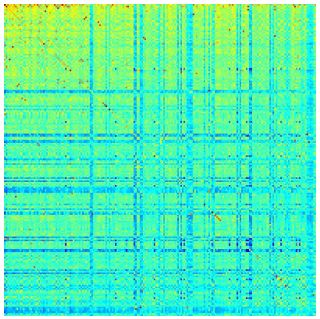
PageRank model of opinion formation and Google Matrix Analysis of DNA Sequences

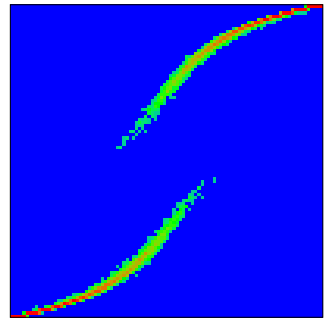
Vivek Kandiah and Dima Shepelyansky

Laboratoire de Physique Théorique, IRSAMC, UMR 5152 du CNRS
Université Paul Sabatier, Toulouse

Supported by EC FET open project NADINE

14 november 2013





Overview

Opinion formation :

- PROF model on webpages network and LiveJournal
- Sznajd model on webpages network

DNA sequences :

- Statistics of Google matrix elements : similarities and differences with WWW.
- Spectrum and PageRank
- PageRank correlations : statistical similarity between species.

Elector networks

- Many studies of opinion formation on regular lattices (voter model, Sznajd model, etc.)
- Real social networks show small-world and scale-free properties
- PageRank is an efficient ranking technique and provides a natural order of importance in a network
- PageRank top nodes represent the elite among the social network

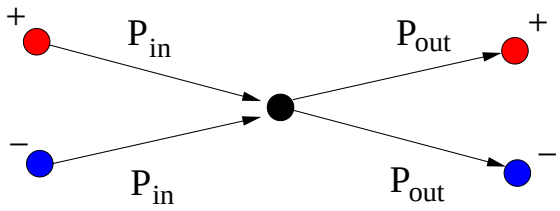
Idea : One's opinion is influenced by the closest members (friends) among the society and influential friend's opinion count more than less important friend's opinion in our environment.

Implementation : Two possible opinions coded by Ising spin variables σ_i taking values 1 or -1. We choose an initial distribution of opinions on the network and observe how the fraction of nodes having the same opinion evolves during time according to a certain rule.

Holley and Liggett (1975)
Sznajd-Weron (2000,2002,2004,2005)

Krapivsky,Redner,Ben-Naim (2010)

PageRank Opinion Formation Model (PROF)



$$\Sigma_i = a\sum_j P^+_{j,in} + b\sum_j P^+_{j,out} - a\sum_j P^-_{j,in} - b\sum_j P^-_{j,out} \quad , \quad a+b=1$$

P_j : PageRank of node j

- defined for one iteration step
- σ_i takes the value 1 or -1 respectively for $\Sigma_i > 0$ or $\Sigma_i < 0$.
- The parameters a and b allow to tune the importance of incoming and outgoing links.

Large b $\rightarrow$ an elector takes the opinion of people he is looking at $\rightarrow$ "conformist" society

Large a $\rightarrow$ an elector takes mainly the opinion of people pointing to him $\rightarrow$ "tenacious" society

Features of society described by PROF model

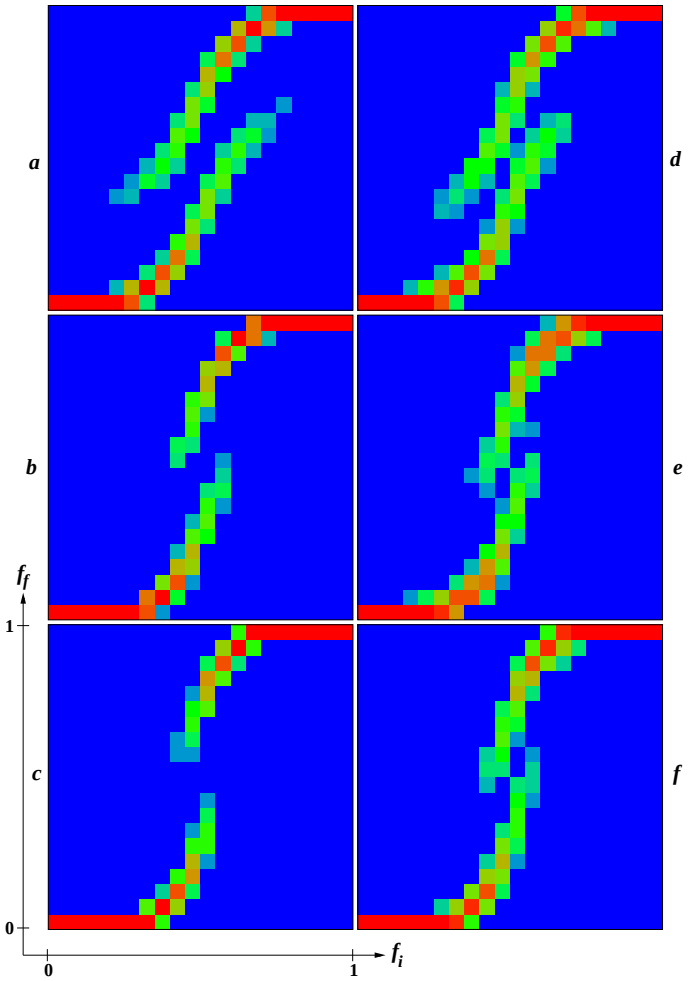


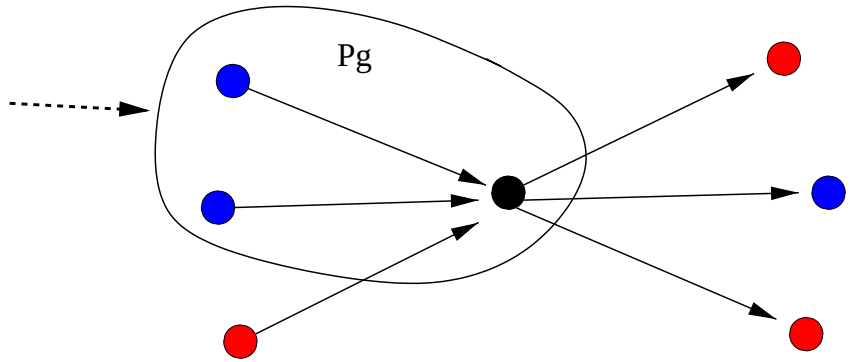
Figure: Density plots of probability to find a final fraction f_f depending on initial fraction f_i of red nodes for Cambridge (left) and Oxford (right). $N_r = 10^4$ random realizations (up to convergence time t=20 iterations) were used on a $20x20$ cells grid. From top to bottom $a = 0.1$,$a = 0.5$ and $a = 0.9$.

- small fraction of red opinion suppressed/larger fraction dominates
- range of bistability phase, wider for low a

A tenacious society has a relatively small range of bistability phase unlike the conformist society where the opinion is strongly influenced by elite. random initial distribution of opinion $\rightarrow$ divided elite $\rightarrow$ divided followers $\rightarrow$ large bistable region

pratically no bistability in LiveJournal and Twitter

PROF-Sznajd model



A group point of view describing the famous principle :

"United we stand, divided we fall"

- pick a random node $\rightarrow$ polarization of N_g -1 highest PageRank nodes pointing to it ?
- if they have the same polarization $\rightarrow$ group with effective PageRank $P_g = \sum_{j=1}^{N_g} P_j$
- consider all nodes pointing to any member of the group
- check all those n nodes, if $P_n < P_g$: the node joins the group by taking the same polarization and P_g is increased by P_n. (preventing small groups to influence high rank members).

Features of society described by PROF-Sznajd model

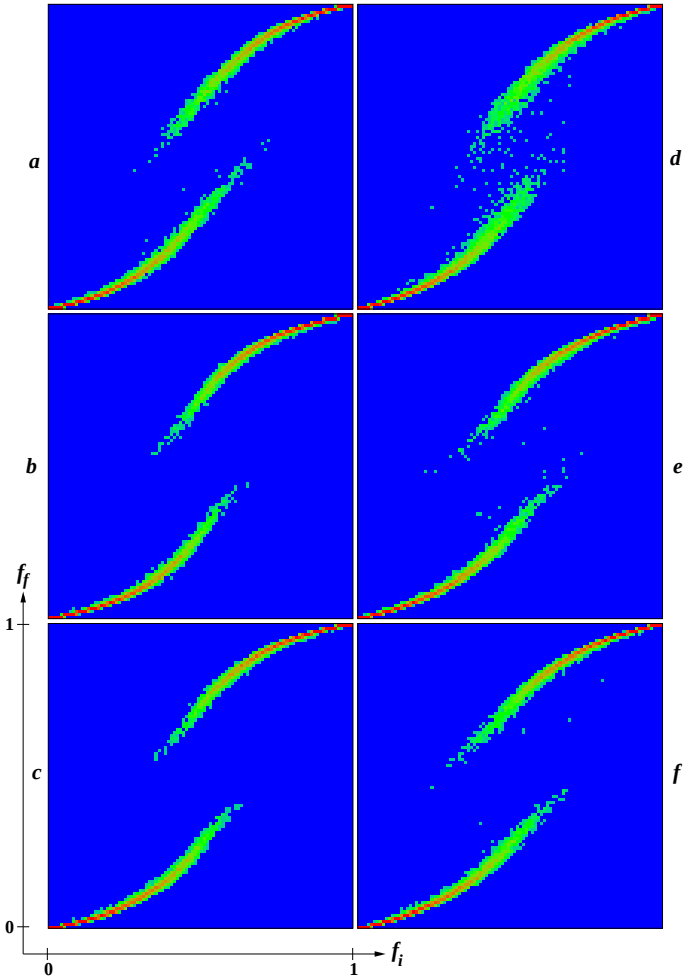


Figure: Density plot of probability constructed using $N_r = 10^4$ random realizations following the evolution up to the convergence time $\tau = 10^7$ iterations for Cambridge (left) and Oxford (right). Here from top to bottom $N_g = 3$, $N_g = 8$ and $N_g = 13$.

- bistability phase
- smaller fluctuations at larger N_g
- finite f_f at small $f_i \rightarrow$ resistance of small groups against totalitarian opinion

Introduction : from DNA sequence to network

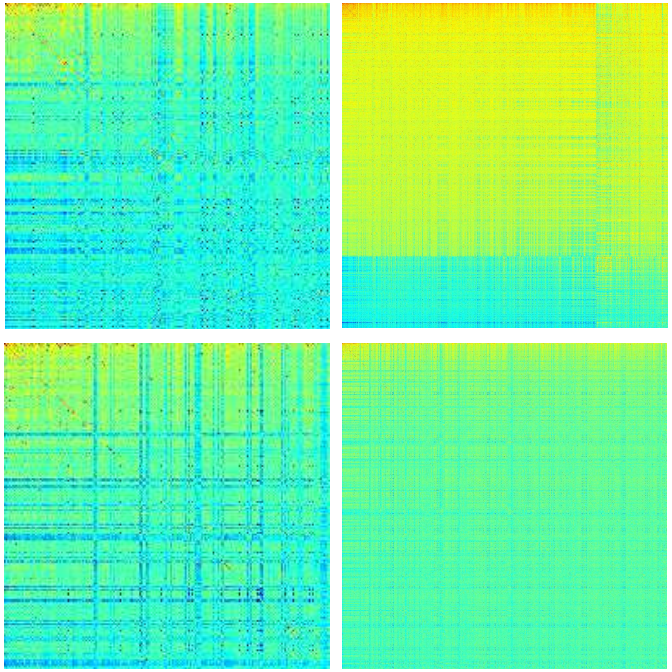
- Interest in detection of specific/rare patterns in a given sequence. New viewpoint of directed network.
- Single string of DNA sequences of length L base pairs. Bull, Dog, Elephant, Human and Zebrafish 10^9 - 10^{10} bp.
- Analysis are performed with $m = 5$, $m = 6$ and $m = 7$ letters words $\rightarrow$ size of the space of states (matrix size) are $N = 4^m = 1024$, $N = 4096$ and $N = 16384$ at $\alpha = 1$.

$$...TCG \underbrace{ATAT}_{W_{k-1}} \underbrace{CTGG}_{W_k} \underbrace{TAAC}_{W_{k+1}} CTA...$$

$$\rightarrow W_{k-1} \rightarrow W_k \rightarrow W_{k+1} \rightarrow$$

- $T_{ij} \rightarrow T_{ij} + 1$ whenever word j points to word i. At the end, all empty columns elements are replaced by $1/N$.

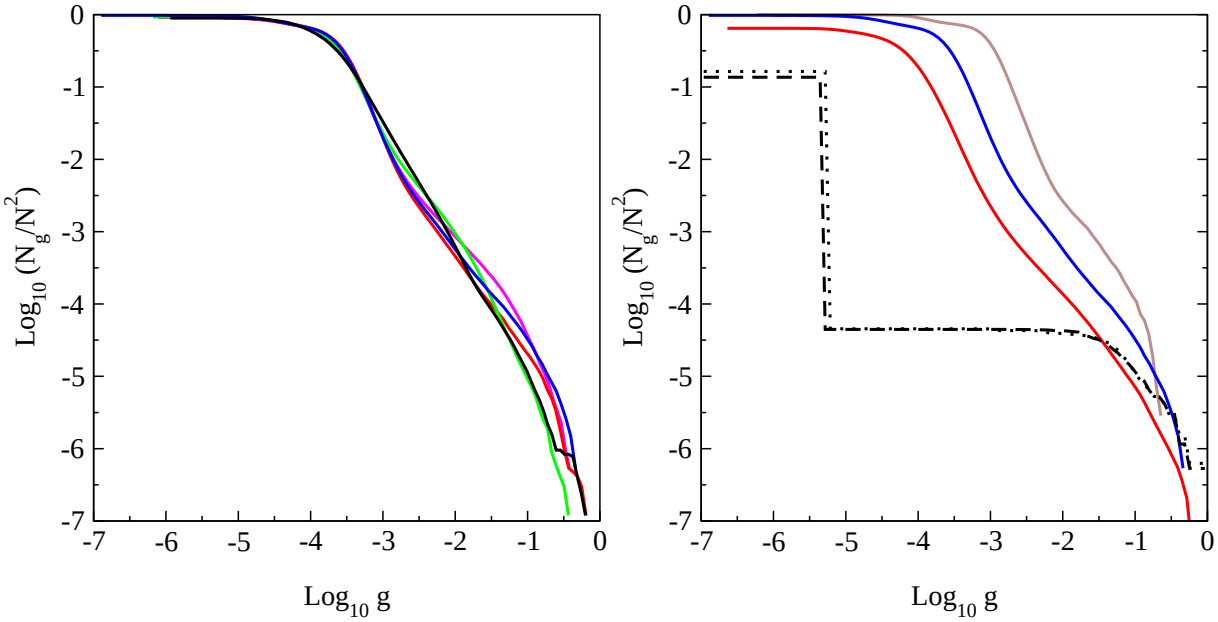
Statistics of Google matrix elements



DNA Google matrix of Homo sapiens (HS) constructed for words of 5-letters (top) and 6-letters (bottom) length. Matrix elements $G_{KK'}$ are shown in the basis of PageRank index K (and K'). Here, x and y axes show K and K' within the range $1 \leq K, K' \leq 200$ (left) and $1 \leq K, K' \leq 1000$ (right). The element G_{11} at $K = K' = 1$ is placed at top left corner. Color marks the amplitude of matrix elements changing from blue for minimum zero value to red at maximum value.

- Full matrix limit, $L/mN^2 \approx 10$ to 100 transitions per elements at $m = 6$.
- Webpages ≈ 10 links per node on average with $N \approx 2 \cdot 10^5$.

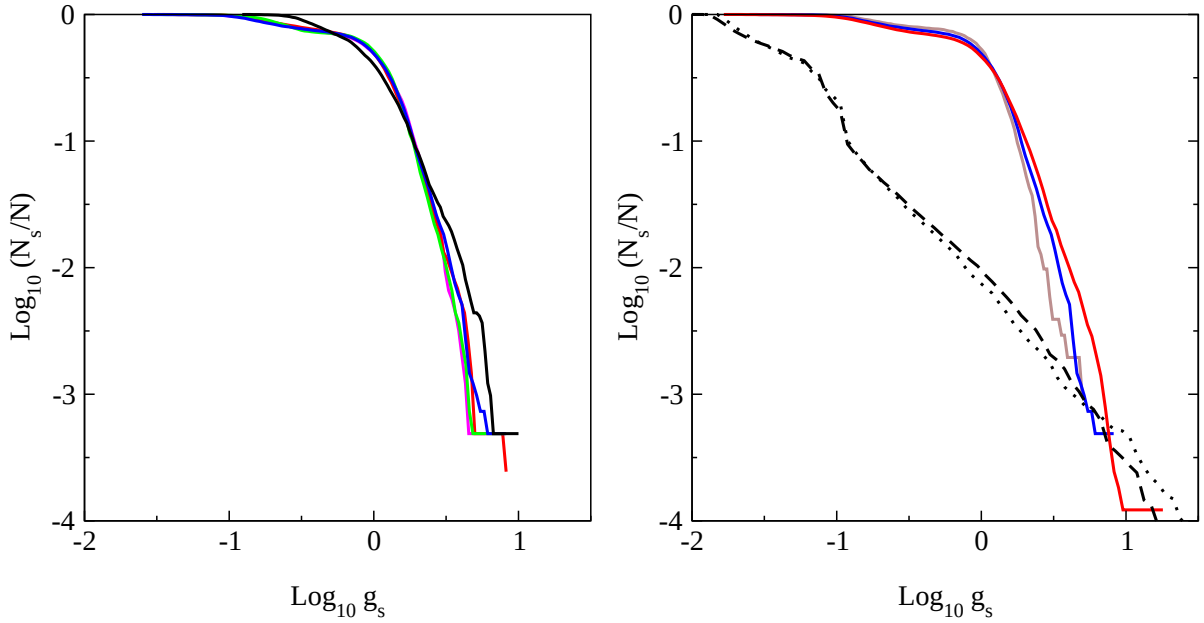
Statistics of Google matrix elements



Integrated fraction N_g/N^2 of Google matrix elements with $G_{ij} > g$ as a function of g. *Left panel :* Various species with 6-letters word length: bull BT (magenta), dog CF (red), elephant LA (green), Homo sapiens HS (blue) and zebrafish DR(black). *Right panel :* Data for HS sequence with words of length $m = 5$ (brown), 6 (blue), 7 (red). For comparison black dashed and dotted curves show the same distribution for the WWW networks of Universities of Cambridge and Oxford in 2006 respectively.

- Long range algebraic decay as $N_g \propto 1/g^{\nu-1}$. Fit in the range $-5.5 < \log_{10} g < -0.5$ gives : $\nu = 2.46 \pm 0.025$ (BT), 2.57 ± 0.025 (CF), 2.67 ± 0.022 (LA), 2.48 ± 0.024 (HS), 2.22 ± 0.04 (DR). For HS : $\nu = 2.68 \pm 0.038$ at $m = 5$ and $\nu = 2.43 \pm 0.02$ at $m = 7$.
- Oscillations but universal decay law with $\nu \approx 2.5$.
- Distribution of outgoing links in WWW networks decay with $\tilde{\nu} \approx 2.7$.

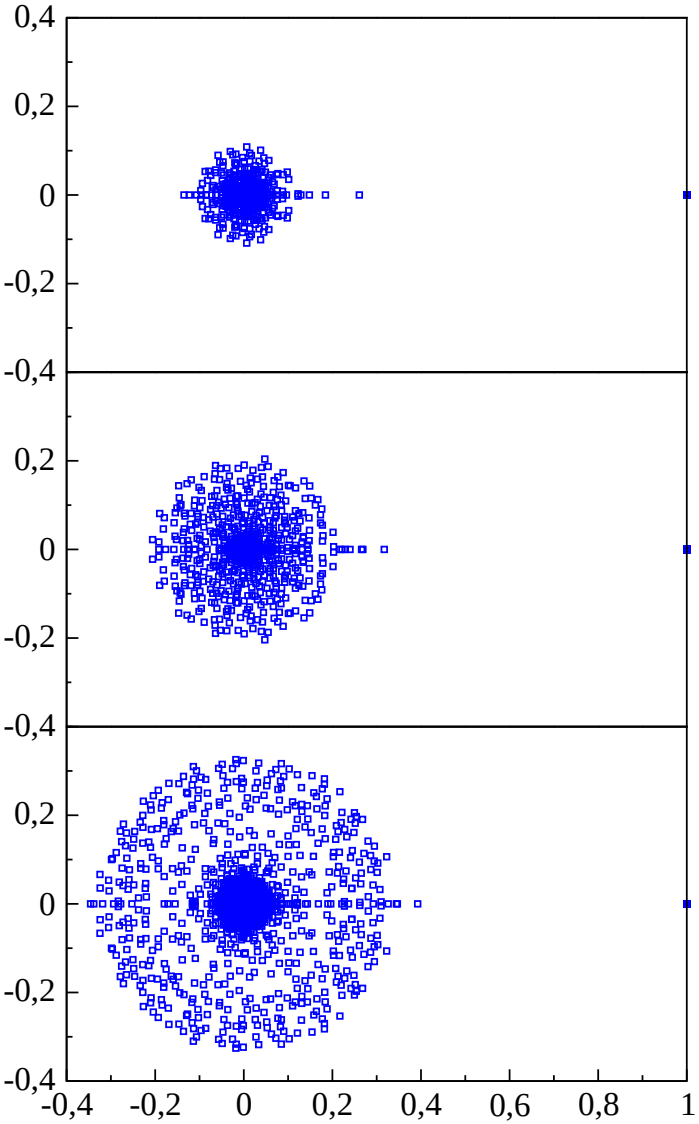
Statistics of Google matrix elements



Integrated fraction N_S/N of sum of ingoing matrix elements with $\sum_{j=1}^{N} G_{i,j} \geq g_S$. Left and right panels show the same cases as above in same colors. The dashed and dotted curves are shifted in x-axis by one unit left to fit the figure scale.

- Power law decay as $N_S \propto 1/g^{\mu-1}$. Fit gives $\mu = 5.59 \pm 0.15$ (BT), 4.90 ± 0.08 (CF), 5.37 ± 0.07 (LA), 5.11 ± 0.12 (HS), 4.04 ± 0.06 (DR). For HS at $m = 5, 7$ we have $\mu = 5.86 \pm 0.14$ and 4.48 ± 0.08.
- Distribution of ingoing links in WWW networks decay with $\tilde{\mu} \approx 2.1$.
- Visible differences between species but close to universal decay curve.

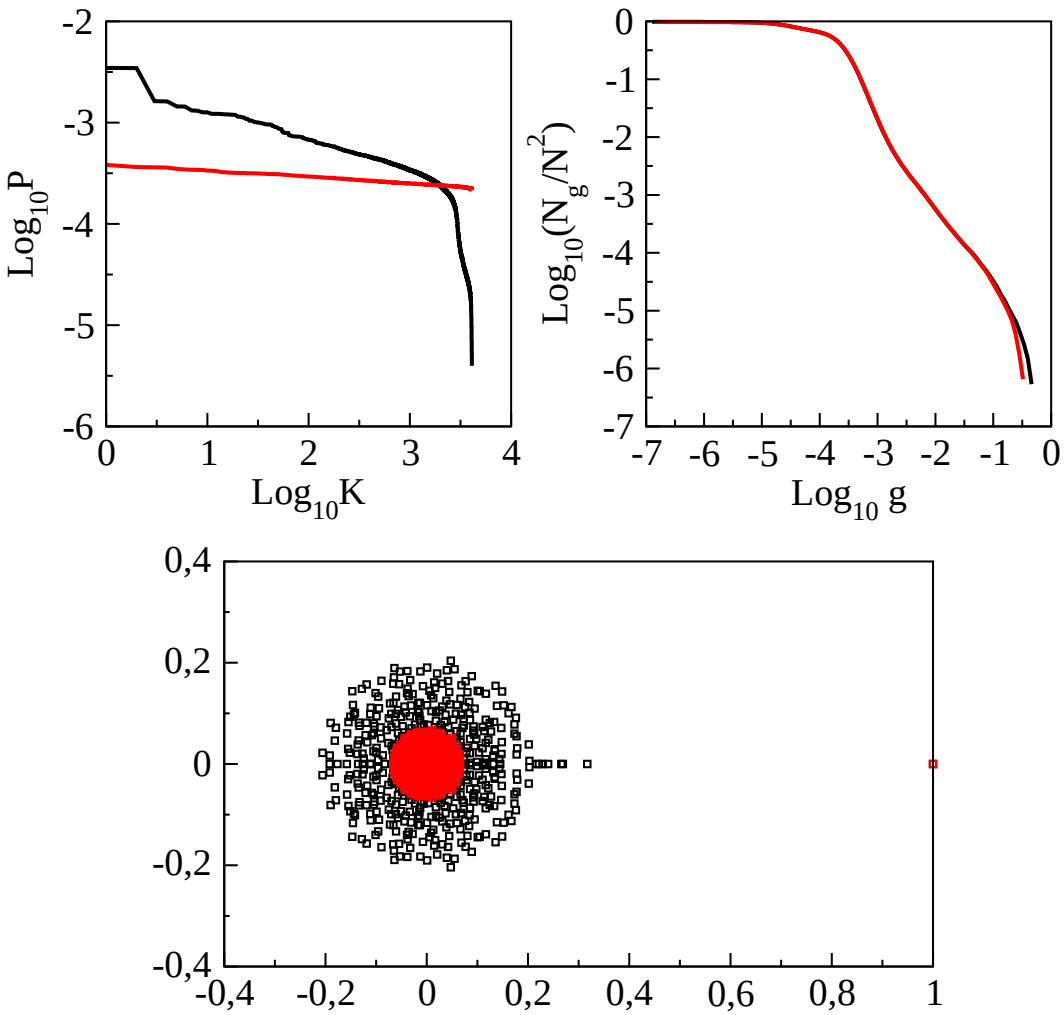
Spectrum and PageRank



Eigenvalue spectrum at $m = 5$, $m = 6$ and $m = 7$ of Homo Sapiens.

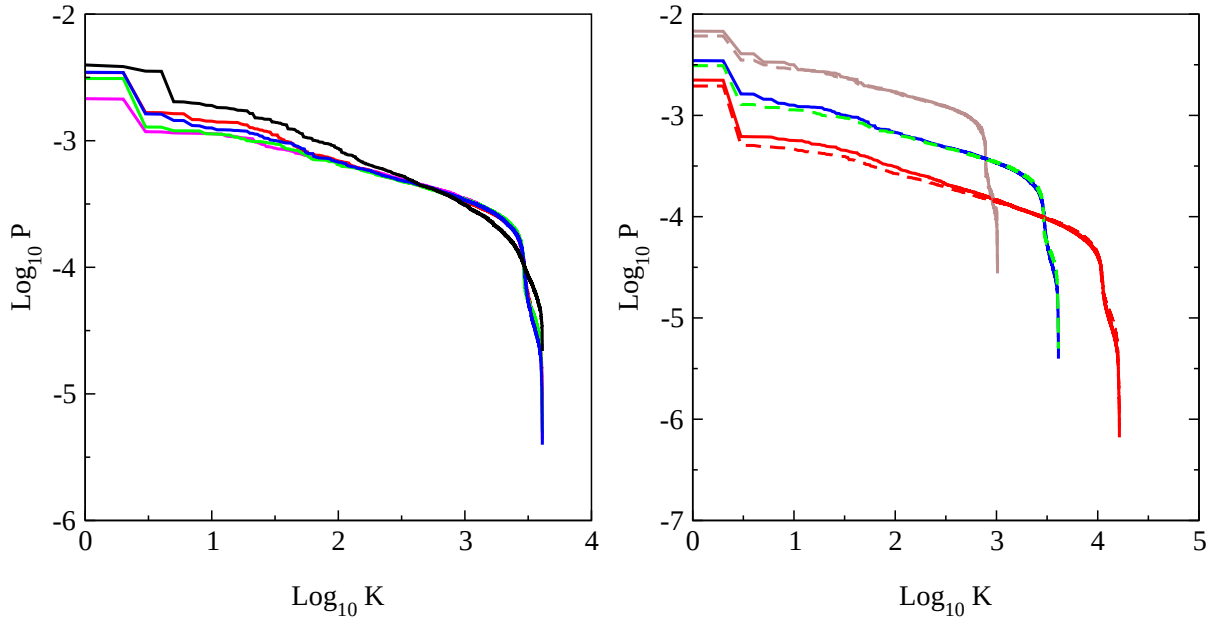
- Increase in word length leads to an increase of eigenvalue cloud radius, $\lambda_c \approx 0.1$, $\lambda_c \approx 0.2$ and $\lambda_c \approx 0.35$ for $m = 5$, $m = 6$ and $m = 7$.
- The spectrum is not reproducible with simple RMT model.

Spectrum and PageRank



Random matrix model with distribution of elements corresponding to HS at $m = 6$.

Spectrum and PageRank



PageRank probability decay of several species at $m = 6$ (left) and Homo Sapiens at $m = 5$, $m = 6$ and $m = 7$ (right).

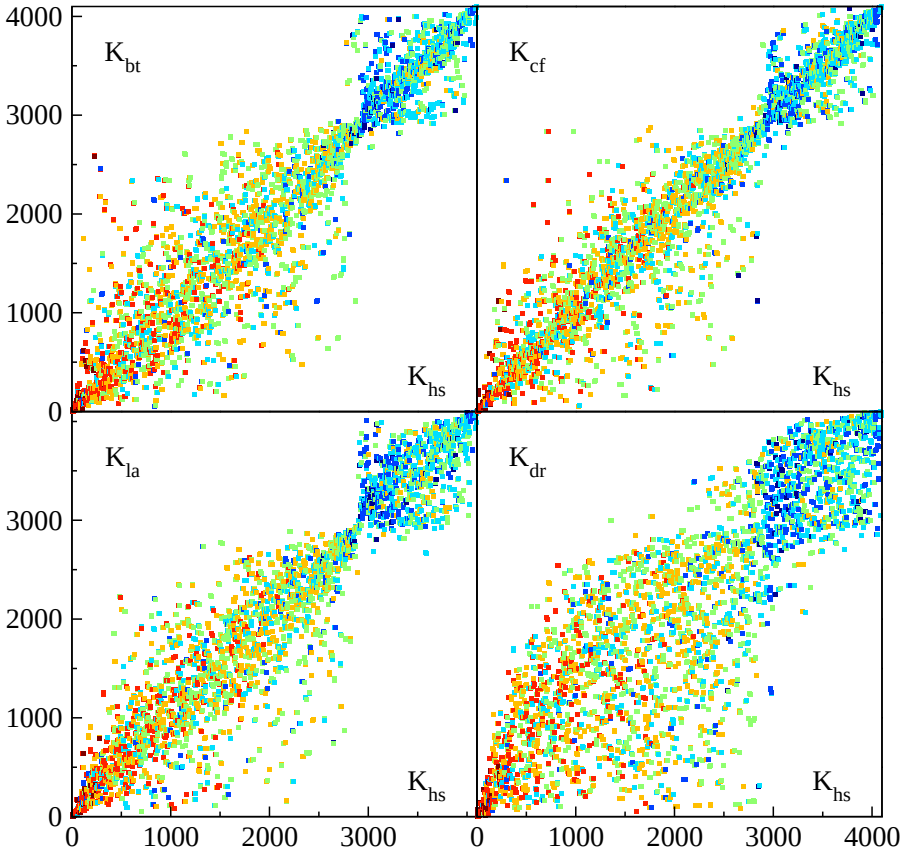
- PageRank $\sim$ frequency of words.
- $P(K) \sim 1/K^{\beta}$ with $\beta = 1/(\mu - 1)$.
- At $m = 6$: $\beta = 0.273 \pm 0.005$ (BT), 0.340 ± 0.005 (CF), 0.281 ± 0.005 (LA), 0.308 ± 0.005 (HS), 0.426 ± 0.008 (DR) in the range $1 \leq \log_{10} K \leq 3.3$. Small variation between mammalian species, stable with word length.

Top five (top) and last five (bottom) PageRank entries of DNA sequences.

BT	CF	LA	HS	DR
TTTTTT	TTTTTT	AAAAAA	TTTTTT	ATATAT
AAAAAA	AAAAAA	TTTTTT	AAAAAA	TATATA
ATTTTT	AATAAA	ATTTTT	ATTTTT	AAAAAA
AAAAAT	TTTATT	AAAAAT	AAAAAT	TTTTTT
TTCTTT	AAATAA	AGAAAA	TATTTT	AATAAA

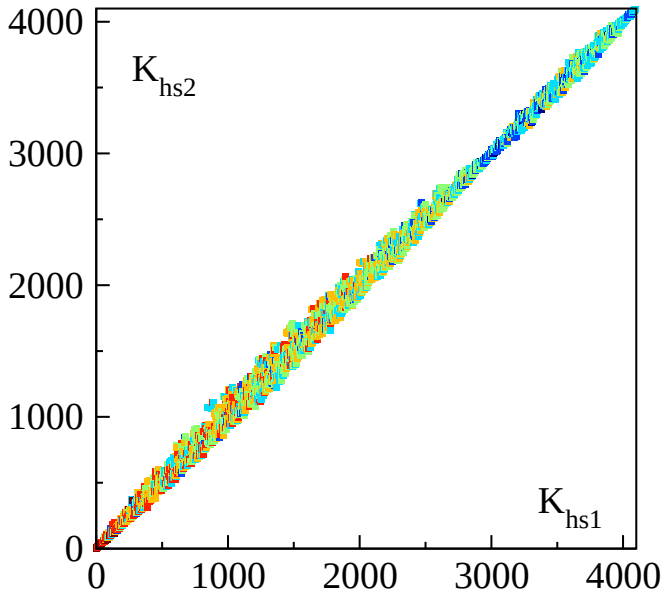
BT	CF	LA	HS	DR
CGCGTA	TACGCG	CGCGTA	TACGCG	CCGACG
TACGCG	CGCGTA	TACGCG	CGCGTA	CGTCGG
CGTACG	TCGCGA	ATCGCG	CGTACG	CGTCGA
CGATCG	CGTACG	TCGCGA	TCGACG	TCGACG
ATCGCG	CGATCG	CGCGAT	CGTCGA	TCGTCG

Statistical proximity



PageRank proximity K — K plane diagrams for different species in comparison with Homo Sapiens.

$$\zeta(s_1, s_2) = \frac{\sqrt{\sum_{i=1}^{N}(K_{s_1}(i) - K_{s_2}(i))^2)/N}}{\sigma_{rnd}}$$



PageRank proximity K — K plane diagrams between two Homo Sapiens individuals.

$$\begin{aligned}
\zeta(HS, CF) &= 0.206, \quad \zeta(HS, LA) = 0.238, \\
\zeta(HS, BT) &= 0.246, \quad \zeta(LA, CF) = 0.303, \\
\zeta(CF, BT) &= 0.308, \quad \zeta(LA, BT) = 0.324, \\
\zeta(DR, HS) &= 0.375, \quad \zeta(DR, CF) = 0.414, \\
\zeta(DR, LA) &= 0.422, \; \zeta(DR, BT) = 0.425
\end{aligned}$$

References

1. V.Kandiah and D.L.Shepelyansky *PageRank model of opinion formation on social networks*, Physica A v.391, p.5779-5793 (2012)
2. V.Kandiah and D.L.Shepelyansky *Google matrix analysis of DNA sequences*, PLOS One v.8(5), p. e61519 (2013)
3. L.Chakhmakhchyan and D.L.Shepelyansky *PageRank model of opinion formation on Ulam networks* Phys. Lett. A v.377, p.3119 (2013)