

Imiona dzieci, prawo Zipfa i mapa Stanów Zjednoczonych

Mateusz Pomorski¹, Małgorzata J. Krawczyk¹,
Jarosław Kwapien², Krzysztof Kułakowski¹,
Marcel Ausloos³

¹*Wydział Fizyki i Informatyki Stosowanej, AGH, al. Mickiewicza 30, 30-059 Kraków*

²*Instytut Fizyki Jądrowej PAN, ul. Radzikowskiego 152, 31-342 Kraków*

³*GRAPES, rue de la Belle Jardiniere, B-4031 Liège*



XLIII Zjazd
Fizyków Polskich

6-11 września 2015
Kielce

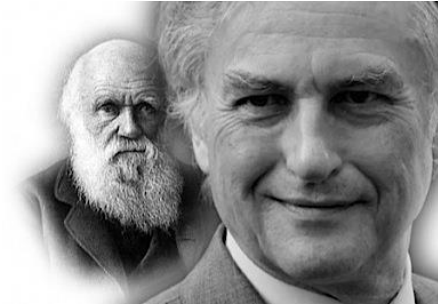
plan

- Motywacja + literatura
- Dane z USA i graf korelacji
- Klastry stanów
- Regiony USA
- Prawo Zipfa
- Dane z Belgii
- Podsumowania

Motywacja

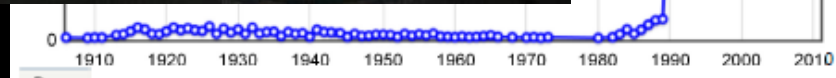
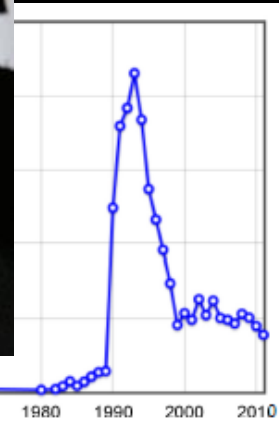
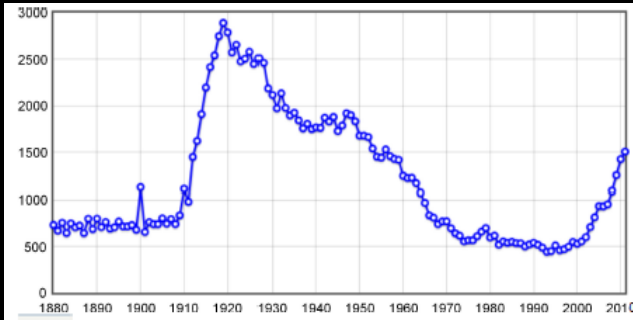
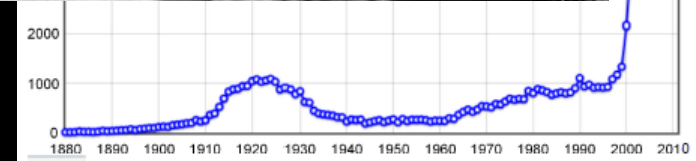
Darwinism is too big a theory to be confined to the narrow context of the gene.

Richard Dawkins



According to Baudrillard (...) consumers continuously (re-)construct their identities, in part, through the consumption of symbols, images, and signs.

Laurie A. Meamber

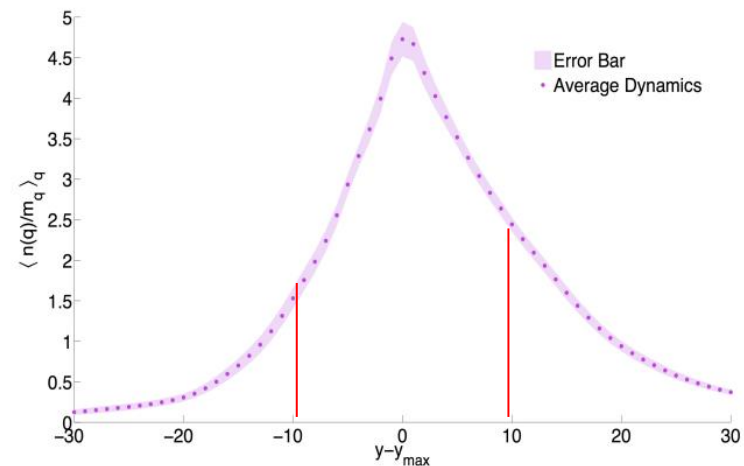


Marta Klekotko



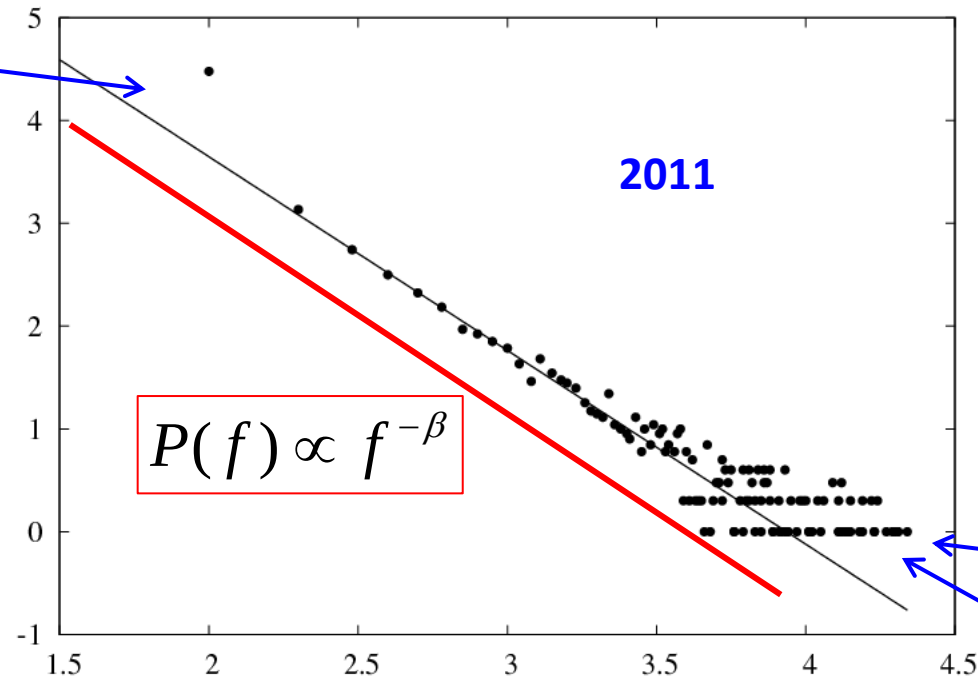
Literatura

- G. Simmel, *Fashion*, International Quarterly 10 (1904) 130 – moda jako naśladownictwo elit w sferze symboli statusu
- R. Pedone, R. Conte, *The Simmel effect: imitation and avoidance in social hierarchies*, LNAI 1979 (2000) 149 – algorytm symulacji w modelu Simmla
- Wentian Li, *Analyses of baby name popularity distribution in US for the last 131 years*, Complexity 18 (2012) 44 – dopasowanie prawa Zipfa do imion
- M. J. Krawczyk, A. Dydejczyk, K. Kułakowski, *The Simmel effect and babies names*, Physica A 395 (2014) 384 – symulacja dynamiki symboli wg Simmla
- P. Barucca, J. Rocchi, E. Marinari, G. Parisi, F. Ricci-Tersenghi, *Cross-correlations of American baby names*, PNAS 112 (2015) 7943
- dynamika korelacji między stanami



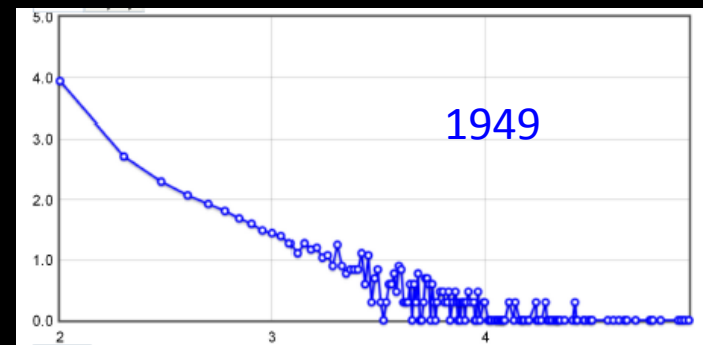
Histogram częstości imion

Aadhya
Aaditya
...



Sophia

Jacob

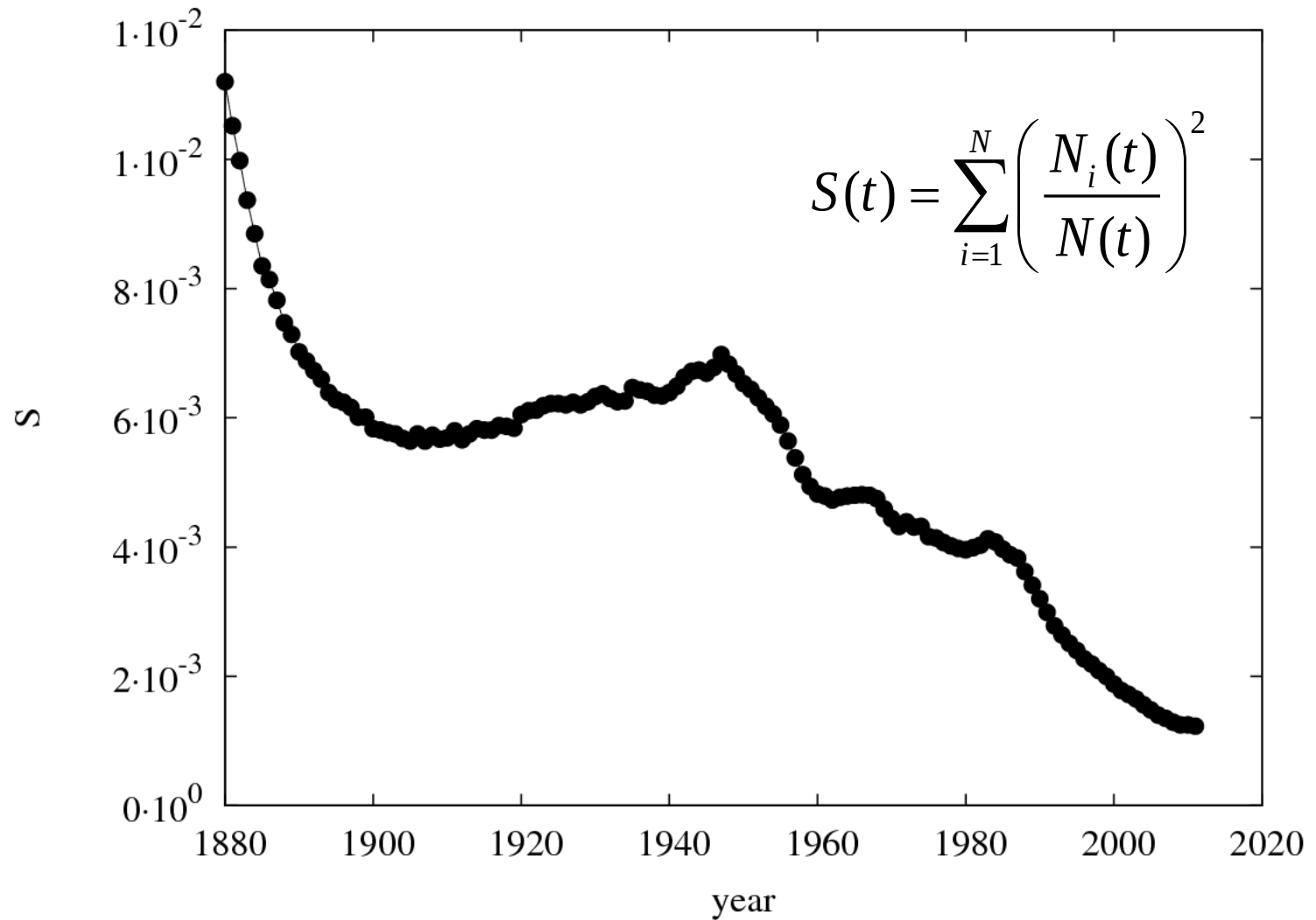


interpretacja β w tekście

- $\beta \cong 2$ - wartość typowa; $\beta \cong 3/2$ dla rzadszych słów w tekstach wieloautorskich
- $\beta > 2$ - wiele tematów, chaotyczne wypowiedzi; także zbyt małe próbki tekstu ; może też oznaczać lekką schizofrenię
- $\beta < 2$ - małe dzieci (1.6), teksty bojowe (1.7); także cięższa schizofrenia

[R. Ferrer i Cancho, Eur. Phys. J. B 44 (2005) 249]

Fragmentacja $S(t)$



dane z USA i graf korelacji

$n(i,t,a)$

www.ssa.gov

i,j – imię

a,b – stan

$t = 1910, \dots, 2011$

Korelacje ρ_{ij} między trochę innymi zmiennymi:

Barucca *et al*, 2015:

$$x(i,t,a) = \frac{n(i,t,a)}{\sum_j n(j,t,a)}$$

$i, j = 1, 19492$

ta praca:

$$y(i,t,a) = \frac{x(i,t,a)}{\sum_b x(i,t,b)}$$

$i, j = 1, 100$

klastry stanów - obliczenia

- Elementy macierzowe sieci stanów USA

$$w_{ab} = \frac{1 + \rho_{ab}}{2}$$

- Ewolucja macierzy *

$$\frac{dw_{ab}}{dt} = G(w_{ab}) \sum_c (w_{ac} w_{cb} - \beta)$$

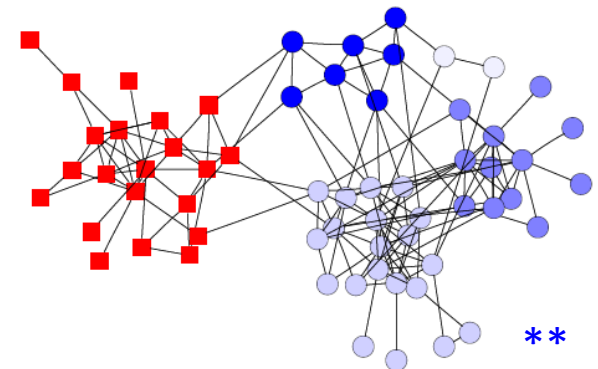
gdzie $G(x) = \Theta(x)\Theta(1-x)$

a β jest parametrem maksymalizującym modularność Q^{**} :

$$Q = \frac{1}{m} \sum_{ab} \left(w_{ab} - \frac{k_a k_b}{m} \right) \delta(a, b)$$

$$k_a = \sum_c w_{ac}$$

$$m = \sum_a k_a$$



* M. J. Krawczyk, PRE 77 (2008) 065701

** M. E. J. Newman, M. Girvan, PRE 69 (2004) 026113

test metody

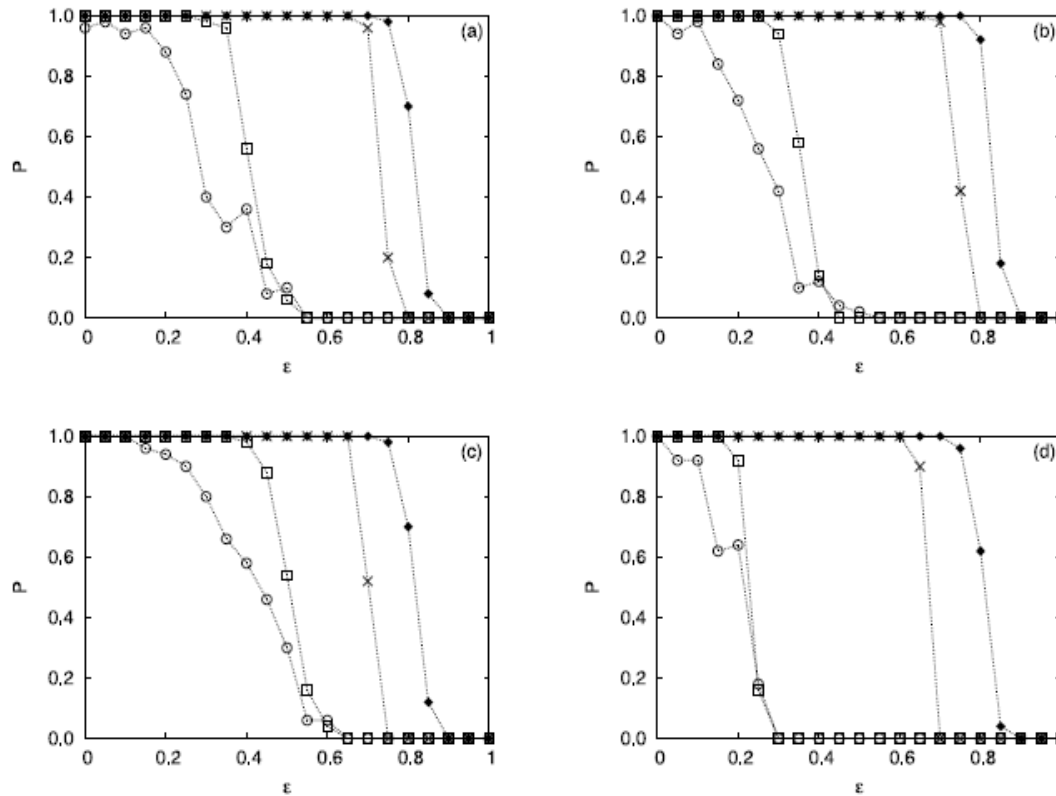
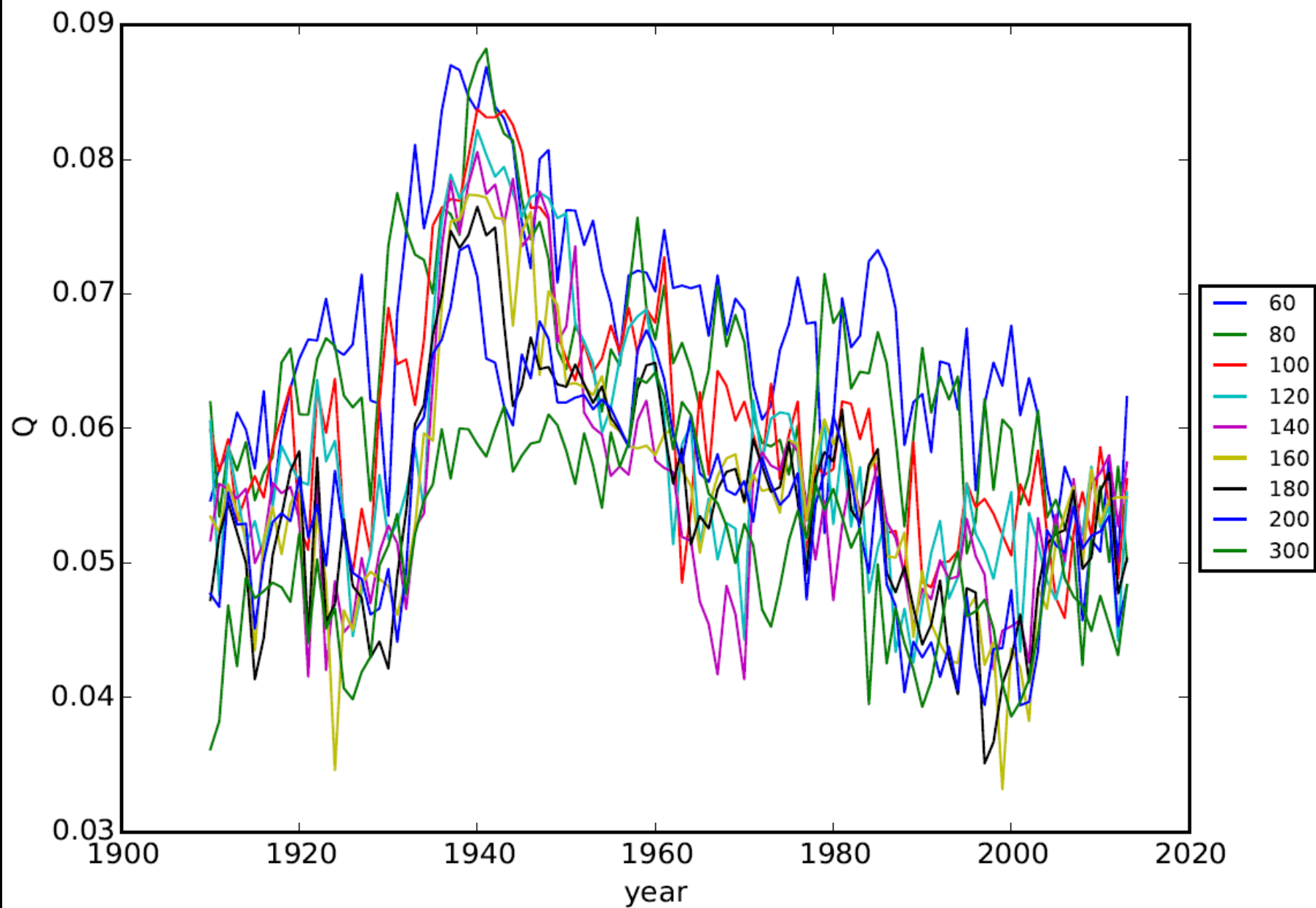


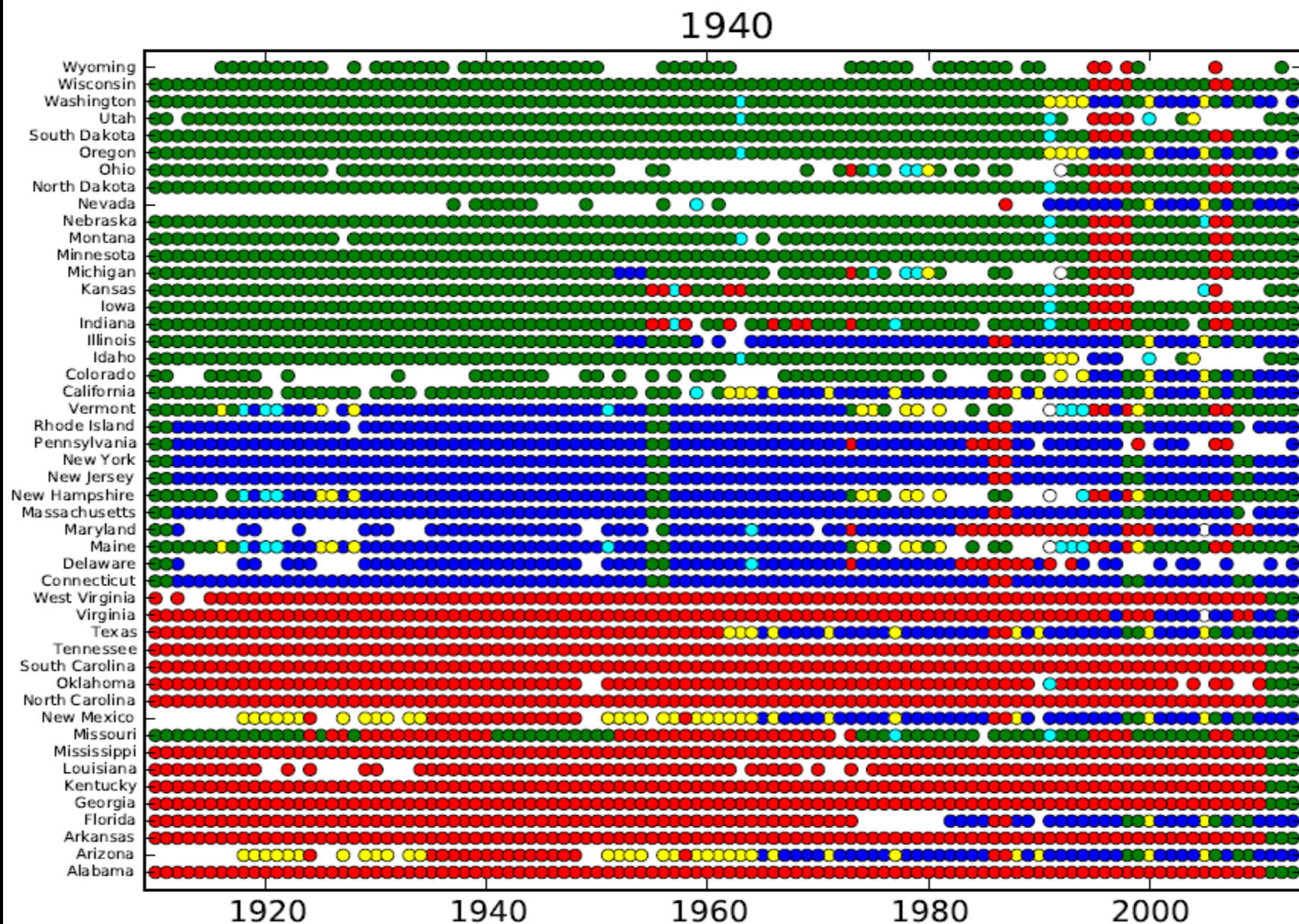
Fig. 2. Results for the network of size $N = 300$, clusters of: (a) 165, 93 and 42 nodes with $pz = 1/3$, (b) 175, 75 and 50 nodes with $pz = 1/3$, (c) 148, 117 and 35 nodes with $pz = 1/4$, (d) 191, 74 and 35 nodes with $pz = 1/4$ (symbols denote: $\circ$ - the Duch and Arenas method, $\square$ - the Newman method, $\times$ - Eq. (1) with $\beta = 0.25$, $\blacklozenge$ - Eq. (1) with $\beta = 0.4$).

[Krawczyk M.J., *PRE* 77 (2008) 065701(R); *Comp. Phys. Comm.* 181 (2010) 1702]

Q(t)



klastry stanów - wyniki



Klasy stanów - wyniki

Region 1: Northeast		
Division 1: New England Connecticut (09) Maine (23) Massachusetts (25) New Hampshire (33) Rhode Island (44) Vermont (50)	9	Division 2: Middle Atlantic New Jersey (34) New York (36) Pennsylvania (42)
Region 2: Midwest*		
Division 3: East North Central Indiana (18) Illinois (17) Michigan (26) Ohio (39) Wisconsin (55)	12	Division 4: West North Central Iowa (19) Kansas (20) Minnesota (27) Missouri (24) Nebraska (31) North Dakota (38) South Dakota (46)
Region 3: South		
Division 5: South Atlantic Delaware (10) District of Columbia (11) Florida (12) Georgia (13) Maryland (14) North Carolina (37) South Carolina (45) Virginia (51) West Virginia (54)	16	Division 6: East South Central Alabama (01) Kentucky (21) Mississippi (28) Tennessee (47)
		Division 7: West South Central Arkansas (05) Louisiana (22) Oklahoma (40) Texas (48)
Region 4: West		
Division 8: Mountain Arizona (14) Colorado (08) Idaho (16) New Mexico (35) Montana (30) Utah (49) Nevada (32) Wyoming (56)	11	Division 9: Pacific Alaska (02) California (09) Hawaii (45) Oregon (41) Washington (53)

- wszystkie 9 stanów *Northeast* są przypisane do wspólnego klastra;
- 18 stanów *Midwest + West* należą do wspólnego klastra. Jednak status Arizony, Kalifornii, Missouri, Nowego Meksyku i Nevady zmienia się w czasie, więc ich przynależność pozostaje otwarta;
- 13 stanów *South* należą do wspólnego klastra. Jednak status Teksasu zmienia się w czasie. Poza tym Delaware i Maryland przypisano do *NorthEast*.

PACIFIC

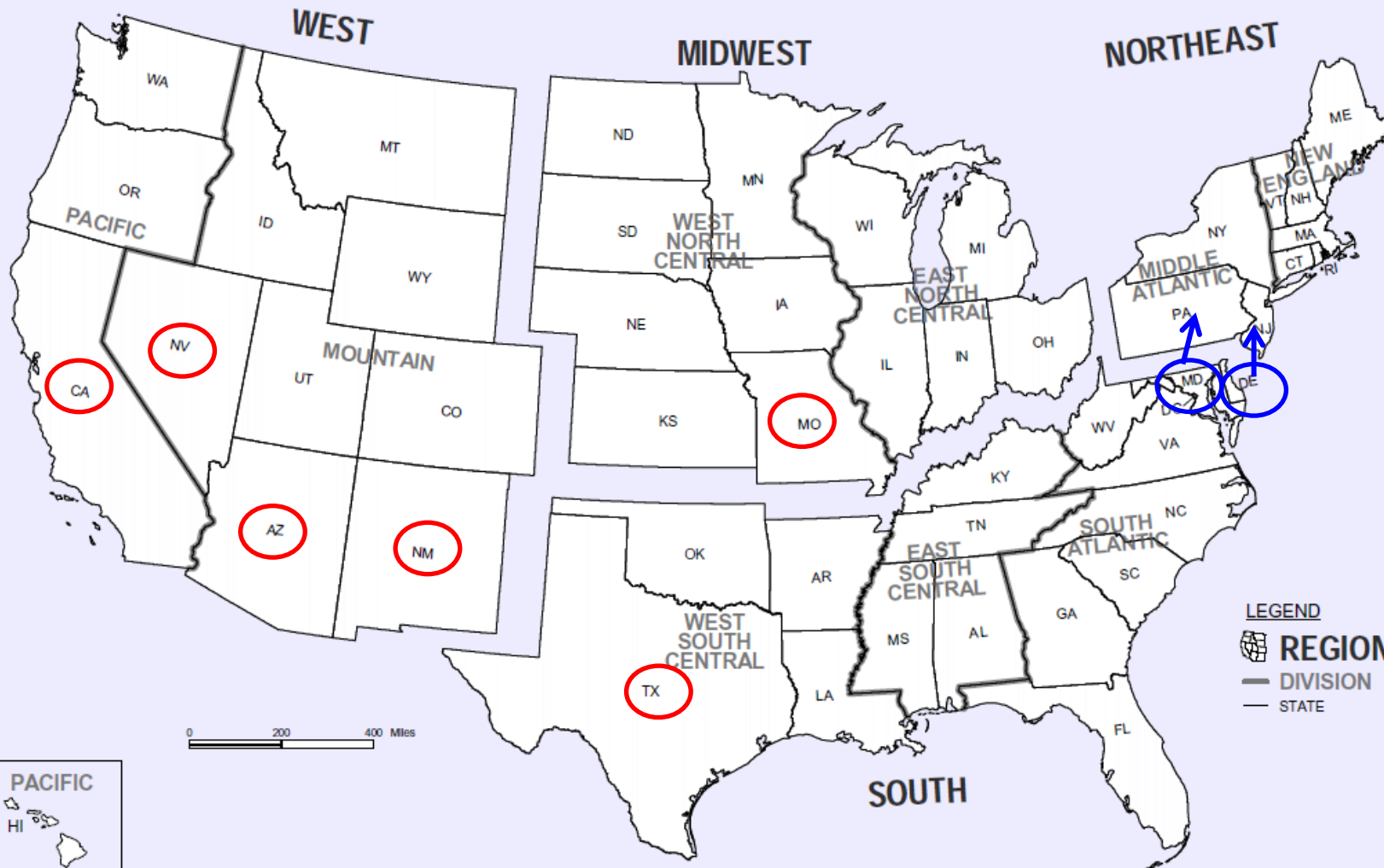


Census Regions and Divisions of the United States

WEST

MIDWEST

NORTHEAST



LEGEND



REGION

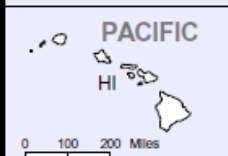


DIVISION

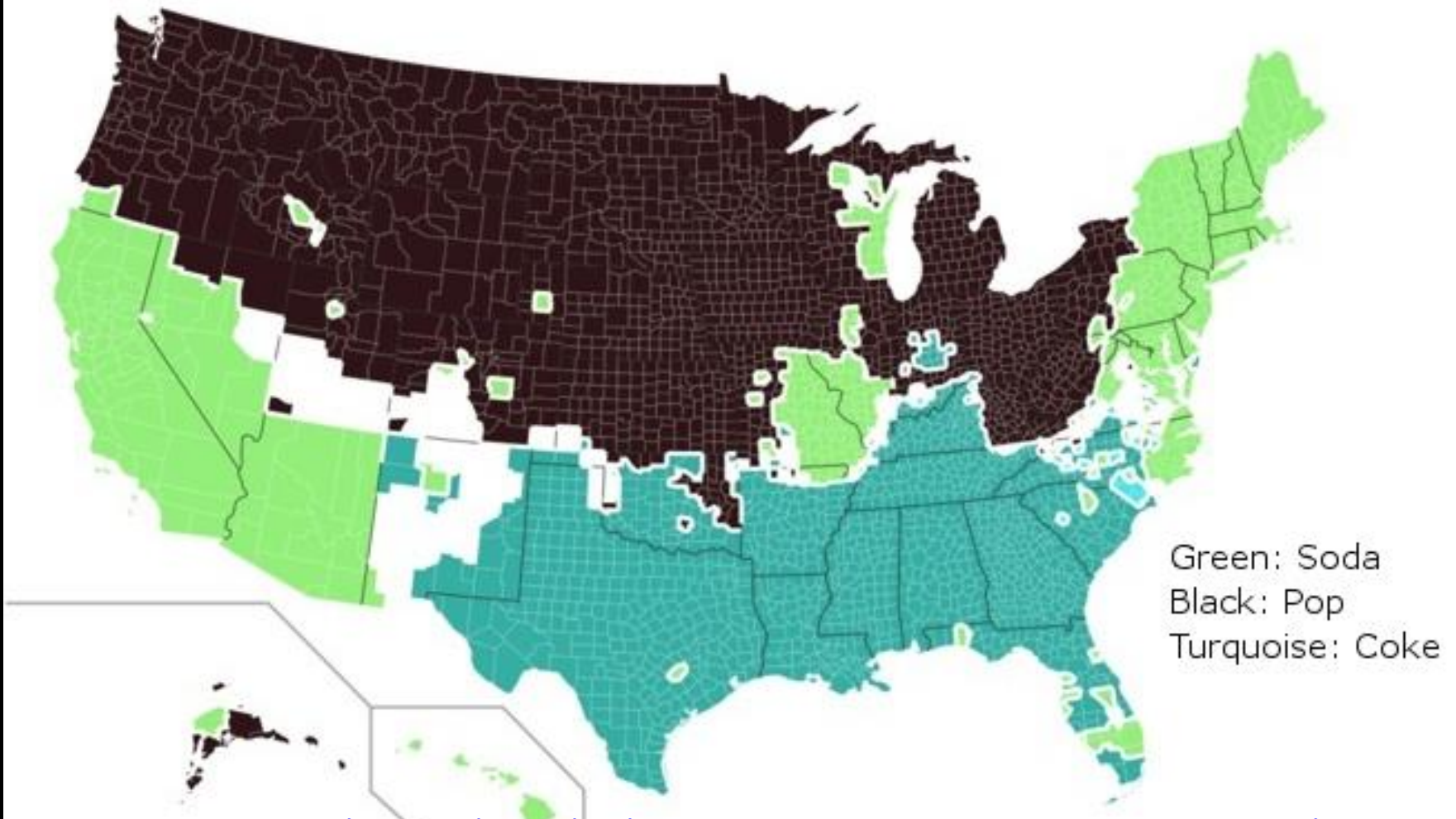


STATE

PACIFIC

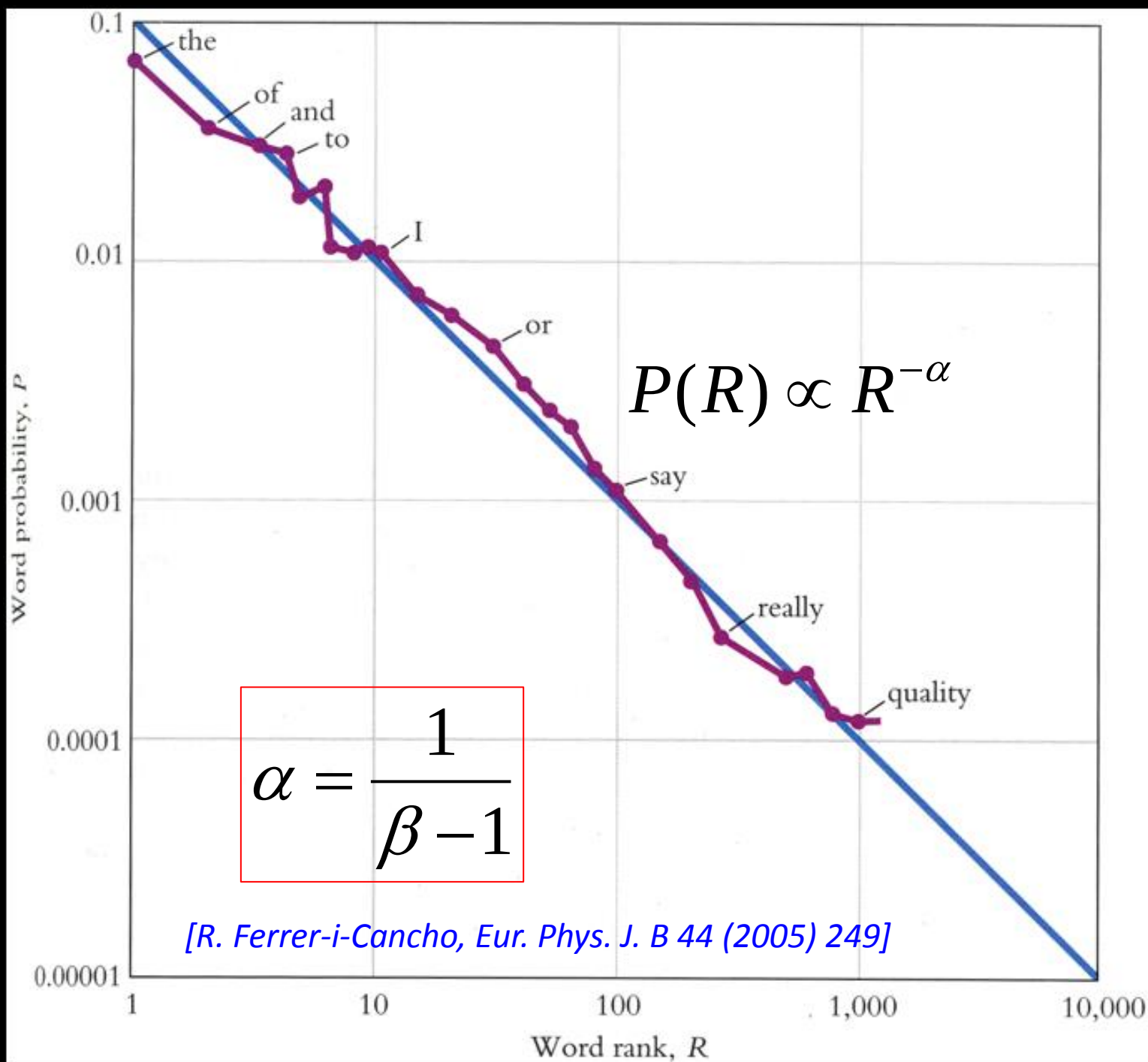


The Soft Drink Borders: Soda, Pop, or Coke?



Prawo Zipfa

www.prismnet.com/~dierdorf/ww-27.html



Uzasadnienie związku α i β

$m(n) = N P(n)$ – ilość imion o populacji n

$$m(n) \propto n^{-\beta} \qquad n(R) \propto R^{-\alpha}$$

Pozycja R w rankingu = ilość imion o większej m

$$R(n) = \int_n^{\infty} m(n') dn' \propto n^{1-\beta} \qquad \text{ale} \qquad R \propto n^{-1/\alpha}$$

Porównując, mamy
$$\frac{1}{\alpha} = \beta - 1$$

[R. Ferrer-i-Cancho, R. V. Sole, Journal of Quantitative Linguistics 8, 165 (2001)]

Co oznacza α ?

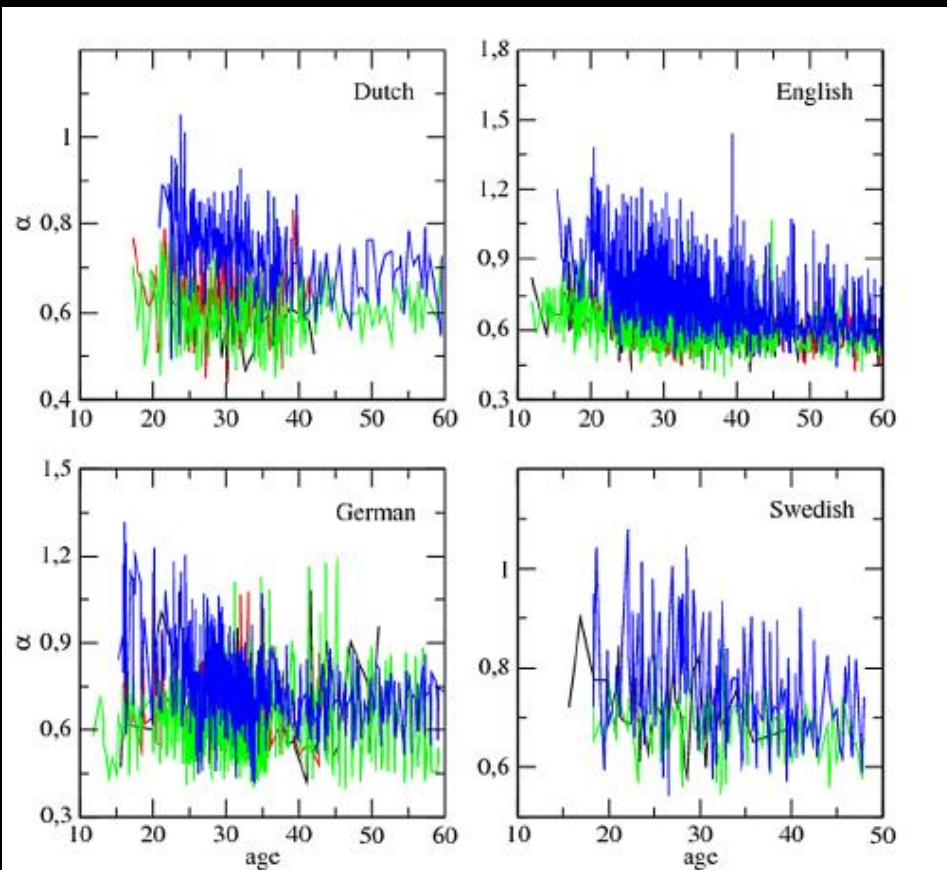


Figure 2. The evolution of the exponent α versus child age (in months): $n^* = 100$. The major classes of roles, i.e. target children (blue), mothers (green), investigators (red) and fathers (black), are shown. Length normalization by prefix with $n^* = 100$ is used. Swedish lacks the class 'investigator'.
doi:10.1371/journal.pone.0053227.g002

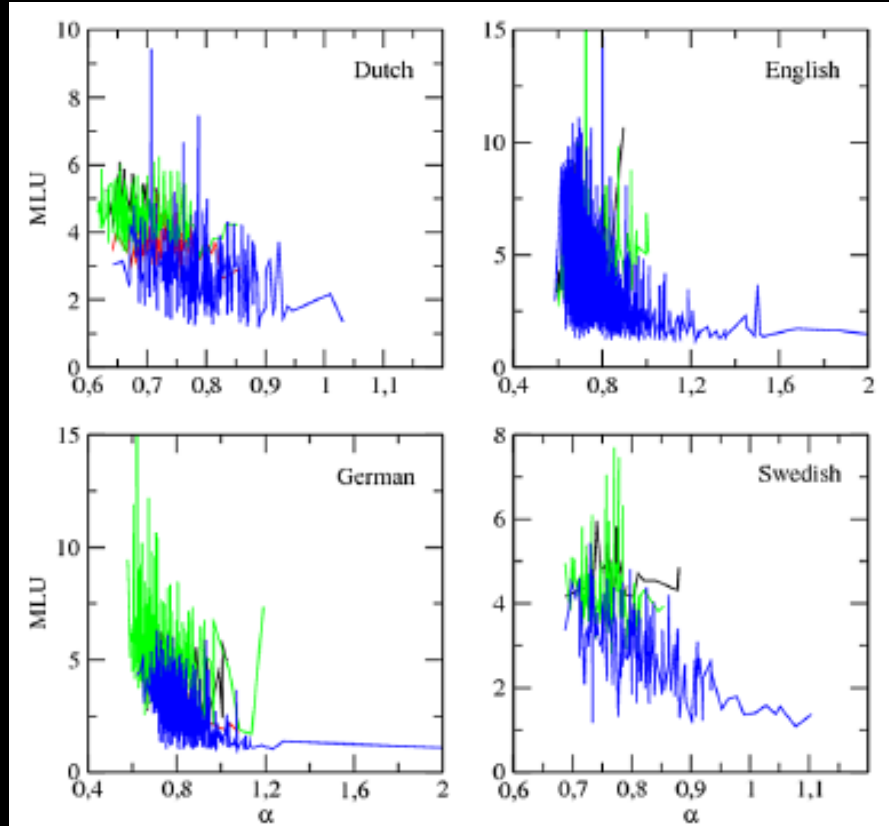
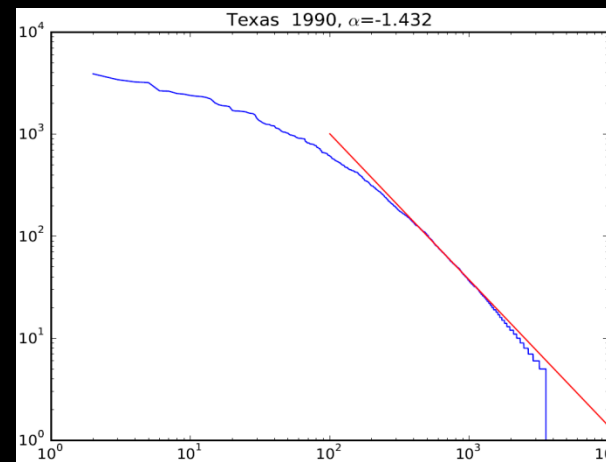
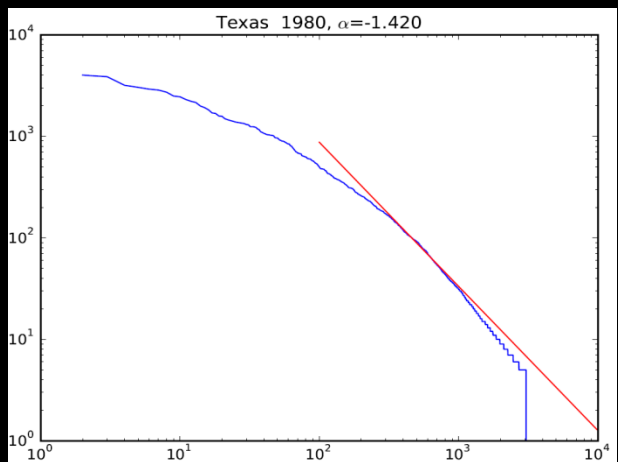
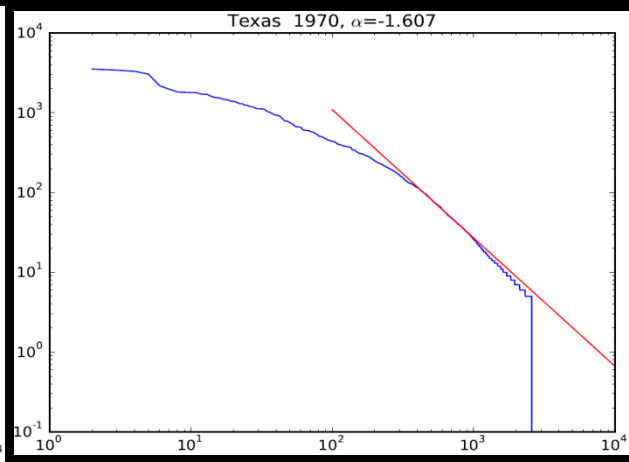
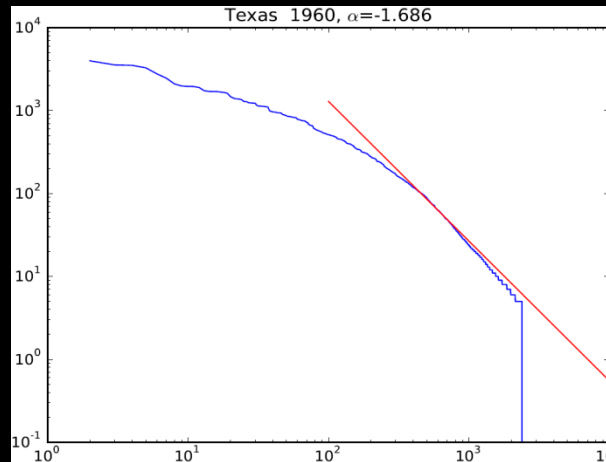
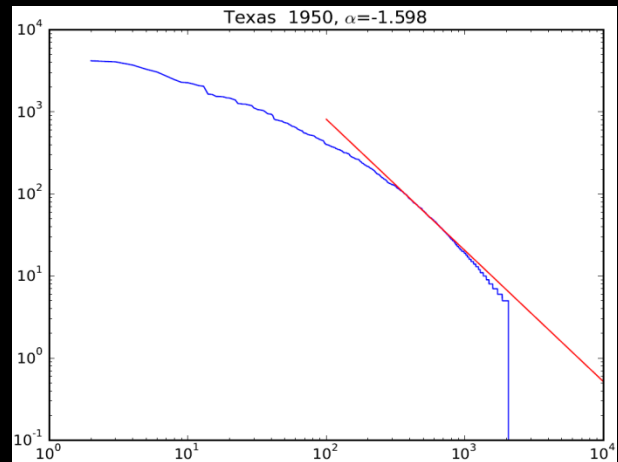
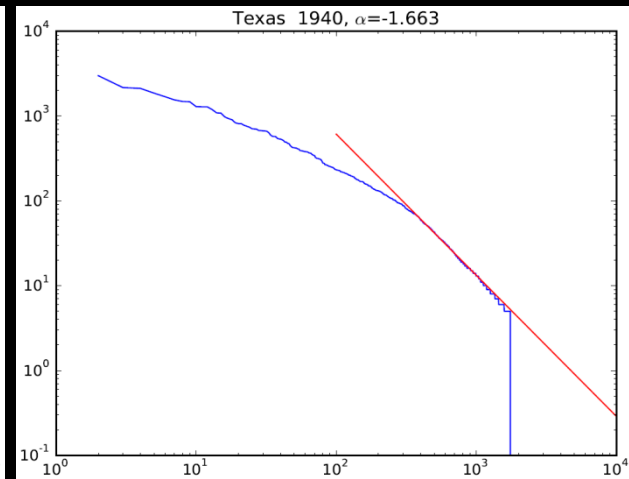
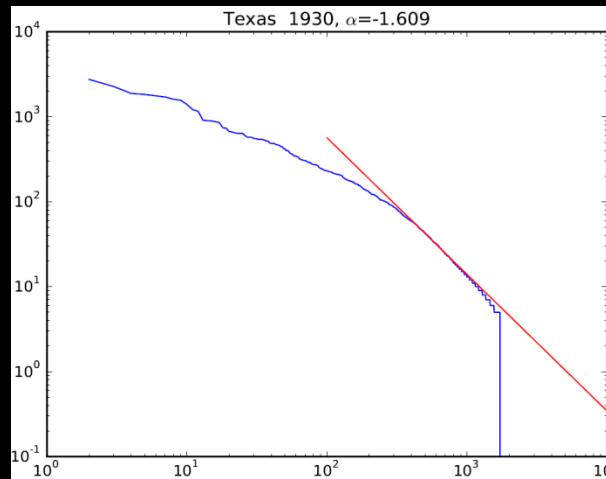
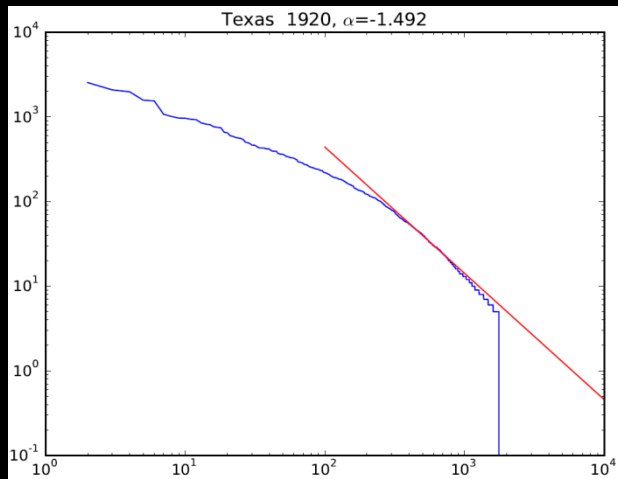


Figure 4. The MLU (in words) versus α : $T^* = 500$. The major classes of roles, i.e. target children (blue), mothers (green), investigators (red) and fathers (black), are shown. Length normalization by prefix with $T^* = 500$ is used. Swedish lacks the class 'investigator'. In order to facilitate the visual inspection of the series, the few points with MLU above 15 or α above 2 are not shown (this concerns English and German).
doi:10.1371/journal.pone.0053227.q004

Dane z Belgii – wykładnik α

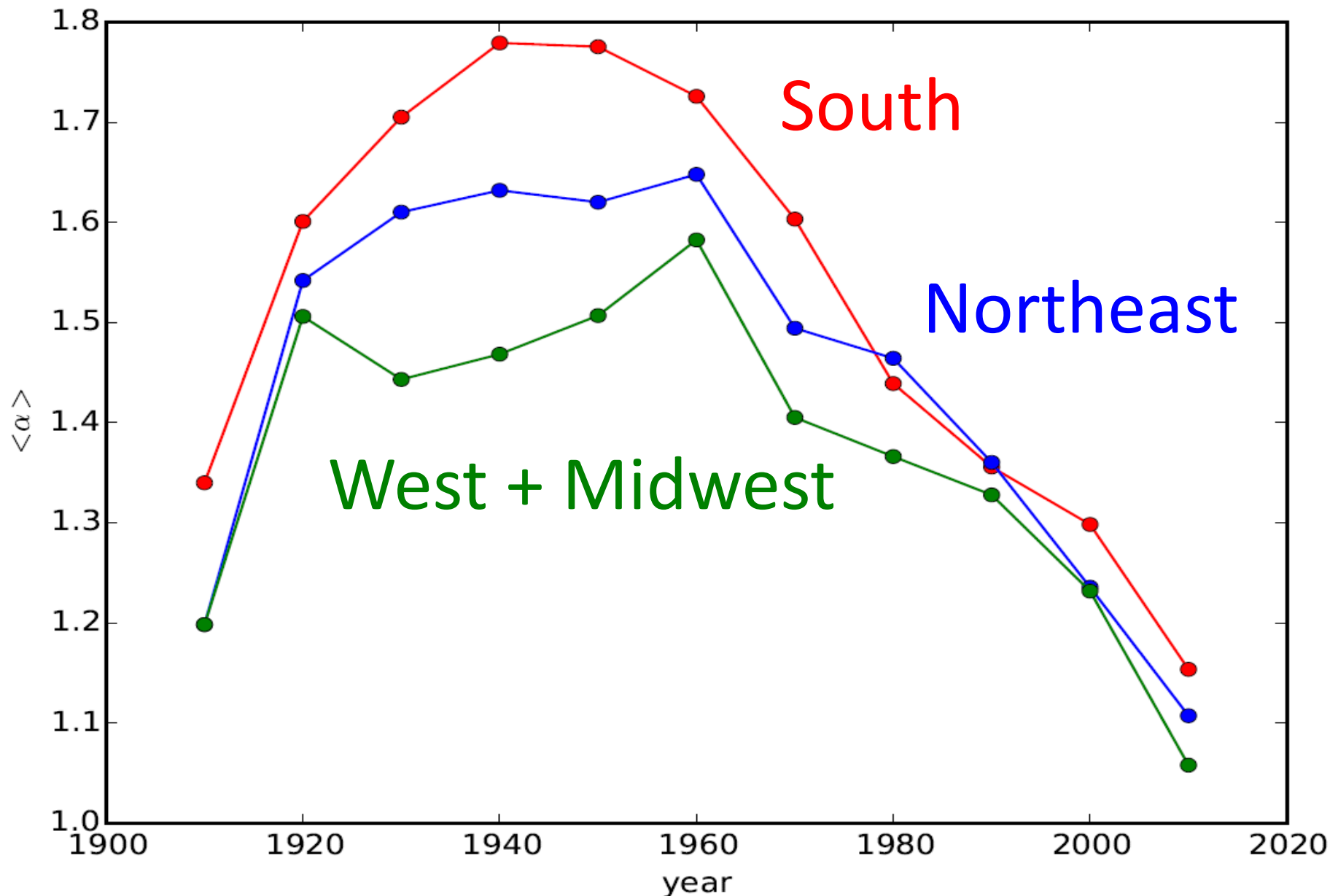
	males	females
Belgium	1.62	1.67
Flanders	1.67	1.57
Wallonia	1.60	1.69
Brussels	1.36	1.37
Belgium < 18	1.42	1.54
Belgium 18 – 64	1.65	1.72
Belgium > 65	1.85	1.83
Flanders < 18	1.41	1.33
Flanders 18 – 64	1.65	1.60
Flanders > 65	2.18*	2.03
Wallonia < 18	1.53	1.42
Wallonia 18 – 64	1.56	1.64
Wallonia > 65	1.58	1.66
Brussels < 18	1.41*	1.30*
Brussels 18 – 64	1.37	1.33
Brussels > 65	1.32	1.40





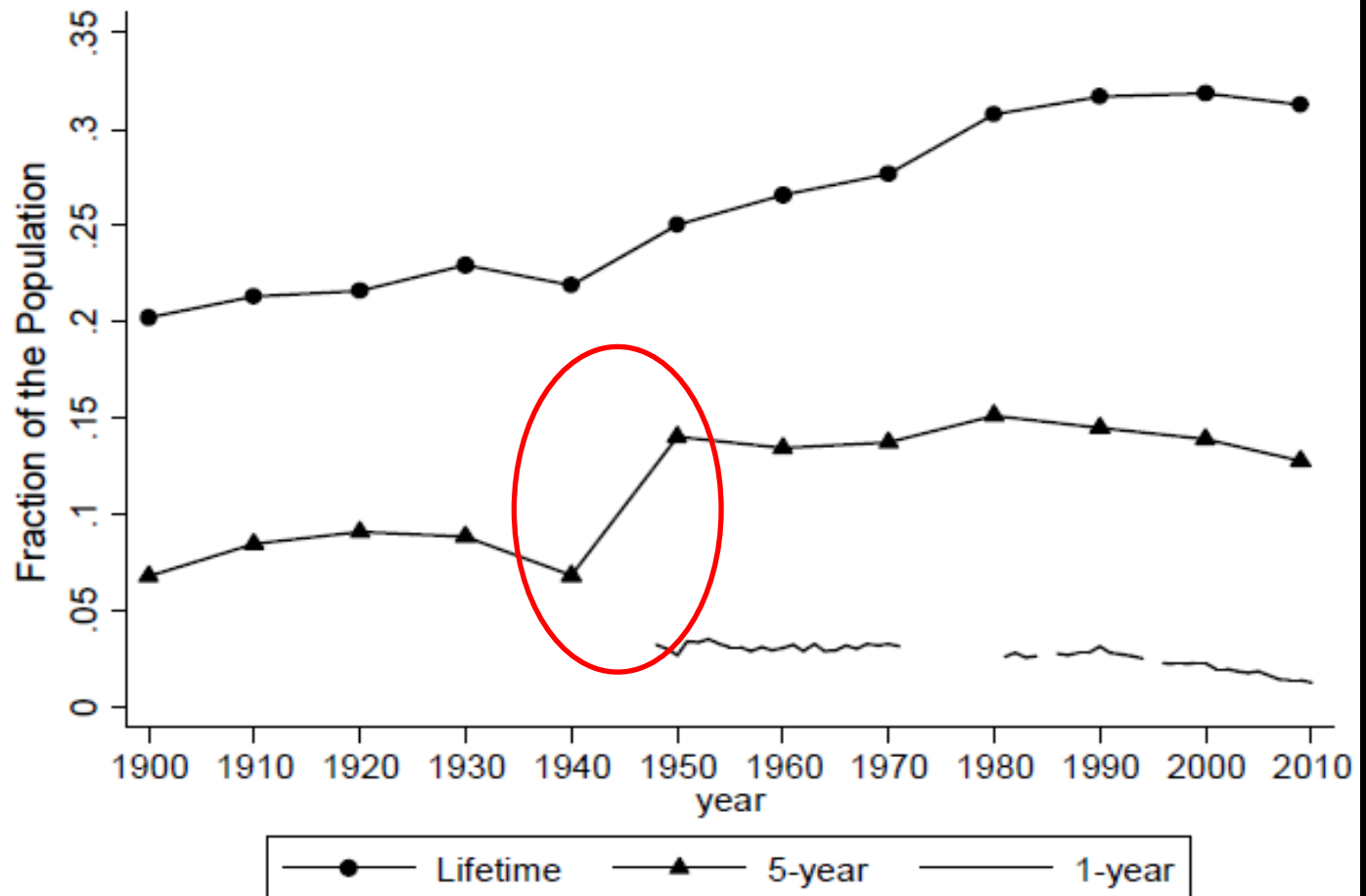
Dopasowanie danych
do prawa Zipfa
dla stanu Texas
1920-1990
home.agh.edu.pl/~gos/ZIPF/

Ewolucja α w klastrach stanów USA



Przyczyny : migracja?

Figure 2
Inter-State Migration Rates Since 1900



Note. Lifetime and 5-year migration rates are from the decennial Census 1900-2000 and from the ACS for 2009. Five-year migration rates are estimated from the fraction of households with a 4- or 5-year old residing outside of their birth state (Rosenbloom and Sundstrom 2004). Annual migration rates are from the Current Population Survey.

podsumowanie

1. Korelacje między imionami nadawanymi w stanach USA odzwierciedlają administracyjną i kulturową strukturę regionalną kraju.
2. Kolejność wartości wykładnika α w prawie Zipfa w otrzymanych klastrach stanów w latach 1920-1980 pozostaje niezmienną:
$$\alpha(\text{South}) > \alpha(\text{Northeast}) > \alpha(\text{West+Midwest}).$$
Interpretacja wykładnika α jest niejasna.

podsumowanie

1. Korelacje między imionami nadawanymi w stanach USA odzwierciedlają administracyjną i kulturową strukturę regionalną kraju.
2. Kolejność wartości wykładnika α w prawie Zipfa w otrzymanych klastrach stanów w latach 1920-1980 pozostaje niezmienną:
$$\alpha(\text{South}) > \alpha(\text{Northeast}) > \alpha(\text{West+Midwest}).$$

podziękowanie

Pani Prof. Dorocie Praszałowicz za dyskusję
Panu Andrzejowi Mleczko za zgodę na użycie plakatu
Państwu za uwagę 😊