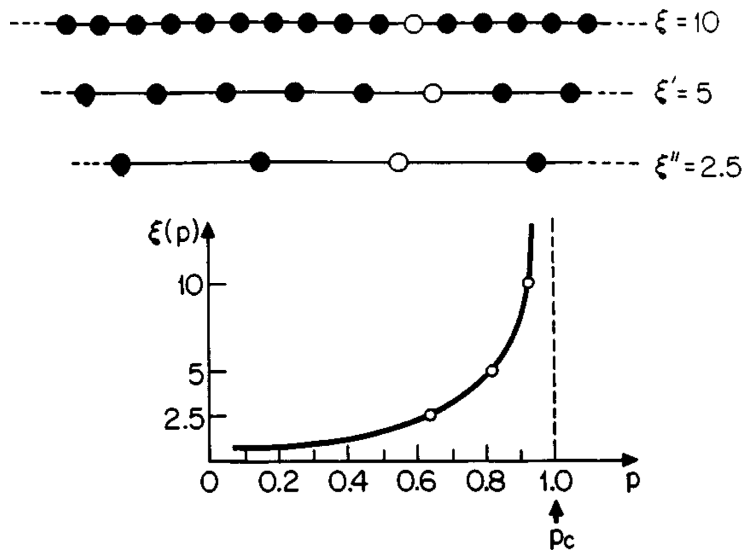


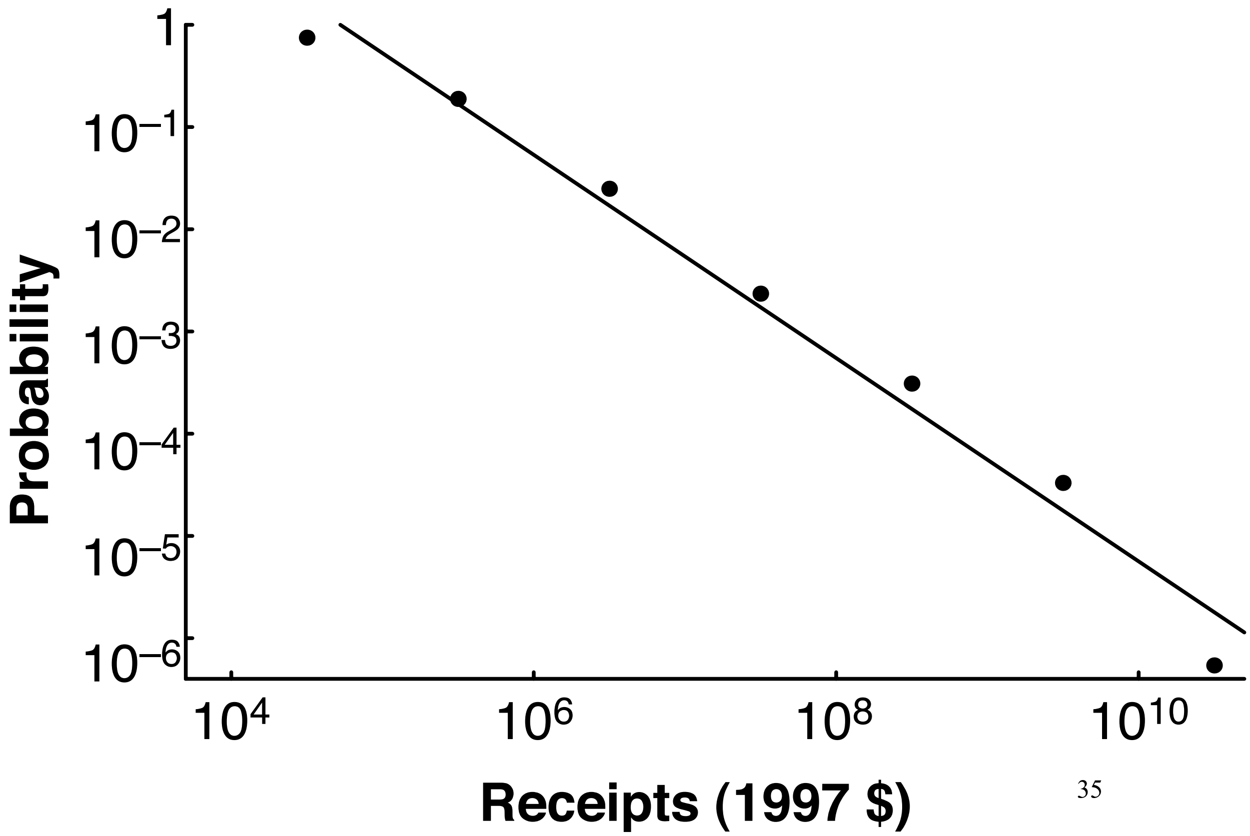
Pillar 3: RENORMALIZATION GROUP



Kadanoff site-to-cell coarse-graining successively tames the problem of an infinite correlation length.

Example: Q=1 Potts model for d=1 (1d percolation)





Probability
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{4}
10^{6}
10^{8}
10^{10}
Receipts (1997 $)