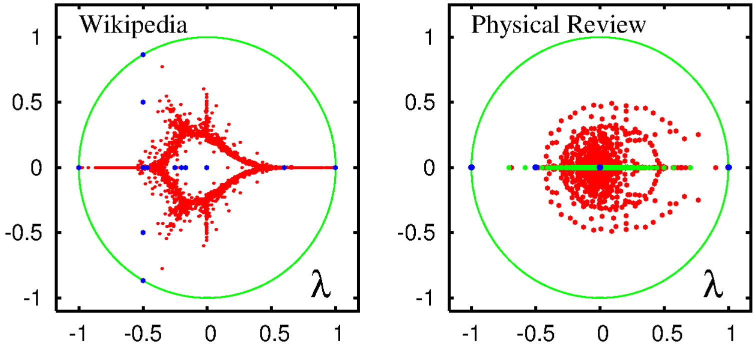


Introduction to Google matrix of directed networks

Klaus Frahm

Quantware MIPS Center Université Paul Sabatier
Laboratoire de Physique Théorique, UMR 5152, IRSAMC

A. D. Chepelianskii, Y. H. Eom, L. Ermann, B. Georgeot, D. Shepelyansky

Applications of Google matrix to directed networks and Big Data

Luchon, May 14 - 18, 2016

Contents

Perron-Frobenius operators

Consider a physical system with N states $i = 1, \ldots, N$ and probabilities $p_i(t) \geq 0$ evolving by a discrete ***Markov process***:

$$p_i(t+1) = \sum_j G_{ij}\, p_j(t) \quad \text{with} \quad \sum_i G_{ij} = 1 \quad , \quad G_{ij} \geq 0 \, .$$

The transition probabilities G_{ij} provide a ***Perron-Frobenius*** matrix. Conservation of probability: $\sum_i p_i(t+1) = \sum_i p_i(t) = 1$.

In general $G^T \neq G$ and eigenvalues λ may be complex and obey $|\lambda| \leq 1$. The vector $e^T = (1, \ldots, 1)$ is left eigenvector with $\lambda_1 = 1$ $\Rightarrow$ existence of (at least) one right eigenvector P for $\lambda_1 = 1$ also called ***PageRank*** in the context of Google matrices: $\boxed{G\, P = 1\, P}$

For non-degenerate λ_1 and finite gap $|\lambda_2| < 1$: $\boxed{\lim_{t\to\infty} p(t) = P}$

$\Rightarrow$ ***Power method*** to compute P with rate of convergence $\sim |\lambda_2|^t$.

PF Operators for directed networks

Consider a directed network with N nodes $1, \ldots, N$ and N_ℓ links.

Adjacency matrix:
$A_{jk} = 1$ if there is a link $k \to j$ and $A_{jk} = 0$ otherwise.

Sum-normalization of each non-zero column of $A \quad \Rightarrow \quad S_0$.

Replacing each zero column (***dangling nodes***) with $e/N \quad \Rightarrow \quad S$.

Eventually apply the ***damping factor*** $\alpha < 1$ (typically $\alpha = 0.85$):

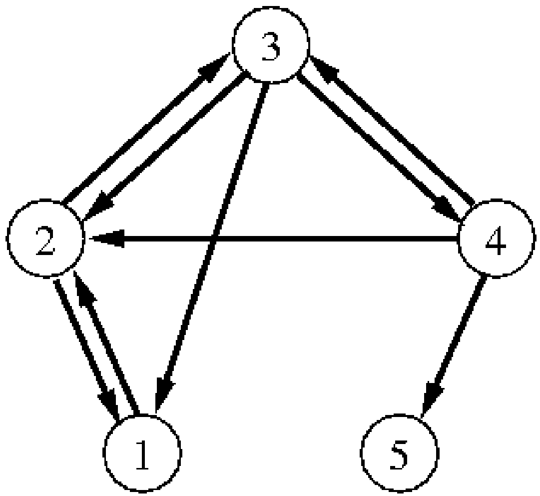
Google matrix:

$$G(\alpha) = \alpha S + (1-\alpha)\,\frac{1}{N}\,ee^T \quad .$$

$\Rightarrow \quad \lambda_1$ is non-degenerate and $|\lambda_2| \le \alpha$.

Same procedure for inverted network: $A^* \equiv A^T$ where S^* and G^* are obtained in the same way from A^*. Note: in general: $S^* \neq S^T$. Leading (right) eigenvector of S^* or G^* is called ***CheiRank***.

Example:

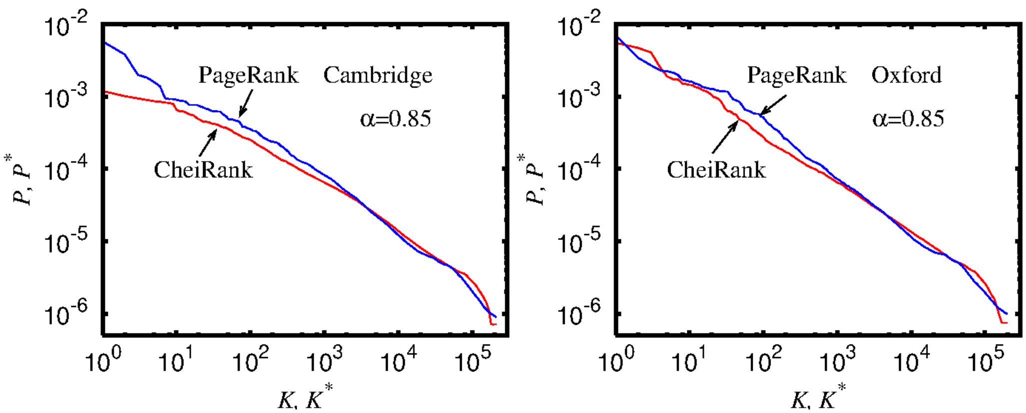


$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$S_0 = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 \\ 1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix} \quad , \quad S = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 0 & \frac{1}{5} \\ 1 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{5} \\ 0 & \frac{1}{2} & 0 & \frac{1}{3} & \frac{1}{5} \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{5} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{5} \end{pmatrix}$$

PageRank

Example for university networks of Cambridge 2006 and Oxford 2006 ($N \approx 2 \times 10^5$ and $N_\ell \approx 2 \times 10^6$).



$$P(i) = \sum_j G_{ij}\, P(j)$$

$P(i)$ represents the “importance” of “node/page i” obtained as sum of all other pages j pointing to i with weight $P(j)$. Sorting of $P(i)$ $\Rightarrow$ index $K(i)$ for order of appearance of search results in search engines such as Google.

Numerical diagonalization

- ***Power method*** to obtain P: rate of convergence for $G(\alpha) \sim \alpha^t$.
- Full "exact" diagonalization ($N \lesssim 10^4$).
- ***Arnoldi method*** to determine largest $n_A \sim 10^2 - 10^4$ eigenvalues. Idea: write

$$G\,\xi_k = \sum_{j=0}^{k+1} H_{jk}\,\xi_j \quad \text{for} \quad k = 0,\,\ldots\,,\, n_A - 1$$

where ξ_{k+1} is obtained from ***Gram-Schmidt*** orthogonalization of $G\xi_k$ to $\xi_0,\,\ldots\,,\,\xi_k$ with ξ_0 being some suitable normalized initial vector. $\xi_0,\,\ldots\,,\,\xi_{n_A-1}$ span a ***Krylov space*** of dimension n_A and the eigenvalues of the "small" representation matrix H_{jk} are (very) good approximations to the largest eigenvalues of G.

Example for Twitter network of 2009: $N \approx 4 \times 10^7$ and $N_\ell \approx 1.5 \times 10^9$ with $n_A = 640$ (lower N in other examples allows for higher n_A).

- Practical problems due to ***invariant subspaces*** of nodes in realistic WWW networks creating large degeneracies of λ_1 (or λ_2 if $\alpha < 1$). Decomposition in subspaces and a core space

$$\Rightarrow \quad S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix}$$

  where S_{ss} is block diagonal according to the subspaces. The subspace blocks of S_{ss} are all matrices of PF type with at least one eigenvalue $\lambda_1 = 1$ explaining the high degeneracies.

  To determine the spectrum of S apply exact (or Arnoldi) diagonalization on each subspace and the Arnoldi method to S_{cc} to determine the largest core space eigenvalues λ_j (note: $|\lambda_j| < 1$).

- Strange numerical problems to determine accurately "small" eigenvalues, in particular for (nearly) ***triangular network structure*** due to large Jordan-blocks (e.g. citation network of Physical Review).

Reduced Google matrix

Consider a sub-network with $N_r \ll N$ nodes providing a decomposition in ***reduced*** and ***scattering*** nodes:

$$G = \begin{pmatrix} G_{rr} & G_{rs} \\ G_{sr} & G_{ss} \end{pmatrix} \quad , \quad P = \begin{pmatrix} P_r \\ P_s \end{pmatrix}$$

$$G\,P = P \quad \Rightarrow \quad G_{\rm R} P_r = P_r$$

with the ***effective reduced Google matrix***:

$$\boxed{G_{\rm R} = G_{rr} + G_{rs}(\mathbf{1} - G_{ss})^{-1} G_{sr}}$$

containing ***direct link contributions*** from G_{rr} and

scattering contributions from $G_{rs}(\mathbf{1} - G_{ss})^{-1} G_{sr}$.

Problem: pratical evaluation of $(\mathbf{1}-G_{ss})^{-1}$ is very difficult for large network sizes and the expansion

$$(\mathbf{1}-G_{ss})^{-1} = \sum_{l=0}^{\infty} G_{ss}^{\,l}$$

typically converges very slowly since the leading eigenvalue λ_c of G_{ss} is very close to unity: $1-\lambda_c \ll 1$.

Proposal of numerical algorithm:

$$\boxed{(\mathbf{1}-G_{ss})^{-1} = \mathcal{P}_c \frac{1}{1-\lambda_c} + \mathcal{Q}_c \sum_{l=0}^{\infty} \bar{G}_{ss}^{\,l}}$$

with $\bar{G}_{ss} = \mathcal{Q}_c G_{ss} \mathcal{Q}_c$, the projectors $\mathcal{P}_c = \psi_R \psi_L^T$, $\mathcal{Q}_c = \mathbf{1} - \mathcal{P}_c$ and $\psi_{R,L}$ are right/left eigenvectors of G_{ss} for λ_c such that $\psi_L^T \psi_R = 1$.

The leading eigenvalue of $\bar{G}_{ss}$ is close to $\alpha = 0.85$

$\Rightarrow$ rapid convergence of the matrix series.

Additional damping factor:

$$G_{\rm mod} = \begin{pmatrix} \mathbf{1} & (1-\eta)U_{rs} \\ 0 & \eta\mathbf{1} \end{pmatrix} \times \begin{pmatrix} G_{rr} & G_{rs} \\ G_{sr} & G_{ss} \end{pmatrix}$$

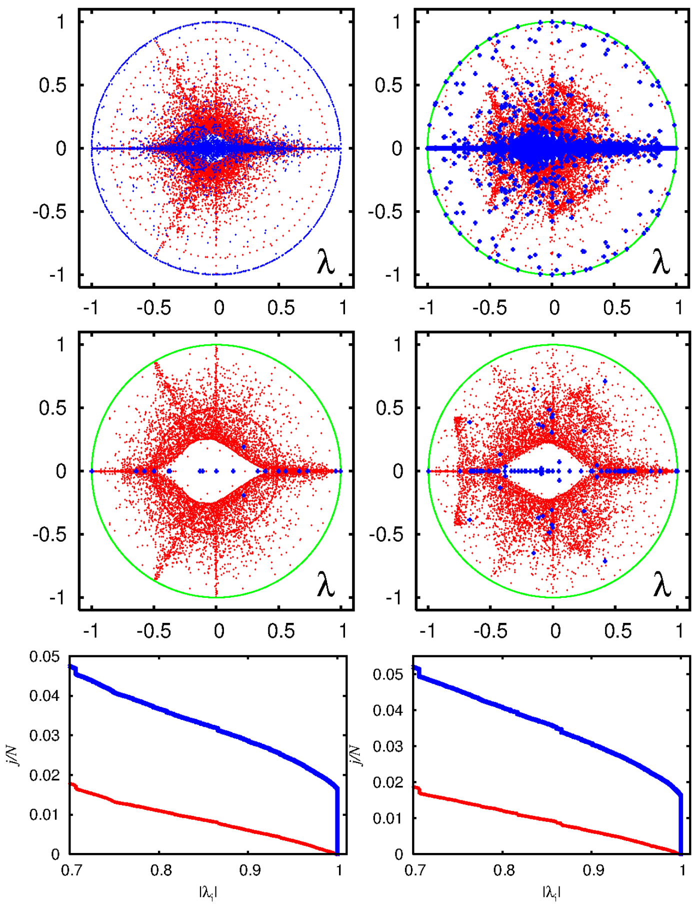
with $0.5 \leq \eta < 1$ and $U_{rs} = (1/N_r)e_r e_s^T$.

$$\Rightarrow \qquad \boxed{(G_{\rm mod})_{ss} = \eta G_{ss}}$$

$\Rightarrow$ no convergence problem for

$$(\mathbf{1} - \eta G_{ss})^{-1} = \sum_{l=0}^{\infty} \eta^l \, G_{ss}^l \quad \text{if} \quad \eta < 1 \; .$$

University Networks



Cambridge 2006 (left),
$N = 212710$, $N_s = 48239$

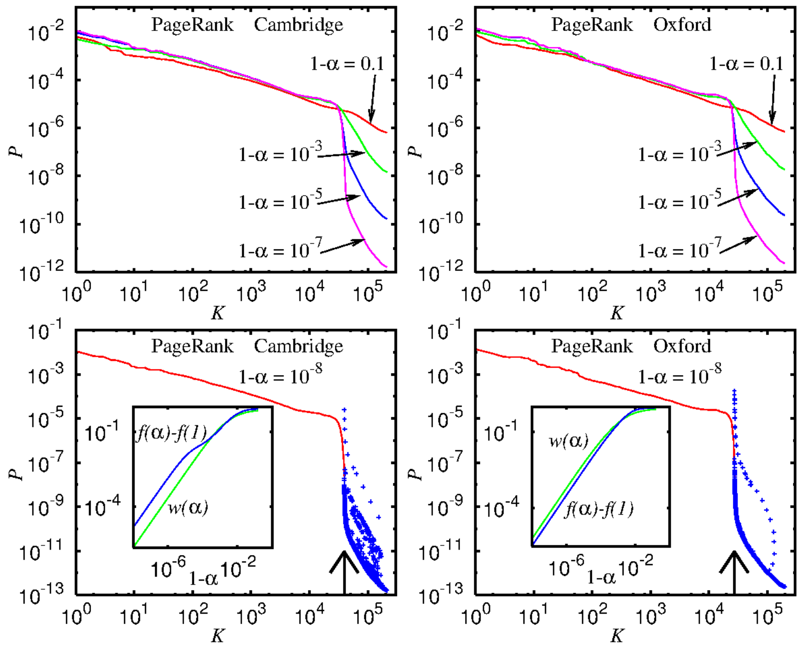
Oxford 2006 (right),
$N = 200823$, $N_s = 30579$

Spectrum of S (upper panels), S^* (middle panels) and dependence of rescaled level number on $|\lambda_j|$ (lower panels).

Blue: subspace eigenvalues

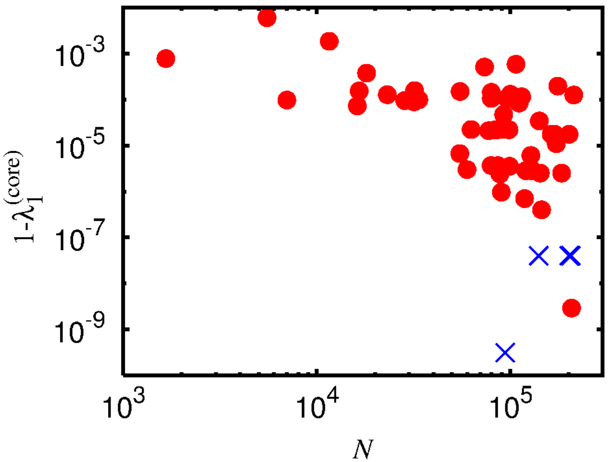
Red: core space eigenvalues (with Arnoldi dimension $n_A = 20000$)

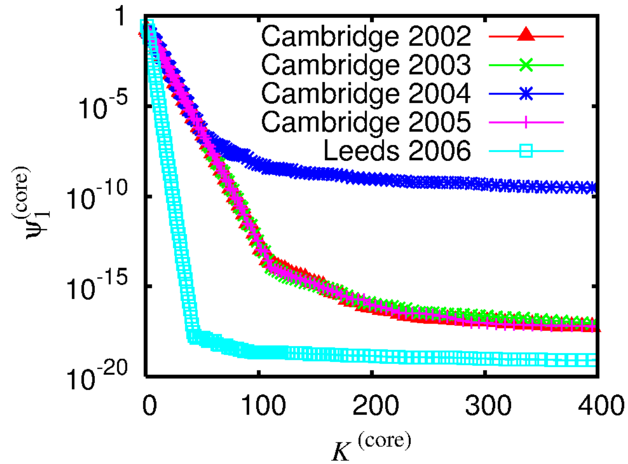
PageRank for $\alpha \to 1$:



$$P = \underbrace{\sum_{\lambda_j=1} c_j\,\psi_j}_{\text{subspace contributions}} + \sum_{\lambda_j \neq 1} \frac{1-\alpha}{(1-\alpha)+\alpha(1-\lambda_j)}\, c_j\,\psi_j \ .$$

Core space gap and quasi-subspaces





Left: Core space gap $1 - \lambda_1^{(\mathrm{core})}$ vs N for certain british universities.

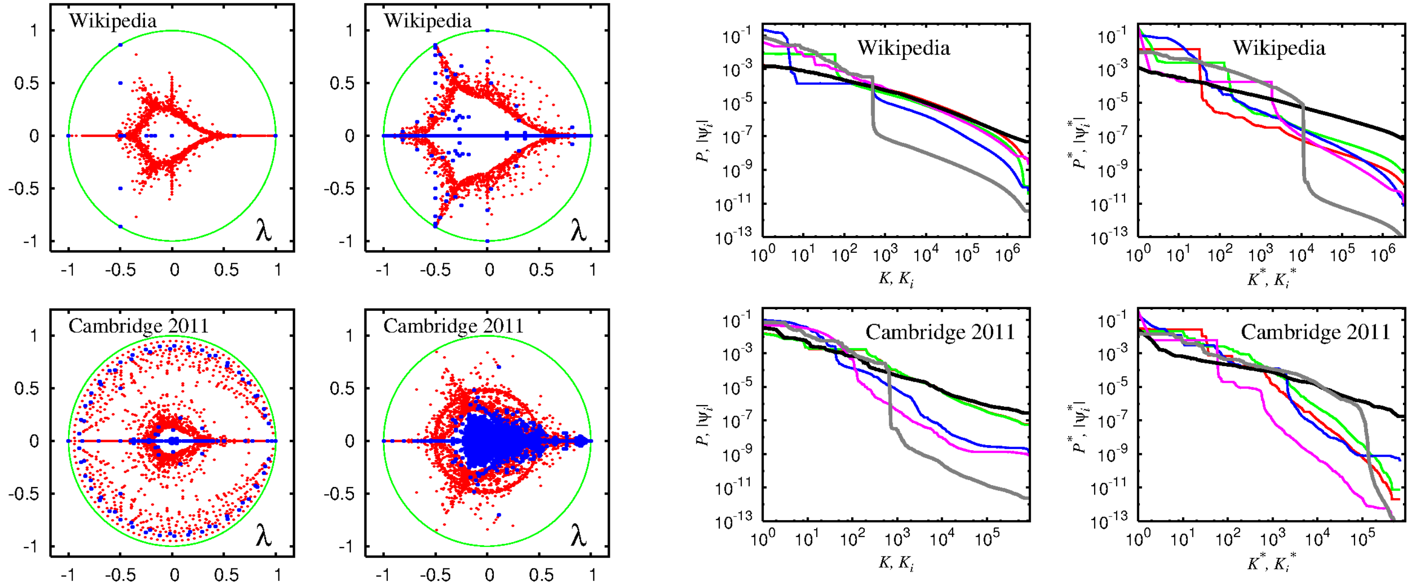
Red dots for gap $> 10^{-9}$; blue crosses (moved up by 10^9) for gap $< 10^{-16}$.

Right: first core space eigenvecteur for universities with gap $< 10^{-16}$ or gap $= 2.91 \times 10^{-9}$ for Cambridge 2004.

Core space gaps $< 10^{-16}$ correspond to ***quasi-subspaces*** where it takes quite many “iterations” to reach a dangling node.

Wikipedia

Wikipedia 2009 : $N = 3282257$ nodes, $N_\ell = 71012307$ network links.

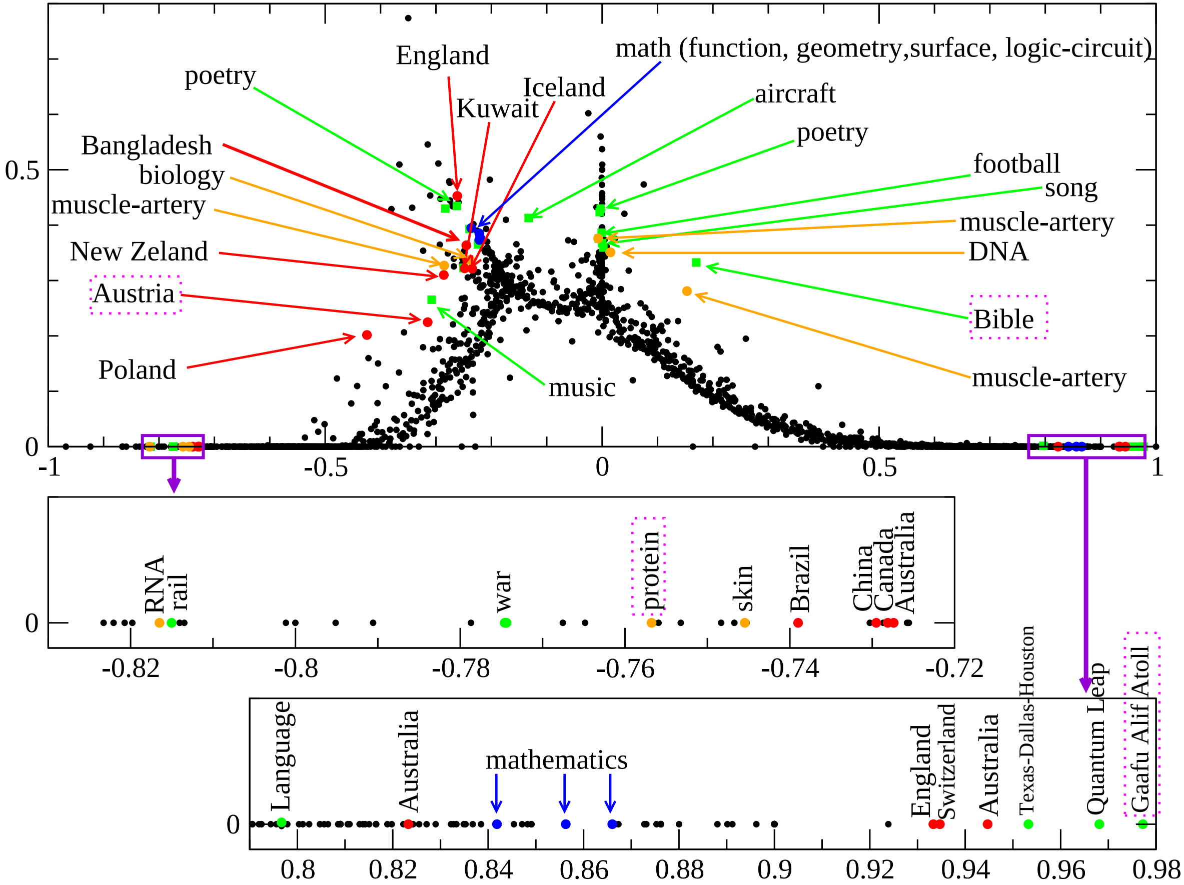


left (right): PageRank (CheiRank)

black: PageRank (CheiRank) at $\alpha = 0.85$
grey: PageRank (CheiRank) at $\alpha = 1 - 10^{-8}$
red and green: first two core space eigenvectors

blue and pink: two eigenvectors with large imaginary part in the eigenvalue

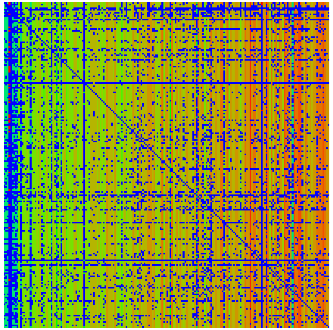
“Themes” of certain Wikipedia eigenvectors:

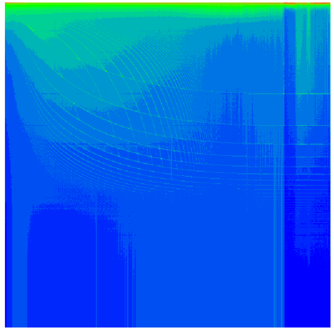


Twitter network

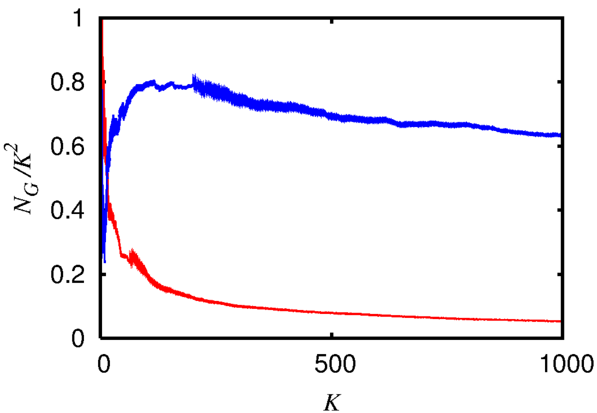
Twitter 2009 : $N = 41652230$ nodes, $N_\ell = 1468365182$ network links.

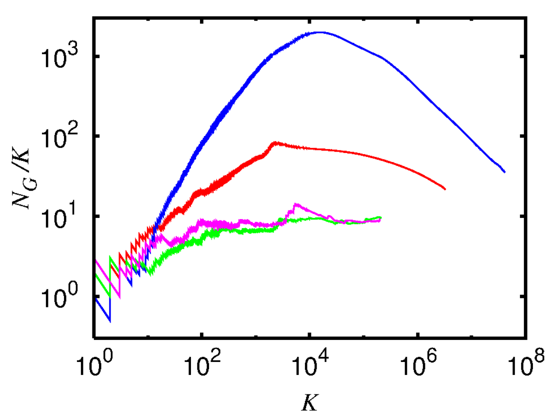
Matrix structure in K-rank order:



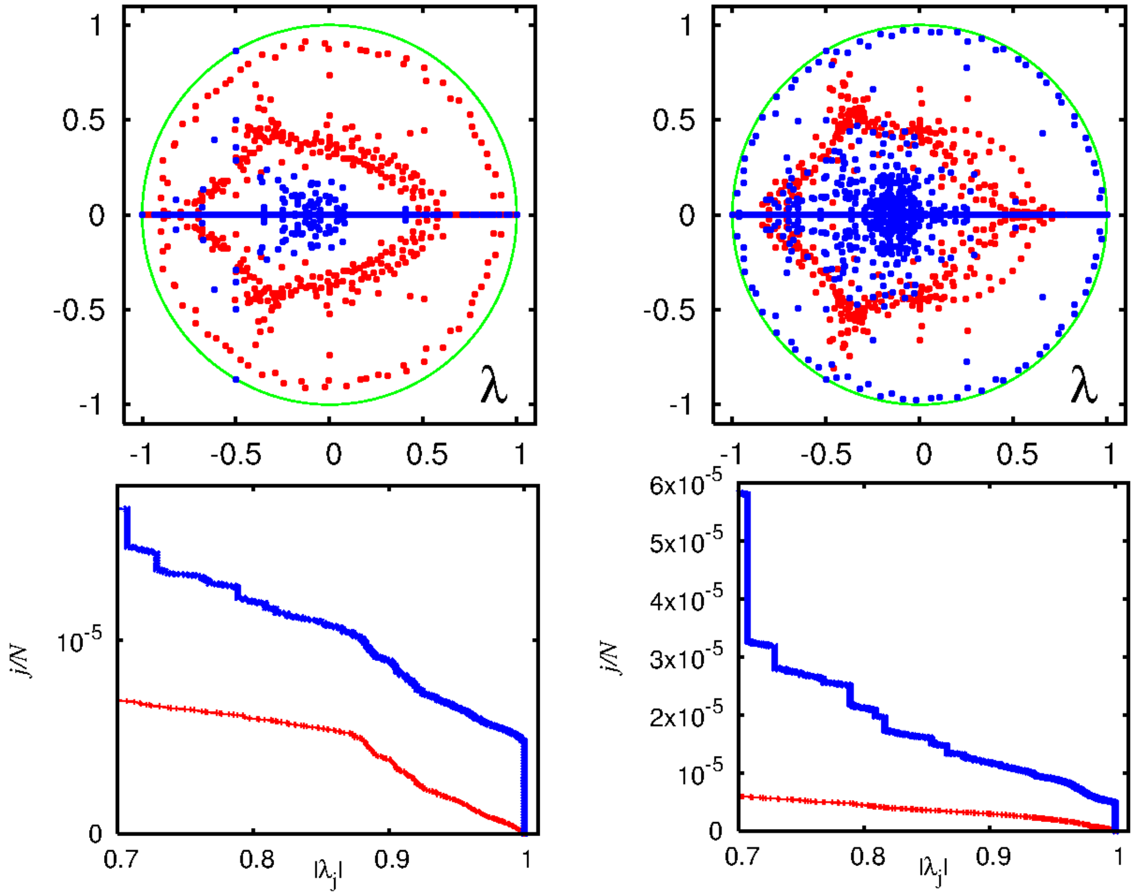


Number N_G of non-empty matrix elements in $K \times K$-square:





Spectrum for the Twitter network



$n_A = 640 \quad \Rightarrow \quad$ requires ~ 200 GB of RAM memory.

Random Perron-Frobenius matrices

Construct random matrix ensembles G_{ij} such that:

$G_{ij} \geq 0$, G_{ij} are (approximately) non-correlated and distributed with the same distribution $P(G_{ij})$ (of finite variance σ^2),

$$\sum_j G_{ij} = 1 \quad \Rightarrow \quad \langle G_{ij} \rangle = 1/N$$

$\Rightarrow$ average of G has one eigenvalue $\lambda_1 = 1$ ($\Rightarrow$ “flat” PageRank) and other eigenvalues $\lambda_j = 0$ (for $j \neq 1$).

degenerate perturbation theory for the fluctuations $\Rightarrow$ circular eigenvalue density with $R = \sqrt{N}\sigma$ and one unit eigenvalue.

Different variants of the model:

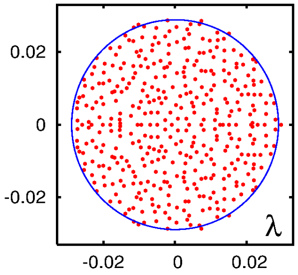
full $\quad \Rightarrow \quad R = 1/\sqrt{3N}$

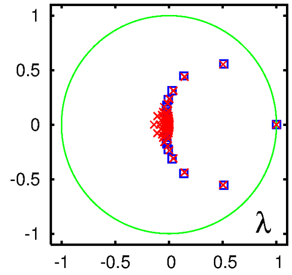
sparse with Q non-zero elements per column $\quad \Rightarrow \quad R \sim 1/\sqrt{Q}$

power law with $P(G) \sim G^{-b}$ for $2 < b < 3 \quad \Rightarrow \quad R \sim N^{1-b/2}$

Numerical verification:

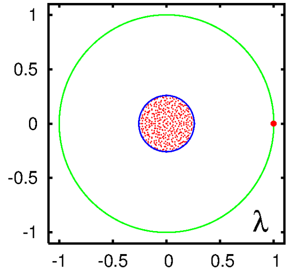
uniform full:
$N = 400$

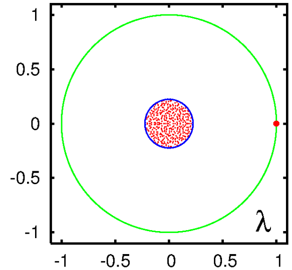




triangular
random and
average

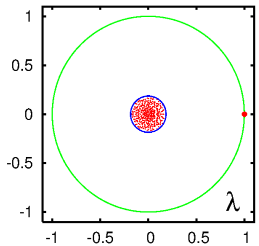
uniform sparse:
$N = 400$,
$Q = 20$

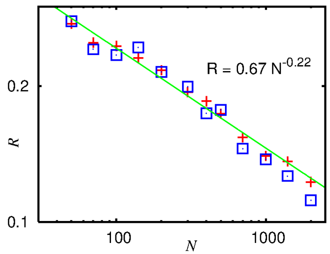




constant sparse:
$N = 400$,
$Q = 20$

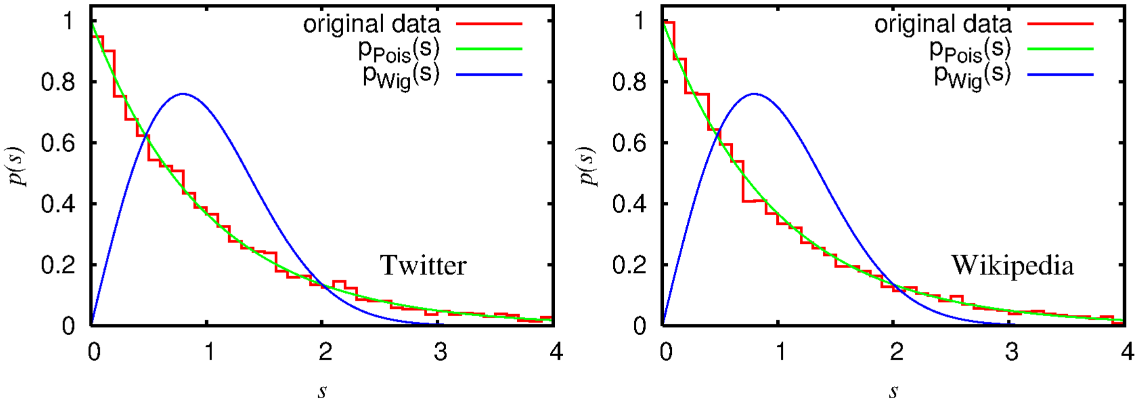
power law:
$b = 2.5$





power law case:
$R_{\mathrm{th}} \sim N^{-0.25}$

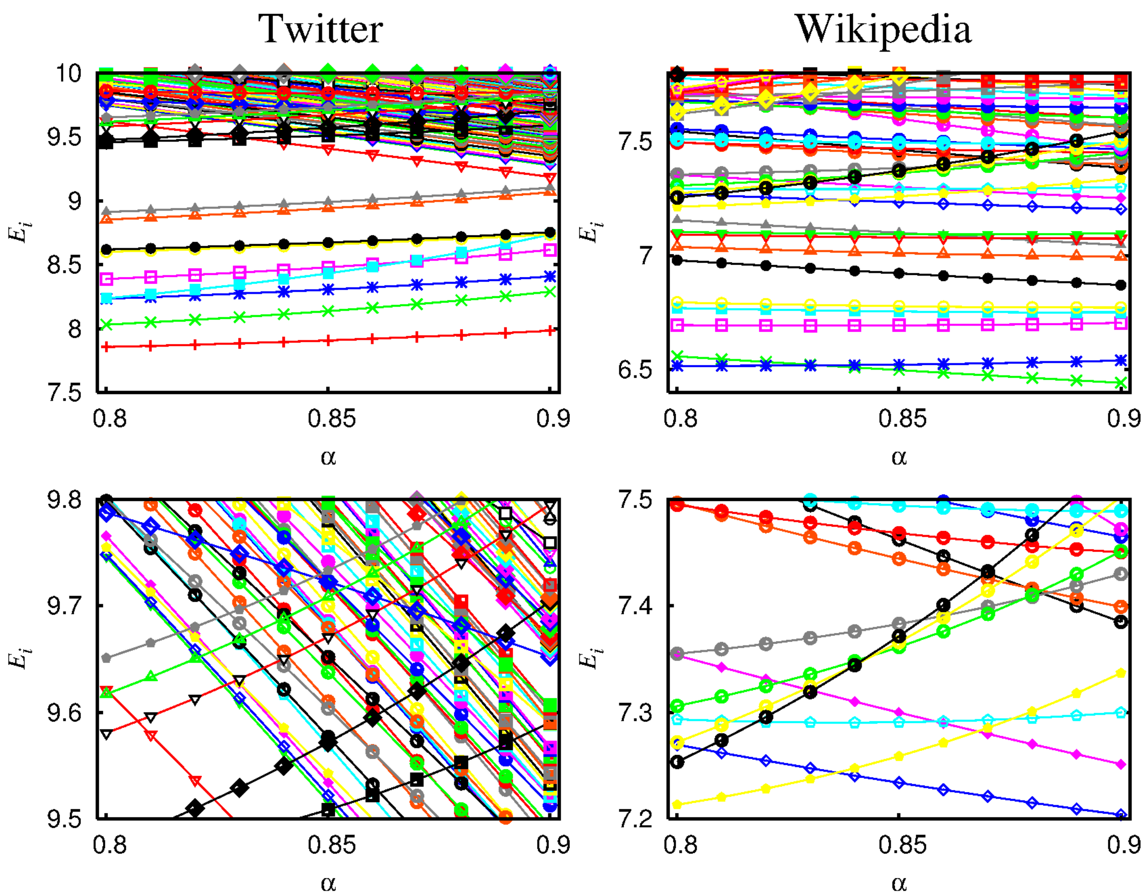
Poisson statistics of PageRank



Identify PageRank values to “energy-levels”:

$$P(i) = \exp(-E_i/T)/Z$$

with $Z = \sum_i \exp(-E_i/T)$ and an effective temperature T (can be choosen: $T = 1$).

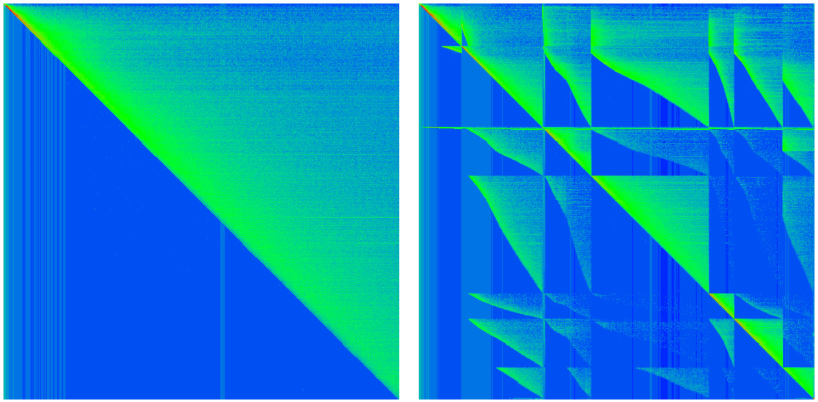


Parameter dependance of $E_i = -\ln(P(i))$ on the damping factor α.

Physical Review network

$N = 463347$ nodes and $N_\ell = 4691015$ links.

Coarse-grained matrix structure (500×500 cells):



left: time ordered, right: journal and then time ordered

“11” Journals of Physical Review: (Phys. Rev. Series I), Phys. Rev., Phys. Rev. Lett., (Rev. Mod. Phys.), Phys. Rev. A, B, C, D, E, (Phys. Rev. STAB and Phys. Rev. STPER).

$\Rightarrow$ nearly triangular matrix structure of adjacency matrix: most citations links $t \to t'$ are for $t > t'$ (“past citations”) but there is a small number ($12126 = 2.6 \times 10^{-3} N_\ell$) of links $t \to t'$ with $t \leq t'$ corresponding to ***future citations***.

Strong numerical problems due to large Jordan subspaces!

Triangular approximation

Remove the small number of links due to "future citations".

Semi-analytical diagonalization is possible:

$$\boxed{S = S_0 + e\,d^T/N}$$

where $e_n = 1$ for all nodes n, $d_n = 1$ for dangling nodes n and $d_n = 0$ otherwise. S_0 is the pure link matrix which is ***nil-potent***:

$$\boxed{S_0^l = 0} \quad \text{with } l = 352.$$

Let ψ be an eigenvector of S with eigenvalue λ and $C = d^T\psi$.

If $C = 0 \;\Rightarrow\; \psi$ eigenvector of $S_0 \;\Rightarrow\; \lambda = 0$ since S_0 nil-potent.

These eigenvectors belong to large Jordan blocks and are responsible for the numerical problems.

If $C \neq 0 \;\Rightarrow\; \lambda \neq 0$ since the equation $S_0\psi = -C\, e/N$ does not have a solution $\Rightarrow\; \lambda\mathbf{1} - S_0$ invertible.

$$\Rightarrow\; \psi = C\,(\lambda\mathbf{1} - S_0)^{-1}\, e/N = \frac{C}{\lambda}\sum_{j=0}^{l-1}\left(\frac{S_0}{\lambda}\right)^j e/N \quad .$$

$$\text{From } \lambda^l = (d^T\psi/C)\lambda^l \;\Rightarrow\; \boxed{\mathcal{P}_r(\lambda) = 0}$$

with the ***reduced polynomial*** of degree $l = 352$:

$$\boxed{\mathcal{P}_r(\lambda) = \lambda^l - \sum_{j=0}^{l-1}\lambda^{l-1-j}\, c_j = 0 \quad , \quad c_j = d^T\, S_0^j\, e/N \, .}$$

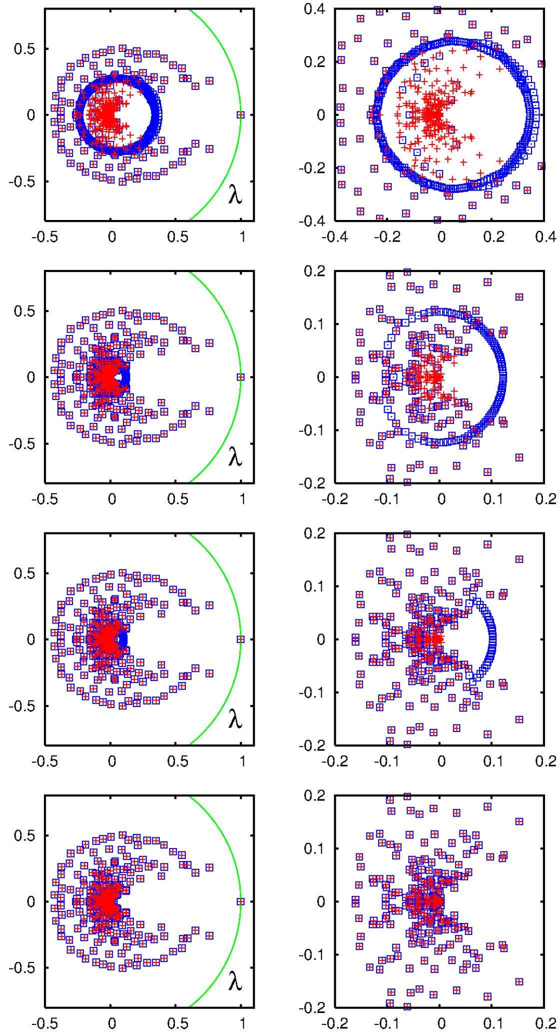
$\Rightarrow$ at most $l = 352$ eigenvalues $\lambda \neq 0$ which can be numerically determined as the zeros of $\mathcal{P}_r(\lambda)$.

<u>However: still numerical problems:</u>

- $c_{l-1} \approx 3.6 \times 10^{-352}$
- alternate sign problem with a strong loss of significance.
- big sensitivity of eigenvalues on c_j

Solution:

Using the multi precision library GMP with 256 binary digits the zeros of $\mathcal{P}_r(\lambda)$ can be determined with accuracy $\sim 10^{-18}$.
Furthermore the Arnoldi method can also be implemented with higher precision.

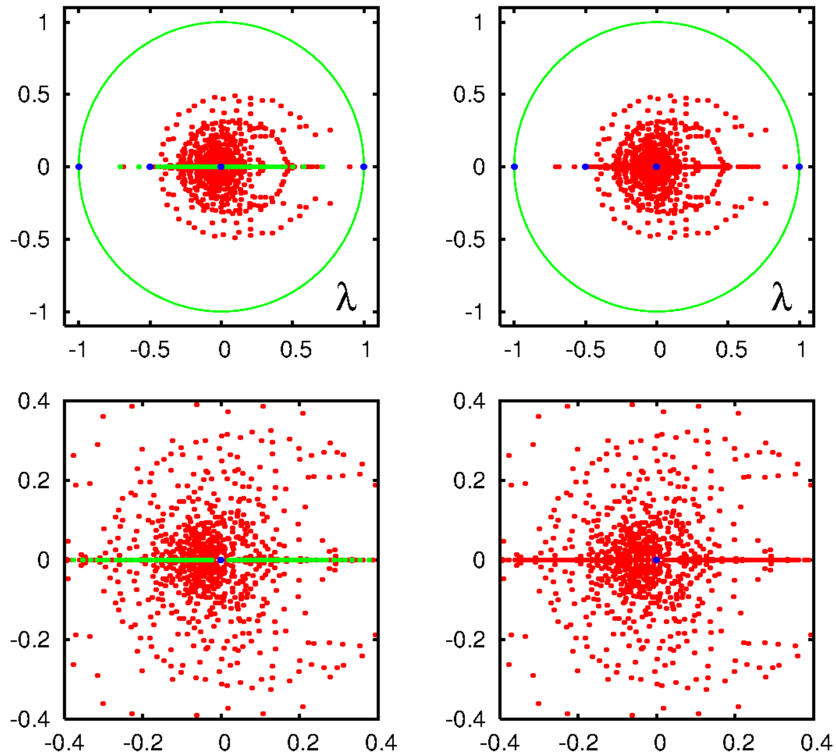


red crosses: zeros of $\mathcal{P}_r(\lambda)$ from 256 binary digits calculation

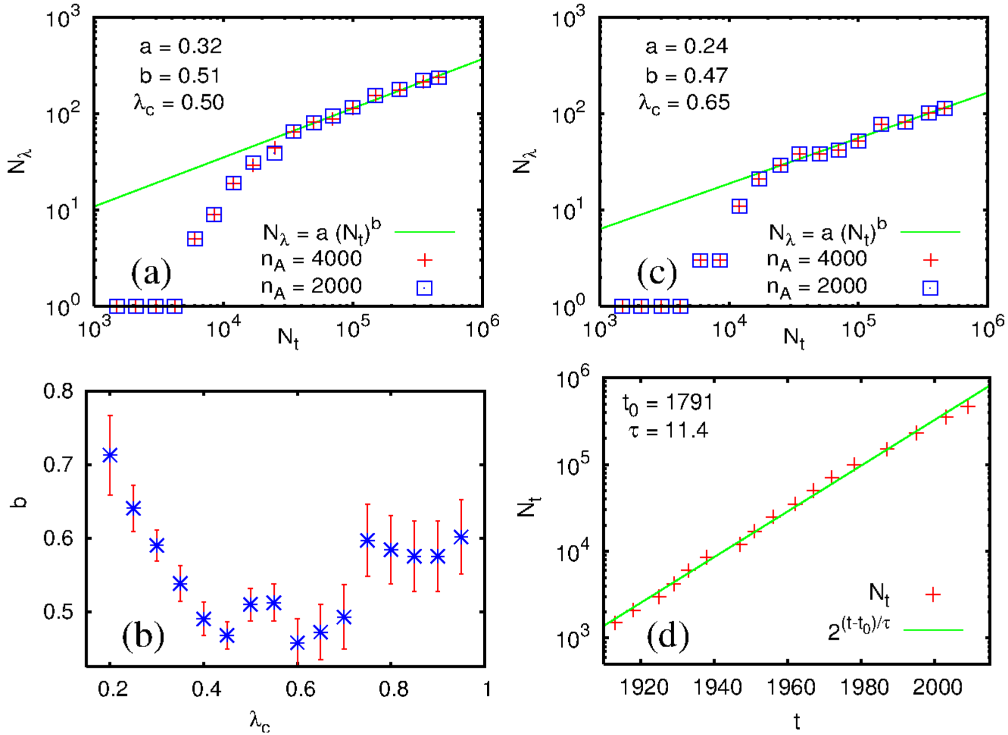
blue squares: eigenvalues from Arnoldi method with 52, 256, 512, 1280 binary digits. In the last case: $\Rightarrow$ break off at $n_A = 352$ with vanishing coupling element.

Full Physical Review network

Accurate eigenvalue spectrum for the full Physical Review network by a new rational interpolation method (left) and the HP Arnoldi method (right):



Fractal Weyl law



$N_\lambda =$ number of complex eigenvalues with $\lambda_c \leq |\lambda| \leq 1$.
$N_t =$ reduced network size of Physical Review at time t.

$$N_\lambda = a N_t^b$$

Perron-Frobenius matrix for chaotic maps

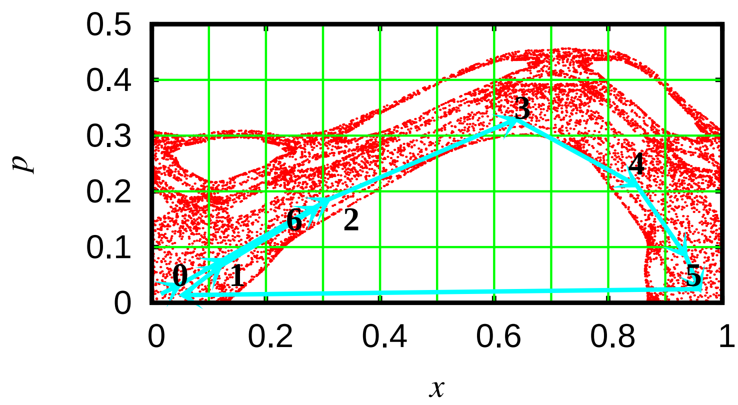
A new variant of the ***Ulam Method*** to construct the ***Perron-Frobenius matrix*** for the case of a mixed phase space:
Subdivide phase space in square cells of size M^{-1} and iterate a classical trajectory ($t \sim 10^{11} - 10^{12}$) and attribute a new number to each new cell which is entered. At the same time count the number of transitions from cell i to cell j ($\Rightarrow\ n_{ji}$) $\Rightarrow$ $N \times N$-PF-Matrix (N=number of non-empty cells) by:

$$\boxed{G_{ji} = \frac{n_{ji}}{\sum_l n_{li}}}$$

Example: Chirikov map at

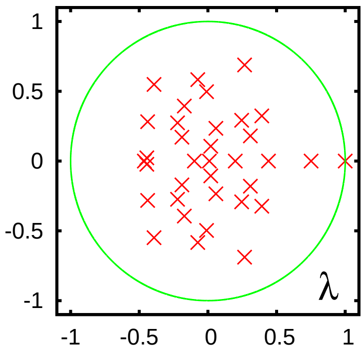
$k = k_c = 0.971635406$

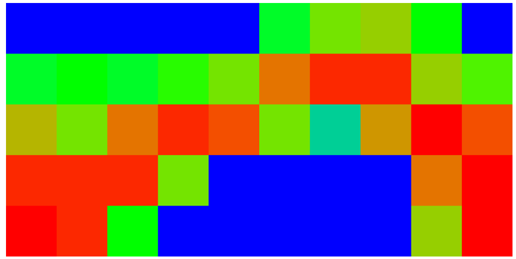
with $M = 10$.



Eigenvalues

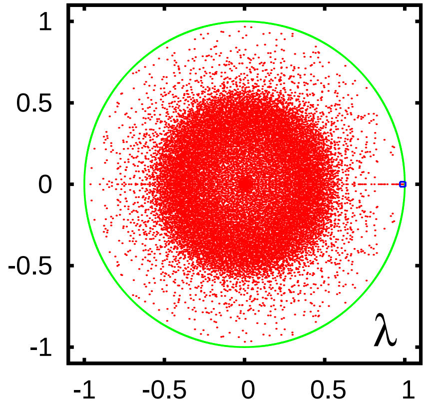
for $M = 10$, $t = 10^6$ and $N = 35$

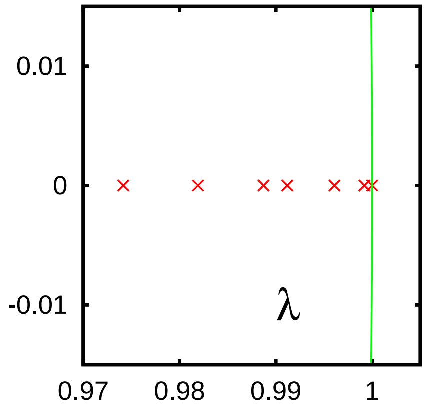




Phase space representation of the eigenvector for $\lambda_0 = 1$.

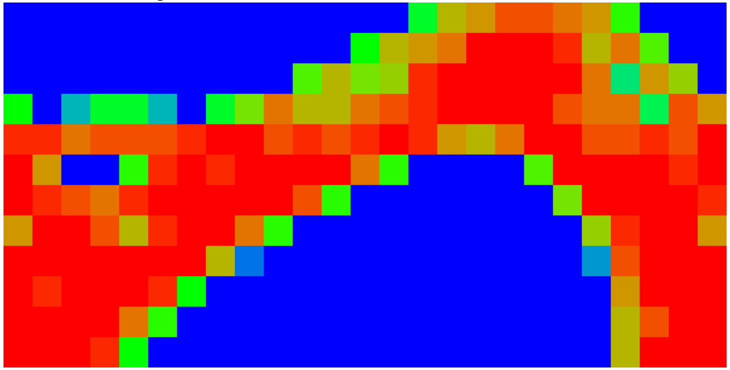
for $M = 280$, $t = 10^{12}$ and $N = 16609$



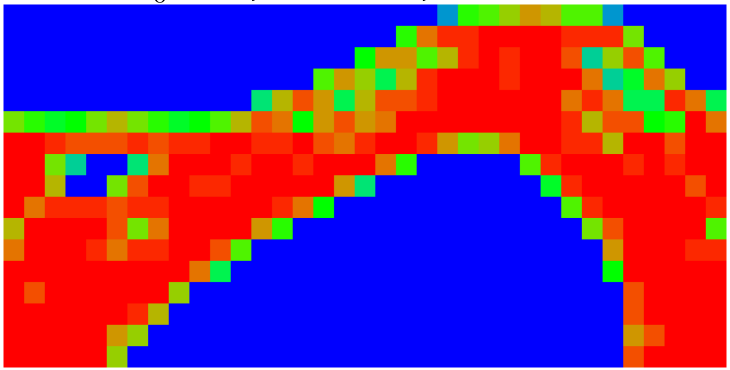


Eigenvectors

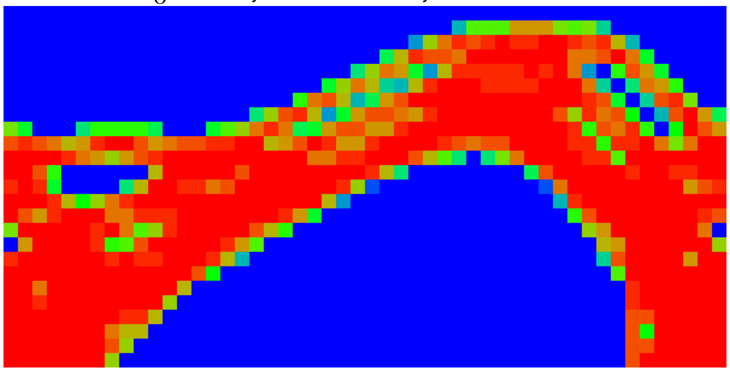
$\lambda_0 = 1,\, M = 25,\, N = 177$



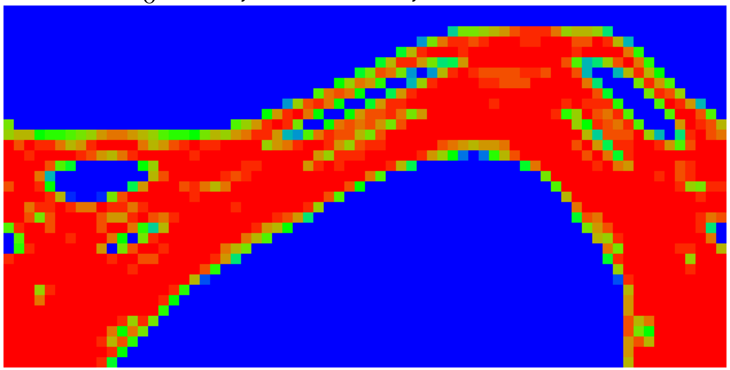
$\lambda_0 = 1,\, M = 35,\, N = 332$

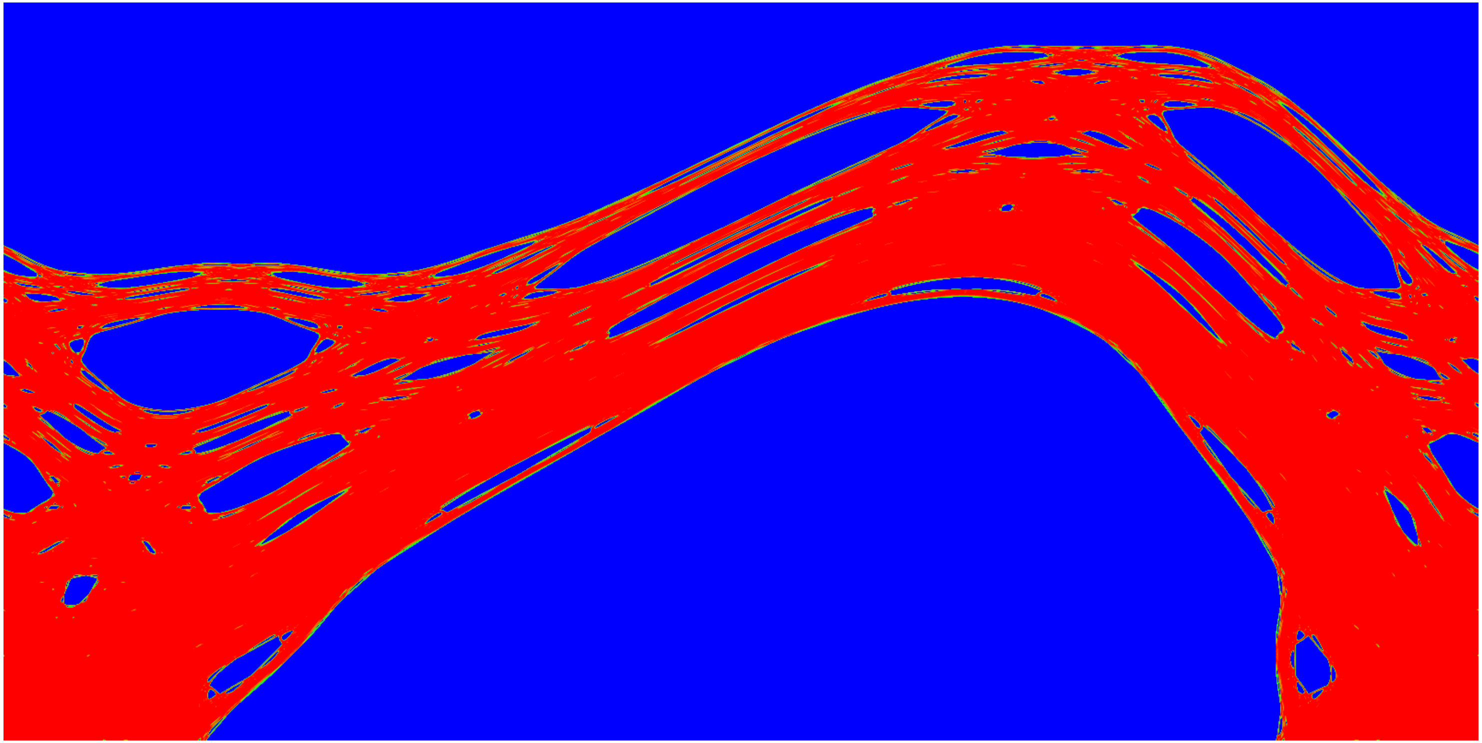


$\lambda_0 = 1,\, M = 50,\, N = 641$

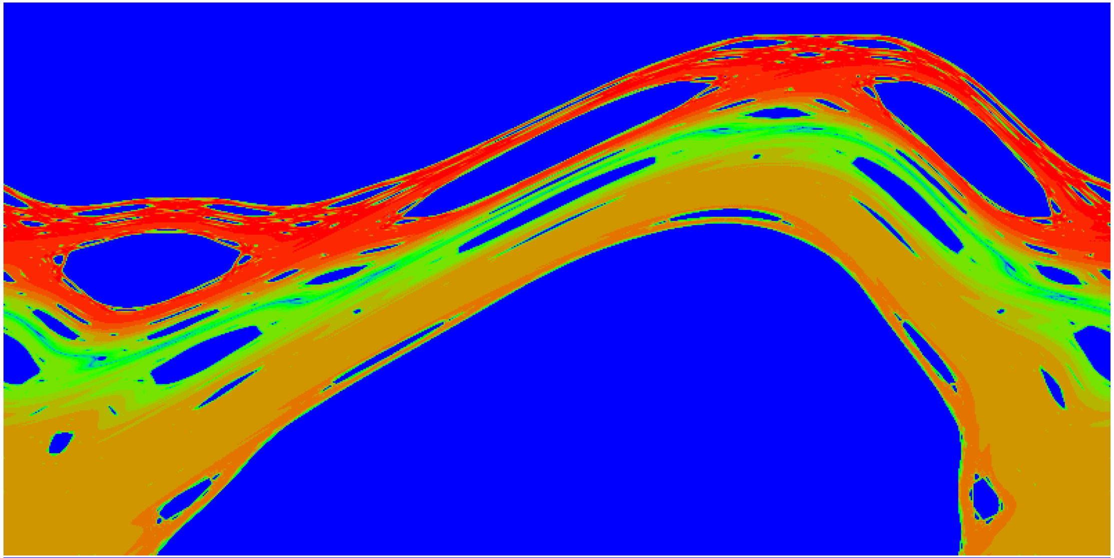


$\lambda_0 = 1,\, M = 70,\, N = 1189$



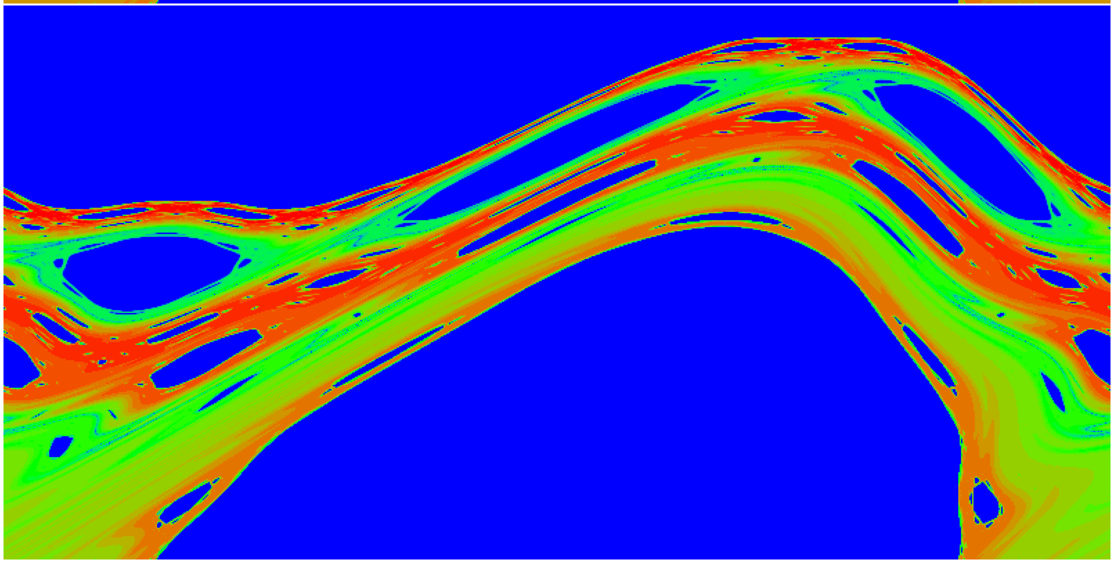


$\lambda_0 = 1$, $M = 1600$, $N = 494964$, $n_A = 3000$



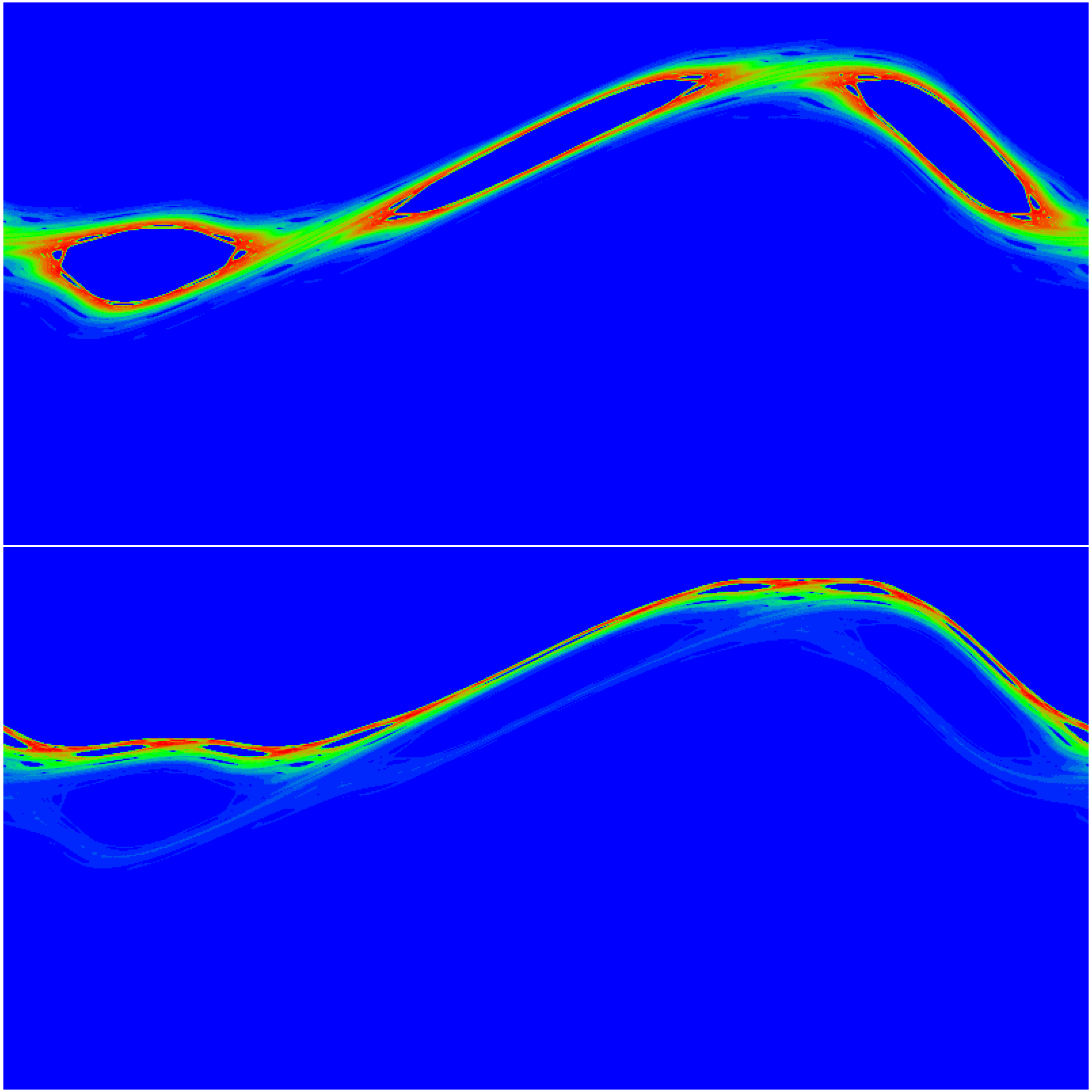
$\lambda_1 =$
0.99980431

$M = 800$
$N = 127282$
$n_A = 2000$



$\lambda_2 =$
0.99878108

$M = 800$
$N = 127282$
$n_A = 2000$

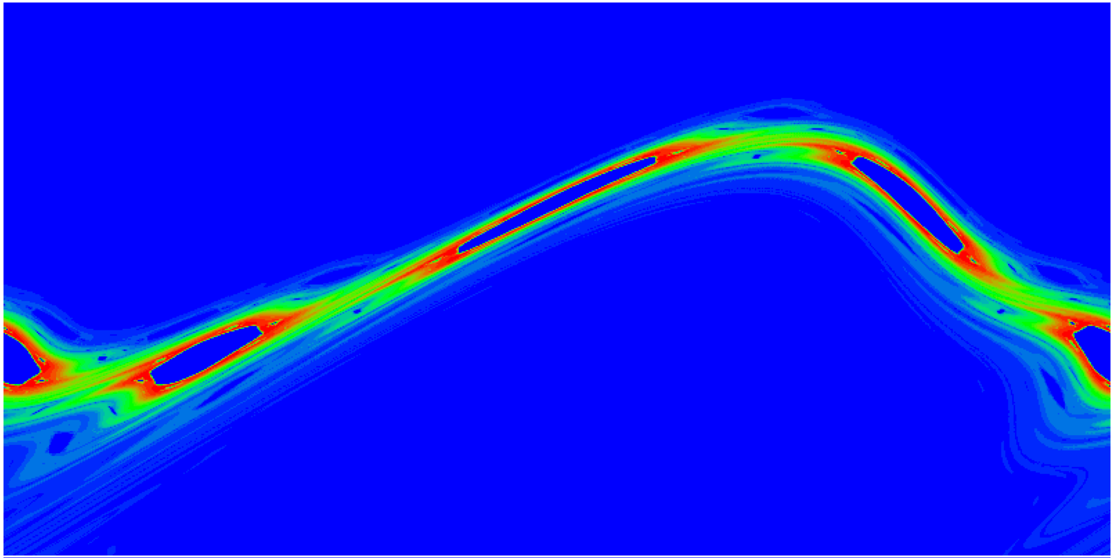


$\lambda_6 =$
-0.49699831
$+i\,0.86089756$
$\approx |\lambda_6|\, e^{i\,2\pi/3}$

$M = 800$
$N = 127282$
$n_A = 2000$

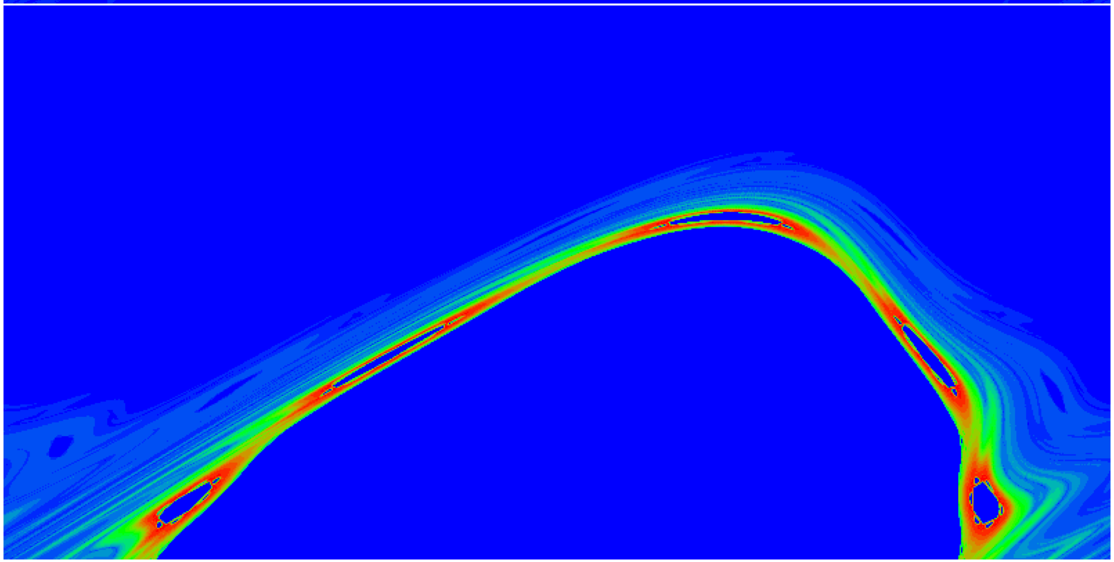
$\lambda_{19} =$
-0.71213331
$+i\,0.67961609$
$\approx |\lambda_{19}|\, e^{i\,2\pi(3/8)}$

$M = 800$
$N = 127282$
$n_A = 2000$



$\lambda_8 =$
0.00024596
$+i\,0.99239222$
$\approx |\lambda_8|\,e^{i\,2\pi/4}$

$M = 800$
$N = 127282$
$n_A = 2000$



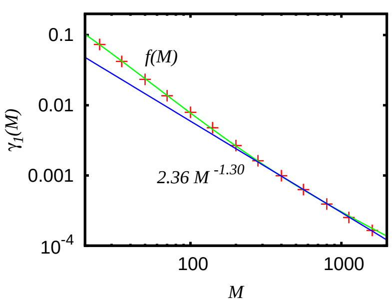
$\lambda_{13} =$
0.30580631
$+i\,0.94120900$
$\approx |\lambda_{13}|\,e^{i\,2\pi/5}$

$M = 800$
$N = 127282$
$n_A = 2000$

Extrapolation of eigenvalues

$(\gamma_j = -2\ln(|\lambda_j|))$

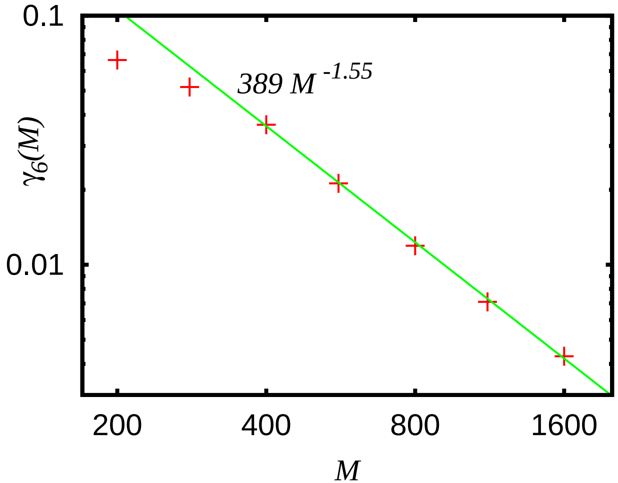
$\gamma_1(M)$ in the limit $M \to \infty$:

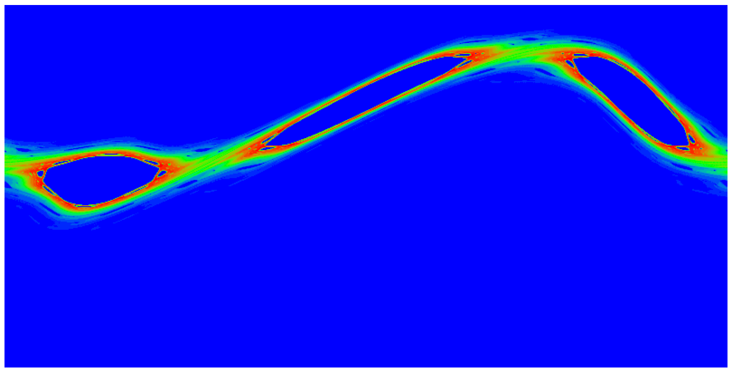


$$f(M) = \frac{D}{M}\,\frac{1+\frac{C}{M}}{1+\frac{B}{M}}$$

$D = 0.245$
$B = 13.1$
$C = 258$

$\gamma_6(M)$ in the limit $M \to \infty$:

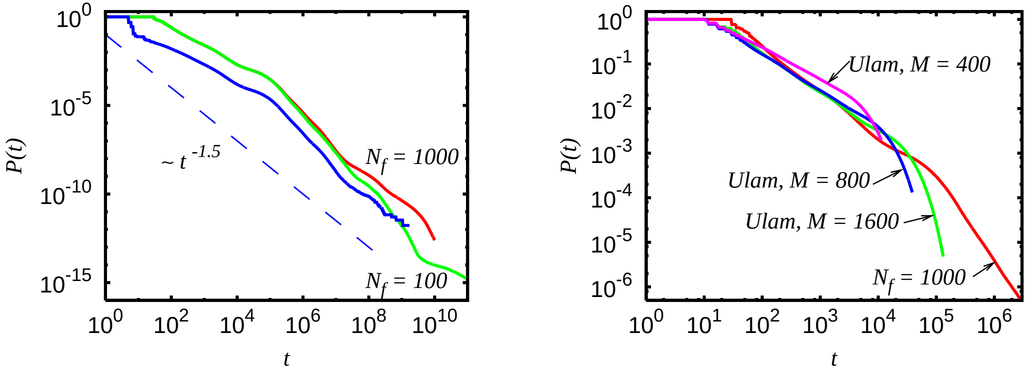




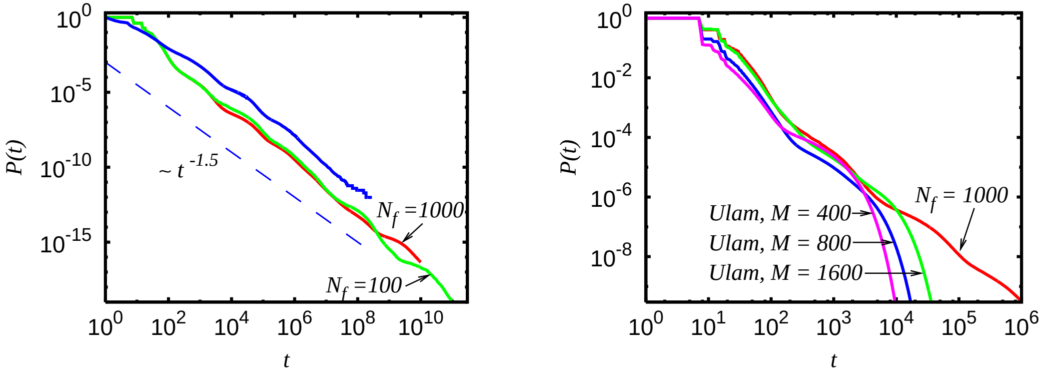
$\gamma_6(M) \approx 389\, M^{-1.55}$ for $M \geq 400$.

Absorption for $p < 0.05$

Chirikov map



Separatrix map



Red, green (left): Survial Monte-Carlo Method

Blue (left): Data of Weiss et al. PRL **89**, 239401 (2002) and Chirikov et al. PRL **89**, 239402 (2002).

References

1. D. L. Shepelyansky ***Fractal Weyl law for quantum fractal eigenstates***, Phys. Rev. E **77**, p.015202(R) (2008).

2. L. Ermann and D. L. Shepelyansky, ***Ulam method and fractal Weyl law for Perron-Frobenius operators***, Eur. Phys. J. B **75**, 299 (2010).

3. K. M. Frahm and D. L. Shepelyansky, ***Ulam method for the Chirikov standard map***, Eur. Phys. J. B **76**, 57 (2010).

4. K. M. Frahm, B. Georgeot and D. L. Shepelyansky, ***Universal emergence of PageRank***, J. Phys. A: Math. Theor. **44**, 465101 (2011).

5. K. M. Frahm, A. D. Chepelianskii and D. L. Shepelyansky, ***PageRank of integers***, arxiv:1205.6343[cs.IR] (2012).

6. K. M. Frahm and D. L. Shepelyansky, ***Google matrix of Twitter***, Eur. Phys. J. B **85**, 355 (2012).

7. L. Ermann, K. M. Frahm and D. L. Shepelyansky, **Spectral properties of Google matrix of Wikipedia and other networks**, Eur. Phys. J. B **86**, 193 (2013).

8. K. M. Frahm and D. L. Shepelyansky, **Poincaré recurrences and Ulam method for the Chirikov standard map**, Eur. Phys. J. B **86**, 322 (2013).

9. K. M. Frahm, and D. L. Shepelyansky, **Poisson statistics of PageRank probabilities of Twitter and Wikipedia networks**, Eur. Phys. J. B, **87**, 93 (2014).

10. K. M. Frahm, Y. H. Eom, and D. L. Shepelyansky, **Google matrix of the citation network of Physical Review**, Phys. Rev. E **89**, 052814 (2014).

11. L. Ermann, K. M. Frahm, and D. L. Shepelyansky, **Google matrix analysis of directed networks**, Rev. Mod. Phys. **87**, 1261 (2015).

12. K. M. Frahm, and D. L. Shepelyansky, **Reduced Google matrix**, Feb. 7 (2016), arXiv:1602.02394 [physics.soc-ph] preprint.