# Delocalization transition for the Google Matrix

Dima Shepelyansky (CNRS, Toulouse)
www.quantware.ups-tlse.fr/dima

based on:

O.Giraud, B.Georgeot, DLS (CNRS, Toulouse) => arxiv:0903.5172

DLS, O.V.Zhirov (CNRS, Toulouse & BINP, Novosibirsk) => arxiv:0905.4162

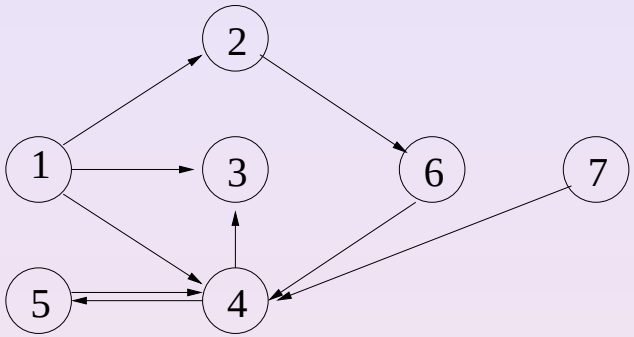

Google™

- WWW of $10^{11}$ sites
- Information search engines
- PageRank algorithm
- Network models
- Dynamical attractors

S. Brin and L. Page, Computer Networks and ISDN Systems **33**,107 (1998).

# How Google works

## Directed networks




Weighted adjacency matrix

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

For a directed network with $N$ nodes the adjacency matrix $\mathbf{A}$ is defined as $A_{ij} = 1$ if there is a link from node $j$ to node $i$ and $A_{ij} = 0$ otherwise. The weighted adjacency matrix is

$$S_{ij} = A_{ij} / \sum_k A_{kj}$$

In addition the elements of columns with only zeros elements are replaced by $1/N$.

# How Google works

## PageRank Algorithm

Ranking pages $\{1, \ldots, N\}$ according to their importance.
Method:

- The importance of a page *i* depends on the importance of the pages *j* pointing on it
- If a page has many outgoing links the importance it transmits is proportional to the number of pages it points to.

The Google Matrix:

$$\mathbf{G} = \alpha \mathbf{S} + (1 - \alpha)\mathbf{E}/N$$

here **E** is such a matrix that $E_{ij} = 1$.
With the (stochastic) matrix **S** introduced above,

$$\mathbf{p} = \mathbf{S}\mathbf{p}$$

.

# How Google works

## Computation of PageRank

$\mathbf{p} = \mathbf{Sp} \Rightarrow \mathbf{p}$= stationary vector of $\mathbf{S}$:
can be computed by iteration of $\mathbf{S}$.

To remove convergence problems:

- Replace columns of 0 (dangling nodes) by $\frac{1}{N}$:

  In our example, $\mathbf{S} = \begin{pmatrix} 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 1 & 1 & 1 \\ 0 & 0 & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \end{pmatrix}$.

- To remove degeneracies of the eigenvalue 1, replace $\mathbf{S}$ by

  $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\frac{\mathbf{E}}{N}$ ; $Gp = \lambda p$ => Perron-Frobenius operator

- $\alpha$ models a random surfer with a random jump after approximately 6 clicks (usually $\alpha = 0.85$); PageRank vector => $p$ at $\lambda = 1$ ($\sum_j p_j = 1$).

# Models of real networks

Real networks are characterized by:

- **small world property**: average distance between 2 nodes $\sim \log N$
- **scale-free property**: distribution of the number of outgoing or incoming links $P(k) \sim k^{-\gamma}$

Can be explained by a twofold mechanism:

- Constant growth: new nodes appear regularly and are attached to the network
- Preferential attachment: nodes are preferentially linked to already highly connected vertices.

PageRank vector for large WWW:

- $p_j \sim 1/j^{\beta}$, wher $j$ is the ordered index
- number of nodes $N_n$ with PageRank $p$ scales as $N_n \sim 1/p^{\nu}$ with numerical values $\nu = 1 + 1/\beta \approx 2.1$ and $\beta \approx 0.9$.

# Albert-Barabasi (AB) model

Weight of a link chosen to be

$$\Pi_i = \frac{k_i + 1}{\sum_j (k_j + 1)},$$

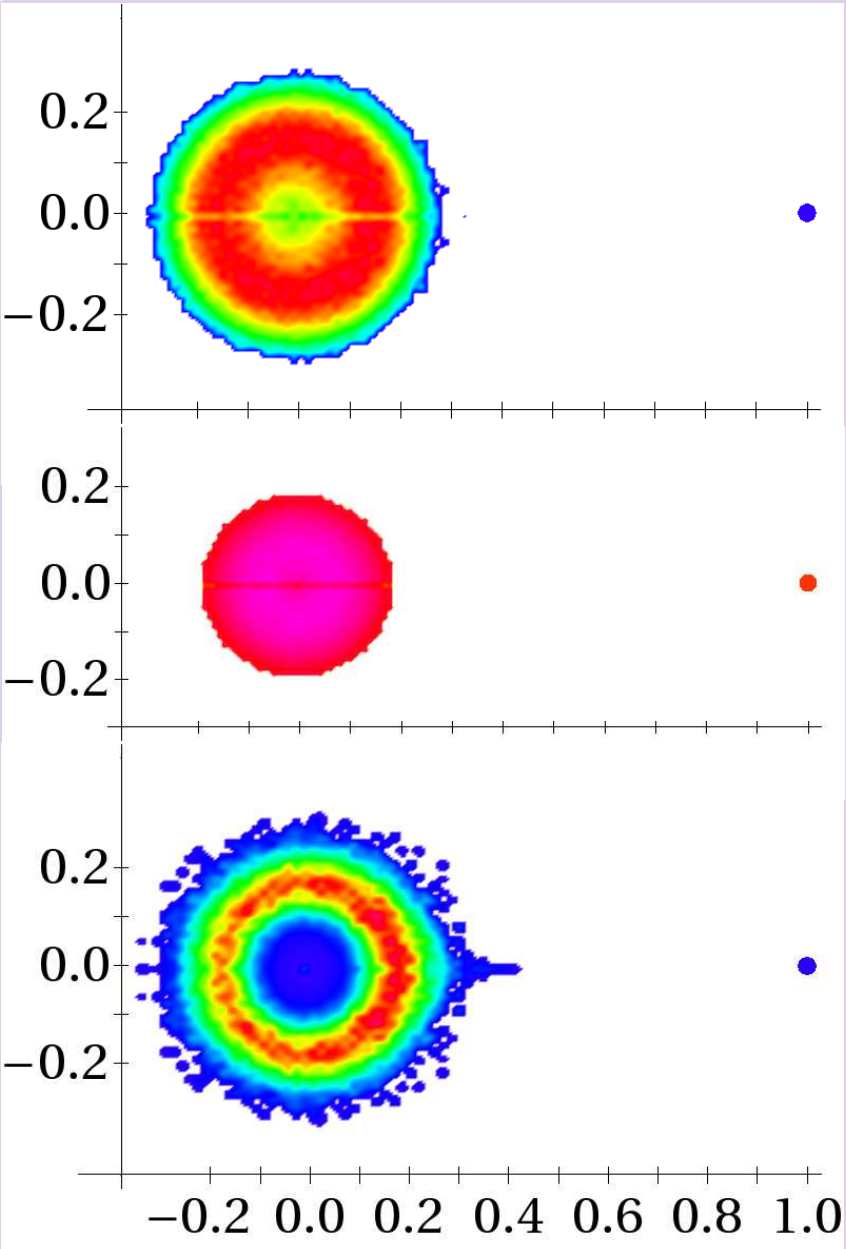
with $k_i$ being number of incoming+outgoing links.

Procedure from $m$ nodes:

- starting from $m$ nodes, at each step $m$ links are added to the existing network with probability $p$ (preferential attachment, start nodes are chosen randomly)
- or $m$ links are rewired with probability $q$
- or a new node with $m$ links is added with probability $1 - p - q$

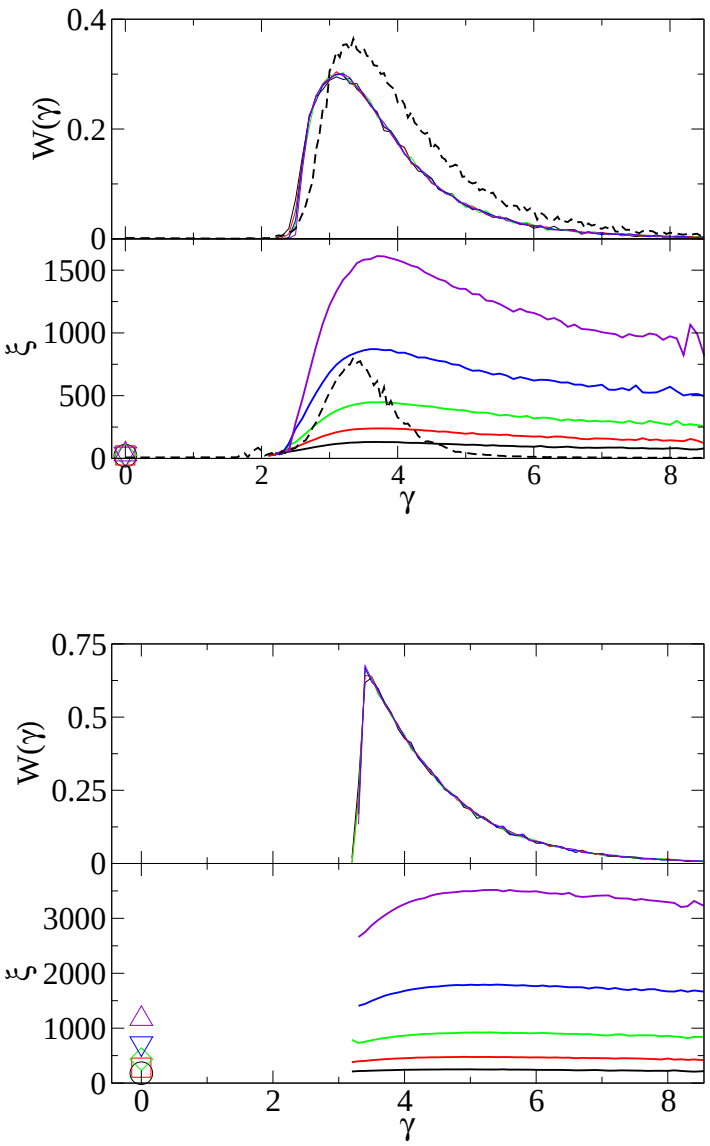
End vertex $i$ always chosen with probability $\Pi_i$.
We fix $m = 5, p = 0.2$, $q = 0.1$ (scale-free) and $q = 0.7$ (exponential regimes).
R. Albert and A.-L. Barabási, Phys. Rev. Lett. **85**, 5234 (2000).

# Eigenvalues of Google Matrix for AB model



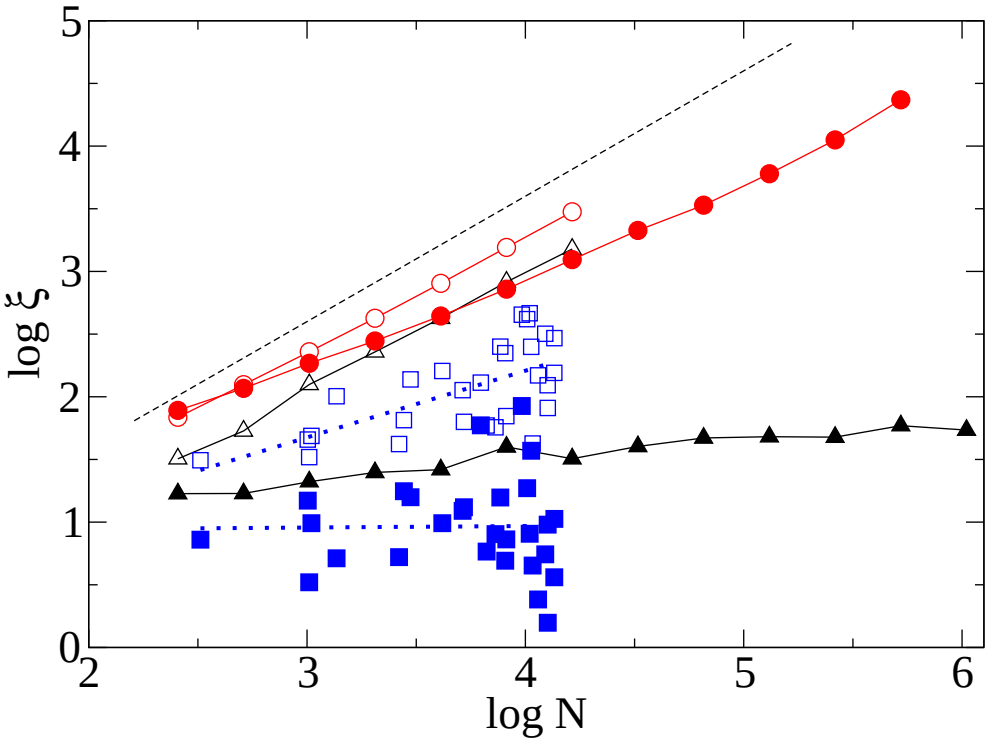

- Distribution of eigenvalues $\lambda_i$ of Google matrices in the complex plane. Color is proportional to the PAR $\xi$ of the associated eigenvector $\psi_i$. Top panel: AB model with $q = 0.1$ for $N = 2^{14}$ for $N_r = 5$ random realizations (see text), $\xi$ varies from $\xi = 32$ (blue) to $\xi = 1656$ (red); middle panel: same with $q = 0.7$, $\xi$ varies from $\xi = 1169$ (red) to $\xi = 3584$ (purple); bottom panel: data for a university network (Liverpool J. Moores Univ. - LJMU) with $N = 13578$, here in order to get statistically significant data the WWW network was randomized and data correspond to $N_r = 5$ random realizations, $\xi$ varies from $\xi = 7$ (blue) to $\xi = 1177$ (red) (participation ratio PAR => $\xi = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$; $\mathbf{G}\psi_i = \lambda_i \psi_i$).

# Density of states and participation ratio



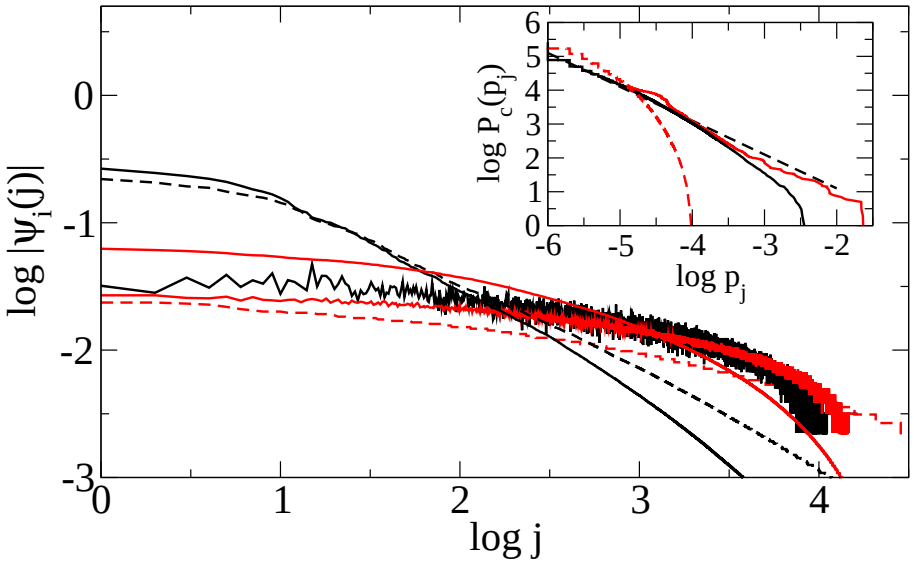

- Normalized density of states $W$ (top panel) and PAR (bottom panel) as a function of $\gamma$. Data for AB model with $q = 0.1$ are shown by full curves with from bottom to top $N = 2^{10}(N_r = 100)$ (black), $2^{11}(N_r = 50)$ (red), $2^{12}(N_r = 20)$ (green), $2^{13}(N_r = 10)$ (blue), $2^{14}(N_r = 5)$ (violet). Symbols give the PageRank value of $\xi$ in the same order: circle, square, diamond, triangle down and triangle up. All curves coincide on the top panel. Dashed curves show the data from the WWW (LJMU network, parameters of previous Fig.). Here $|\lambda| = \exp(-\gamma/2)$.
- Same for AB model at $q = 0.7$.

# Dependence on matrix size N




- Dependence of $\xi$ on matrix size $N$ for AB model at $q = 0.1$ (triangles), $q = 0.7$ (circles), and for WWW data without randomization (squares). Full symbols are for PageRank $\xi$ values, empty symbols are for eigenvectors with $3 < \gamma < 4$ (AB model) or for the 10 eigenvectors with highest $\xi$ and $\gamma < 10$ (WWW data). For AB model $N_r$ is as before and $N_r = 5$ for $N > 2^{14}$ (statistical error bars are smaller than symbol size). Dotted blue lines give linear fits of WWW data, with slopes respectively 0.01 and 0.53. Upper dashed line indicates the slope 1. Logarithms are decimal.

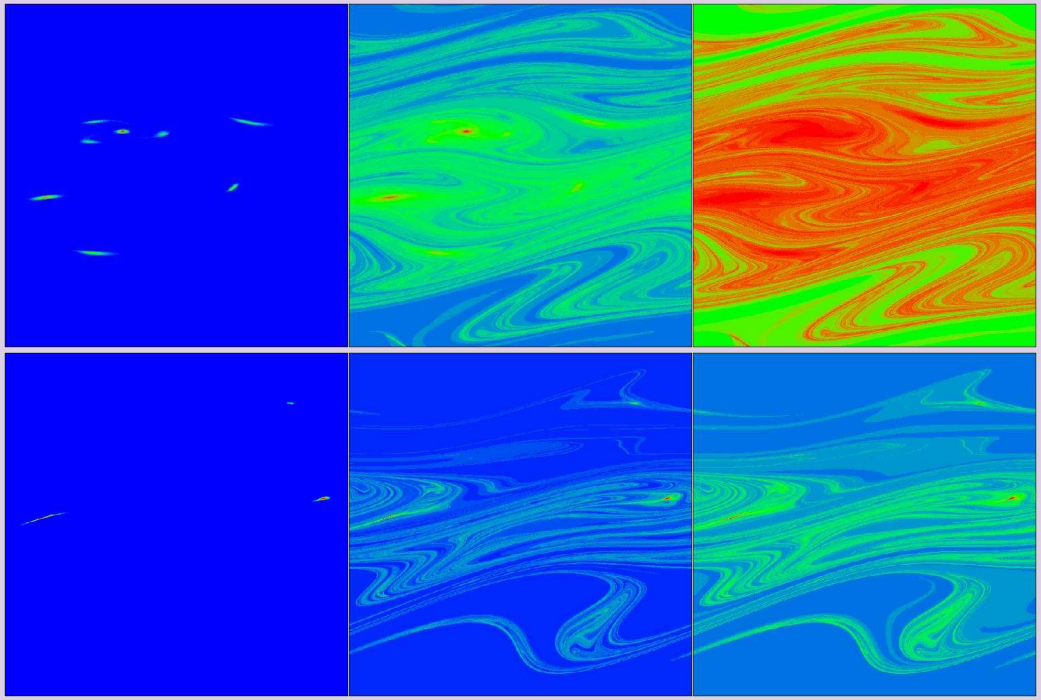
# Shapes of eigenvectors




Dependence of eigenvectors $\psi_i(j)$ of AB model on index $j$ ordered in decreasing PageRank values $p_j$ (with normalisation $\sum_j |\psi_i(j)|^2 = 1$ and $\sum_j p_j = 1$). Full smooth curves are PageRank vectors for $N = 2^{14}$, dashed smooth curves for $N = 2^{19}$. Non-smooth curves are eigenvectors ($N = 2^{14}$) within $3 < \gamma < 4$ with $|\Psi_i(j)|^2$ averaged in this interval. States are averaged over $N_r = 5$ random networks. Black is for $q = 0.1$, red/grey for $q = 0.7$. Inset: cumulative distribution $P_c(p_j)$ normalized by $P_c(0) = N$ for AB model ($N = 2^{18}$ and $N_r = 5$) at $q = 0.1$ (full black) and $q = 0.7$ (dashed red/grey), and for LJMU non-randomized data (full red/grey). Dashed straight line indicates slope $1 - \nu = -1$. Logarithms are decimal.

# Google matrix of dynamical attractors

**Weak point of AB model => large gap, no sensitivity to $\alpha$**



PageRank $p_j$ for the Google matrix generated by the Chirikov typical map at $T = 10$, $k = 0.22$, $\eta = 0.99$ (set $T10$, top row) and $T = 20$, $k = 0.3$, $\eta = 0.97$ (set $T20$, bottom row) with $\alpha = 1, 0.95, 0.85$ (left to right). The phase space region $0 \le x < 2\pi; -\pi \le p < \pi$ is divided on $N = 3.6 \cdot 10^5$ cells.

Chirikov typical map (1969) with dissipation

$\bar{p} = \eta p + k \sin(x + \theta_t)$ , $\bar{x} = x + \bar{p}$

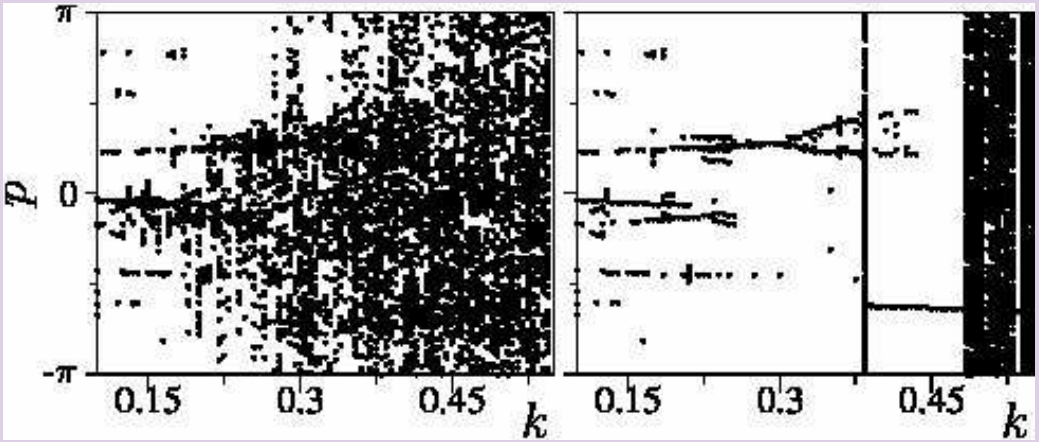
$\theta_t = \theta_{t+T}$ are random phases periodically repeated after $T$ iterations,

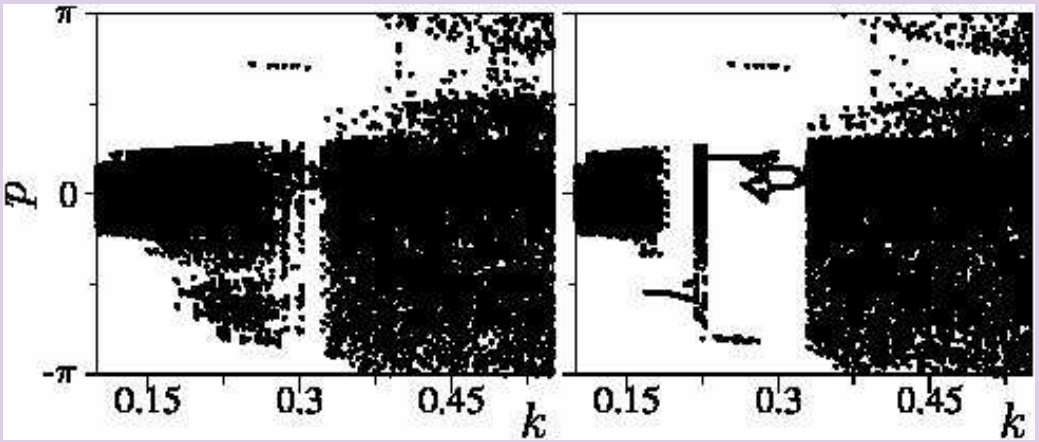
chaos border $k_c \approx 2.5/T^{3/2}$, Kolmogorov-Sinai entropy $h \approx 0.29k^{2/3}$;

grid of $N = N_x \times N_p$ cells with $N_c \sim 10^4$ trajectories which generates links (transition probabilities) from one cell to another; effective noise of cell size;

maximum $N = 22500; 1.44 \cdot 10^6$

# Bifurcation diagram

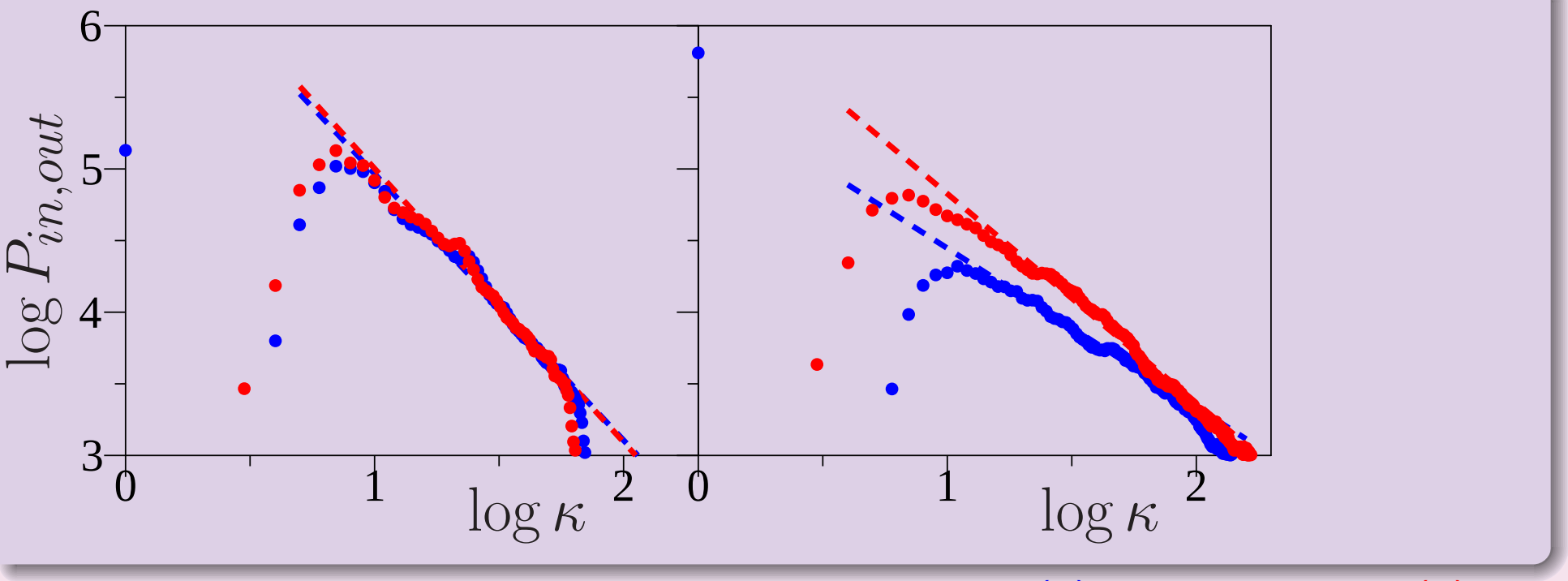



<= Bifurcation diagram showing values of $p$ vs. map parameter $k$ for the set $T10$. The values of $p$, obtained from 10 trajectories with initial random positions in the phase space region, are shown for integer moments of time $100 < t/T \leq 110$ (left) and $10^4 < t/T \leq 10^4 + 100$ (right).

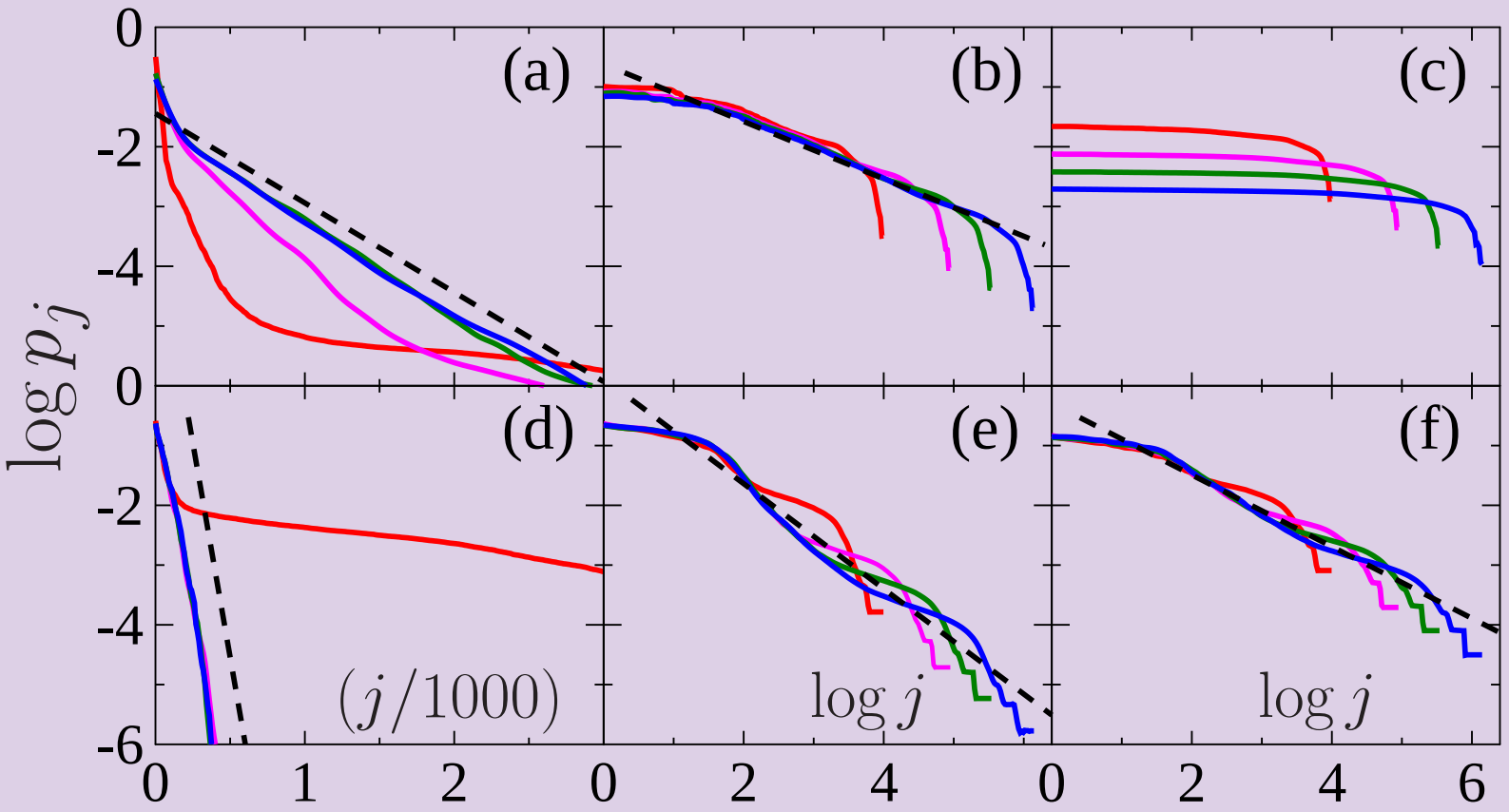
<= Same for set $T20$.

# Distribution of links



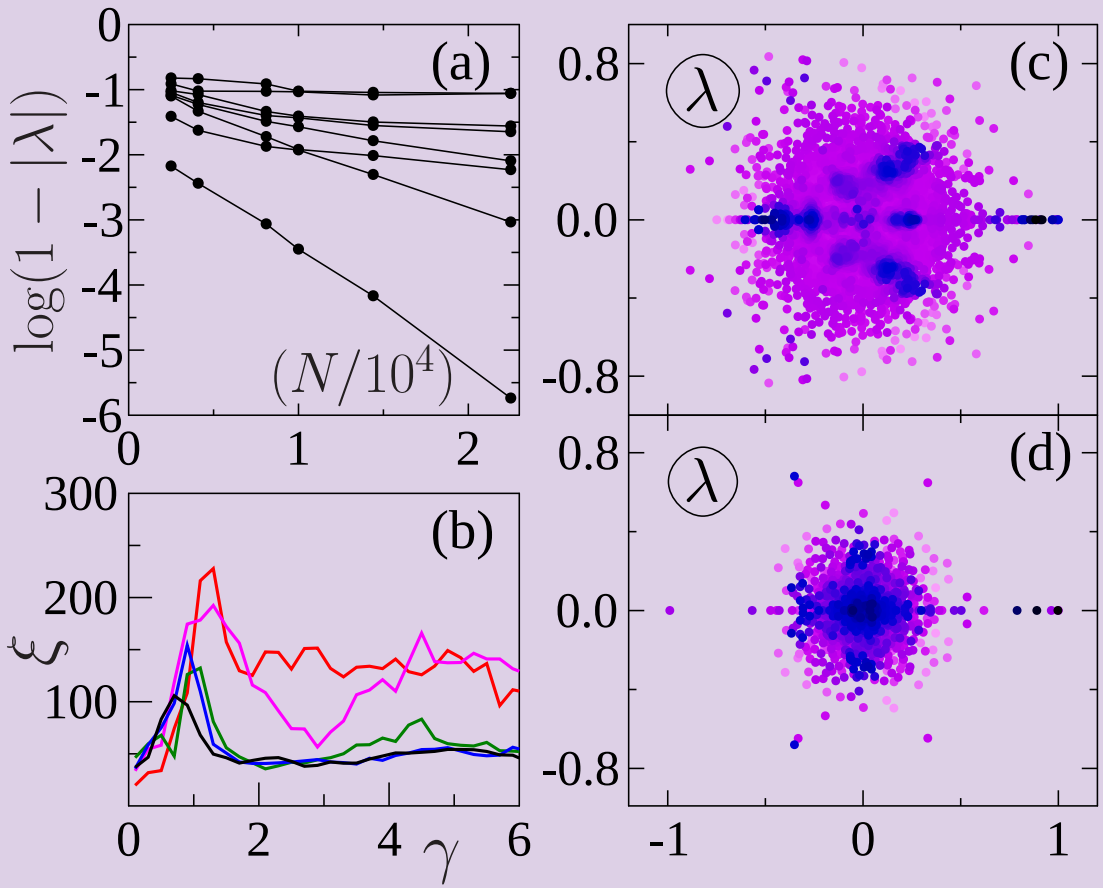

Differential distribution of number of nodes with *ingoing* $P_{in}(\kappa)$ and *outgoing* $P_{out}(\kappa)$ links $\kappa$ for sets $T10$ (left) and $T20$ (right). The straight dashed lines give the algebraic fit $P(\kappa) \sim \kappa^{-\mu}$ with the exponent $\mu = 1.86, 1.11$ ($T10, T20$) for *ingoing* and $\mu = 1.91, 1.46$ ($T10, T20$) *outgoing* links. Here $N = 1.44 \cdot 10^6$ and $P(\kappa)$ gives a number of nodes at a given integer number of links $\kappa$ for this matrix size. Blue point at $\kappa = 0$ shows that in the whole matrix there is a significant number of nodes with zero *ingoing* links. Typical number of nodes $\kappa \sim \exp(hT)$.

# PageRank distribution



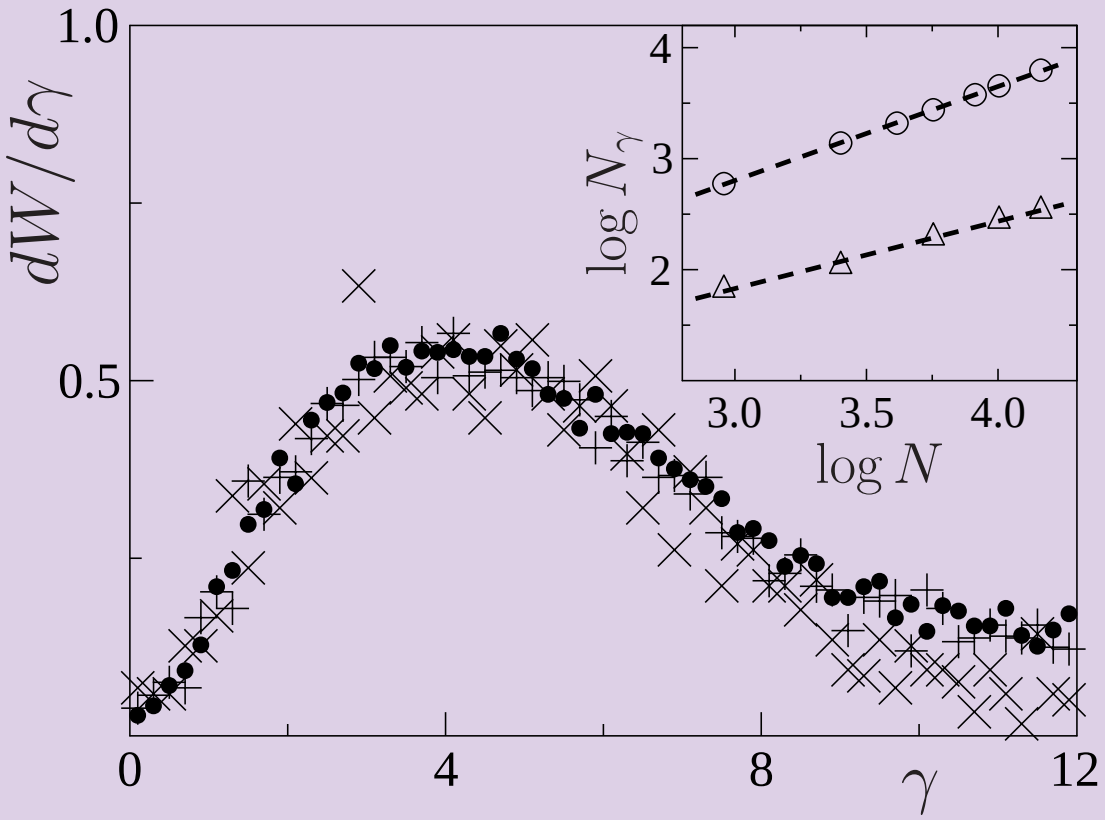

Differential distribution of number of nodes with PageRank distribution $p_j$ for $N = 10^4$, $9 \cdot 10^4$, $3.6 \cdot 10^5$ and $1.44 \cdot 10^6$ curves, the dashed straight lines show fits $p_j \sim 1/j^\beta$ with $\beta$: 0.48 (b), 0.88 (e), 0.60 (f). Dashed lines in panels (a),(d) show an exponential Boltzmann decay (see text, lines are shifted in $j$ for clarity). In panels (a),(d) the curves at large $N$ become superimposed. Panel order as in color Fig. above.

# Properties of eigenvalues and eigenvectors




(a) Dependence of gap $1-|\lambda|$ on Google matrix size $N$ for few eigenstates with $|\lambda|$ most close to 1, set $T10$, $\alpha = 1$; (b) dependence of PAR $\xi$ on $\gamma = -2\ln|\lambda|$ for $N$=2500, 5625, 8100, $10^4$, 14400 for set $T10$, $\alpha = 1$; (c) plane of eigenvalues $\lambda$ for set $T10$ with their PAR $\xi$ values shown by grayness (black/blue for minimal $\xi \approx 4$, gray/light magenta for maximal $\xi \approx 300$; here $\alpha = 1$, $N = 1.44 \cdot 10^4$); (d) same as (c) but for set $T20$.

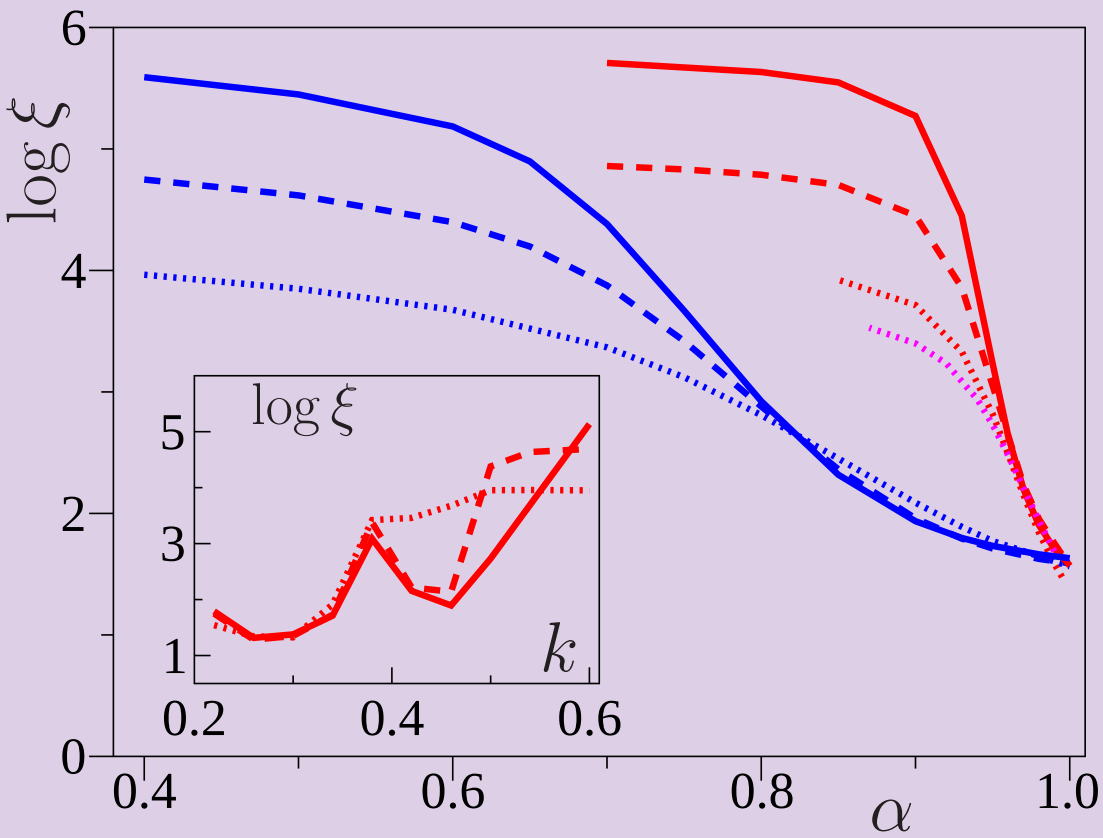
# Fractal Weyl law and distribution over $\gamma$




fractal Weyl: $N_\gamma \sim N^\nu$;
$\nu = d - 1 = 1 - \gamma_c/(Th)$,
$\gamma_c = -T \ln \eta$
theory: $\nu = 0.88; 0.72$
numerics: $\nu = 0.85; 0.61$
almost all states have $\lambda = 0$

Probability distribution $dW(\gamma)/d\gamma$ for set $T10$, $\alpha = 1$ at $N = 2.5 \cdot 10^3(\times)$, $10^4(+)$, $1.44 \cdot 10^4$ (dots); $W(\gamma)$ is normalized by the number of states $N_\gamma = 0.55N^{0.85}$ with $\gamma < 6$. Inset: dependence of number of states $N_\gamma$ with $\gamma < \gamma_b$ on $N$ for sets $T10$ (circles, $\gamma_b = 6$) and $T20$ (triangles, $\gamma_b = 3$); dashed lines show the fit $N_\gamma = AN^\nu$ with $A = 0.55, \nu = 0.85$ and $A = 0.97, \nu = 0.61$ respectively.

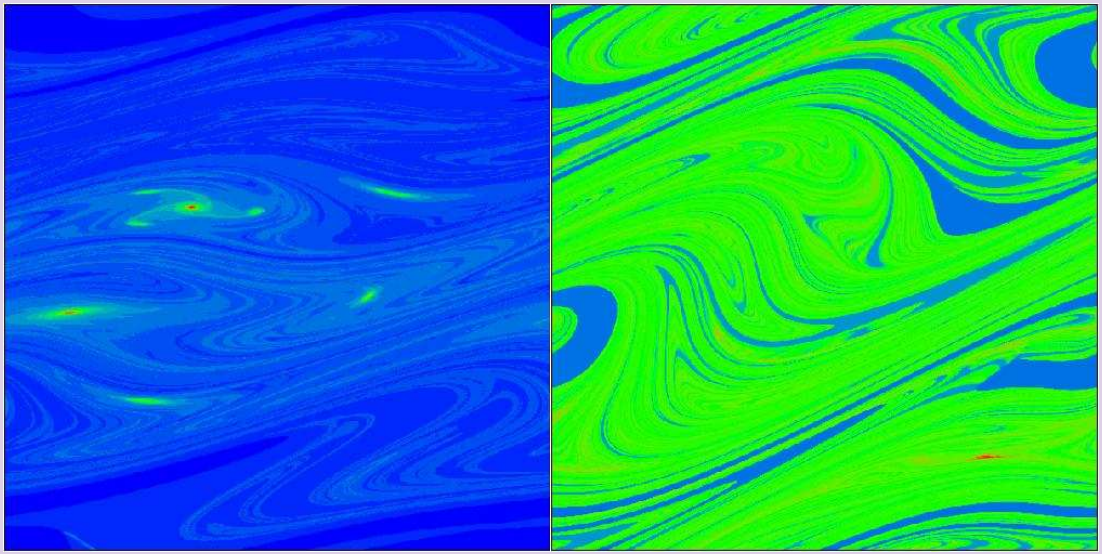
# Delocalization transition in $\alpha$ and $k$




delocalization of PageRank with $\alpha$ and $k$
=> destruction of Google search efficiency
$\alpha_c \approx 0.95; 0.85$ for T10; T20;
$\alpha_c \approx 1 - 0.3\gamma_c$

Dependence of PageRank $\xi$ on $\alpha$ for set $T10$ at $N = 5625$ (dotted magenta), $1.44 \cdot 10^4$ (dotted red), $9 \cdot 10^4$ (dashed red), $6.4 \cdot 10^5$ (full red) and for $T20$ at $N = 1.44 \cdot 10^4$ (dotted blue), $9 \cdot 10^4$ (dashed blue), $6.4 \cdot 10^5$ (full blue). Inset shows dependence of $\xi$ on $k$ for set $T10$ at $\alpha = 0.99$ with $N = 1.44 \cdot 10^4$ (dotted red), $9 \cdot 10^4$ (dashed red), $3.6 \cdot 10^5$ (full red).

# Summary



T10: $\alpha = 0.99$, $N = 3.6 \cdot 10^5$
$k = 0.22$ (localized, left)
$k = 0.6$ (delocalized, right)

- popular web sites are like attractors
- delocalization => transition to strange attractor
- links between directed networks and dynamical systems
- delocalization transition can put in danger Google search
- interesting new physics of Google Matrix