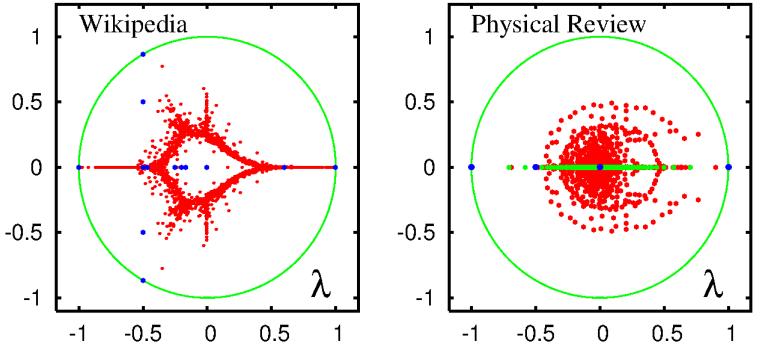


Google matrix analysis of directed networks

Lecture 3

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. L. Shepelyansky

Networks and data mining Luchon, June 27 - July 11, 2015

Contents

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:

- ***uniform full***: $P(G) = N/2$ for $0 \leq G \leq 2/N$

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

- ***uniform sparse*** with Q non-zero elements per column:
  $P(G) = Q/2$ for $0 \leq G \leq 2/Q$ with probability Q/N
  and $G = 0$ with probability $1 - Q/N$

  $\Rightarrow \quad R = 2/\sqrt{3Q}$

- ***constant sparse*** with Q non-zero elements per column:
  $G = 1/Q$ with probability Q/N
  and $G = 0$ with probability $1 - Q/N$

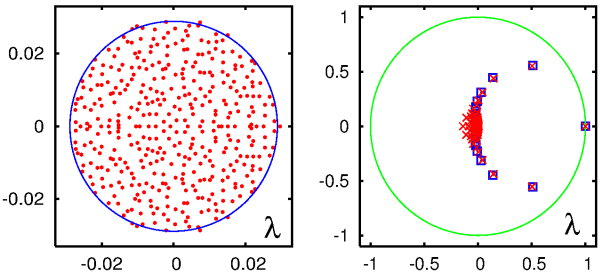
  $\Rightarrow \quad R = 1/\sqrt{Q}$

- ***powerlaw*** with $p(G) = D(1 + aG)^{-b}$ for $0 \leq G \leq 1$ and $2 < b < 3$:

  $\Rightarrow \quad R = C(b)\, N^{1-b/2} \quad , \quad C(b) = (b-2)^{(b-1)/2} \sqrt{\frac{b-1}{3-b}}$

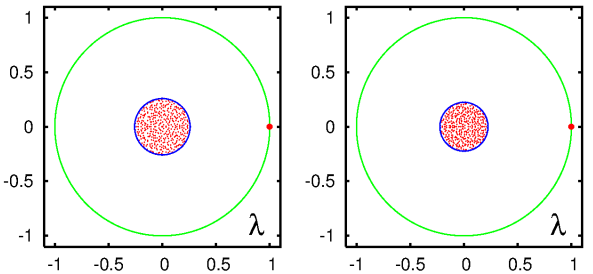
Numerical verification:

uniform full:
$N = 400$



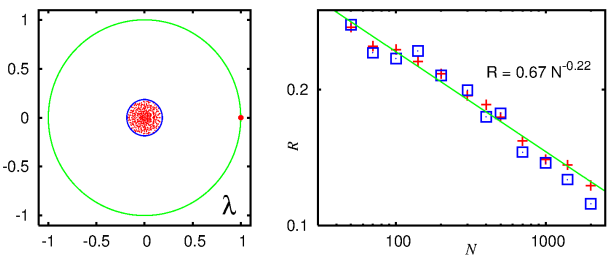
triangular random and average

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



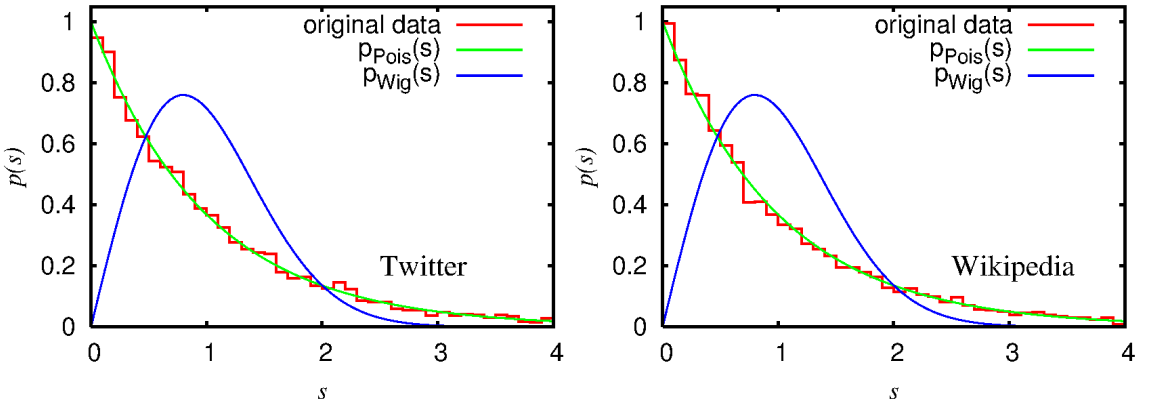
constant sparse:
$N = 400,$
$Q = 20$

power law:
$b = 2.5$



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

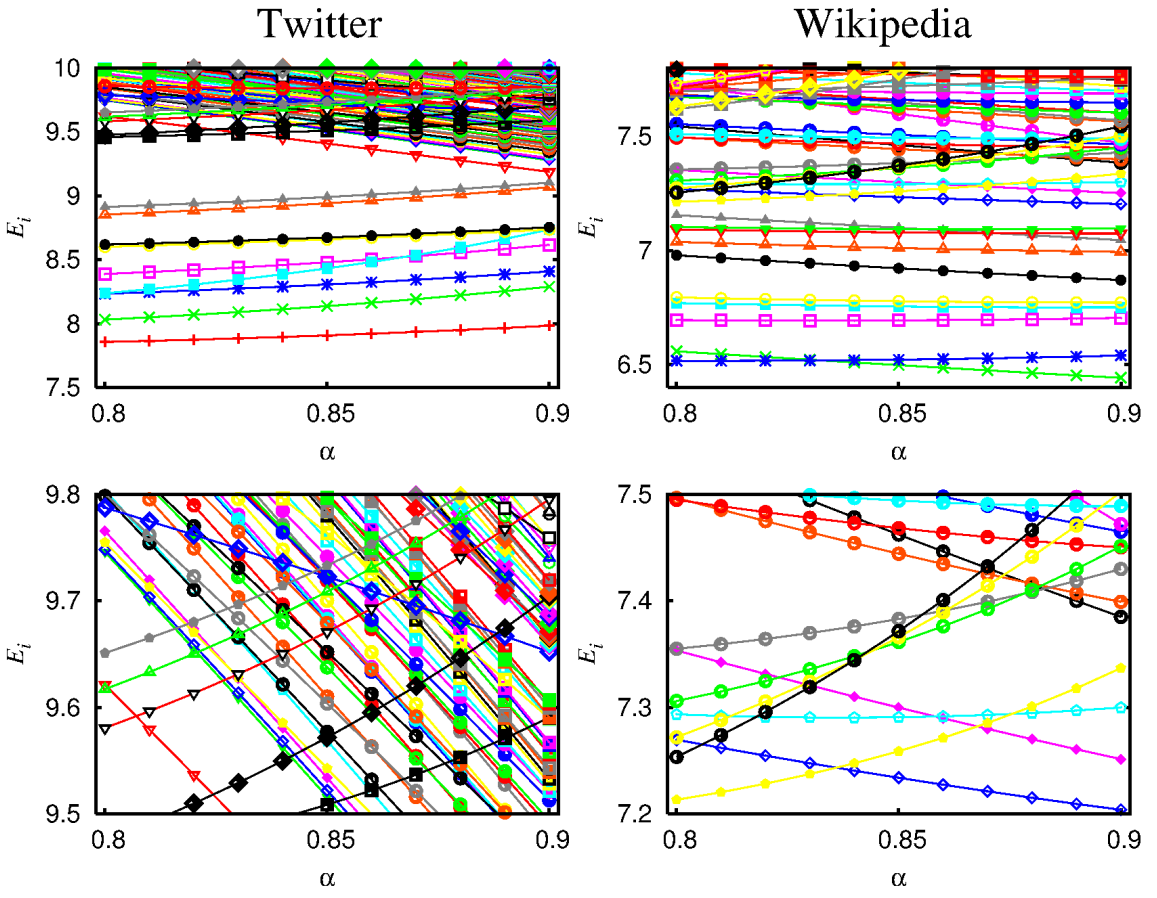
Poisson statistics of PageRank



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: $T = 1$).

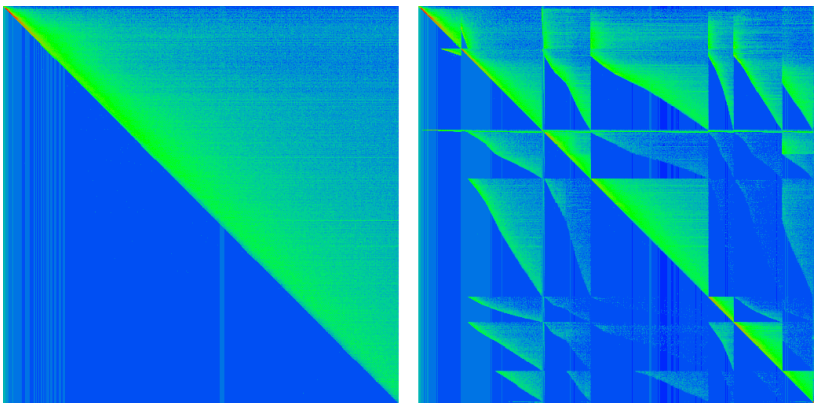


Parameter dependance of $E_i = -\ln(P_i)$ on the damping factor α.

Physical Review network

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

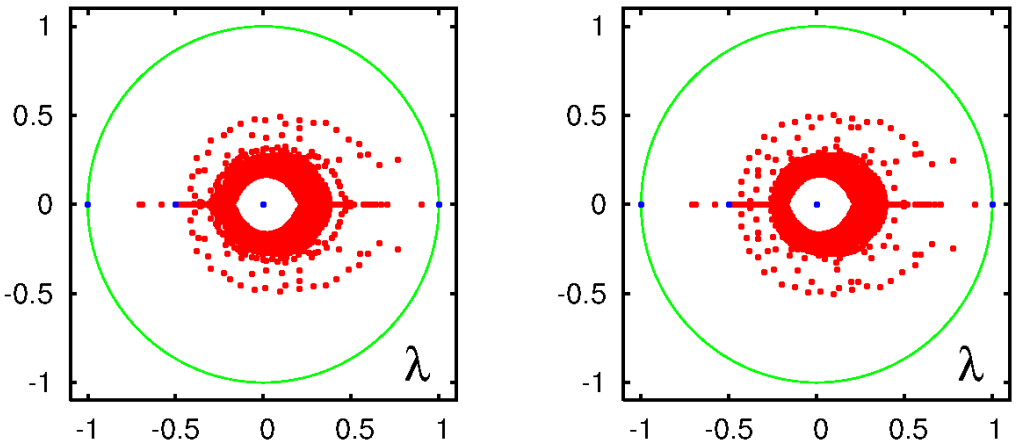
Coarse-grained matrix structure (500×500 cells):



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 small number ($12126 = 2.6 \times 10^{-3} N_\ell$) of links $t \to t'$ with $t \leq t'$ corresponding to ***future citations***.
Spectrum by "double-precision" Arnoldi method with $n_A = 8000$:



Numerical problem: eigenvalues with $|\lambda| < 0.3 - 0.4$ are not reliable!
Reason: large Jordan subspaces associated to the eigenvalue $\lambda = 0$.

"very bad" Jordan perturbation theory:
Consider a "perturbed" Jordan block of size D:

$$\begin{pmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \varepsilon & 0 & \cdots & 0 & 0 \end{pmatrix}$$

characteristic polynomial: $\lambda^D - (-1)^D \varepsilon$
$\varepsilon = 0 \quad \Rightarrow \quad \lambda = 0$
$\varepsilon \neq 0 \quad \Rightarrow \quad \lambda_j = -\varepsilon^{1/D} \exp(2\pi i j/D)$
for $D \approx 10^2$ and $\varepsilon = 10^{-16} \quad \Rightarrow$ "Jordan-cloud" of artifical eigenvalues due to rounding errors in the region $|\lambda| < 0.3 - 0.4$.

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***:

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

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.

  Note: Similar situation as in ***network of integer numbers*** where $l = [\log_2(N)]$ and numerical instability for $|\lambda| < 0.01$.

- 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 $l = 352$:

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

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

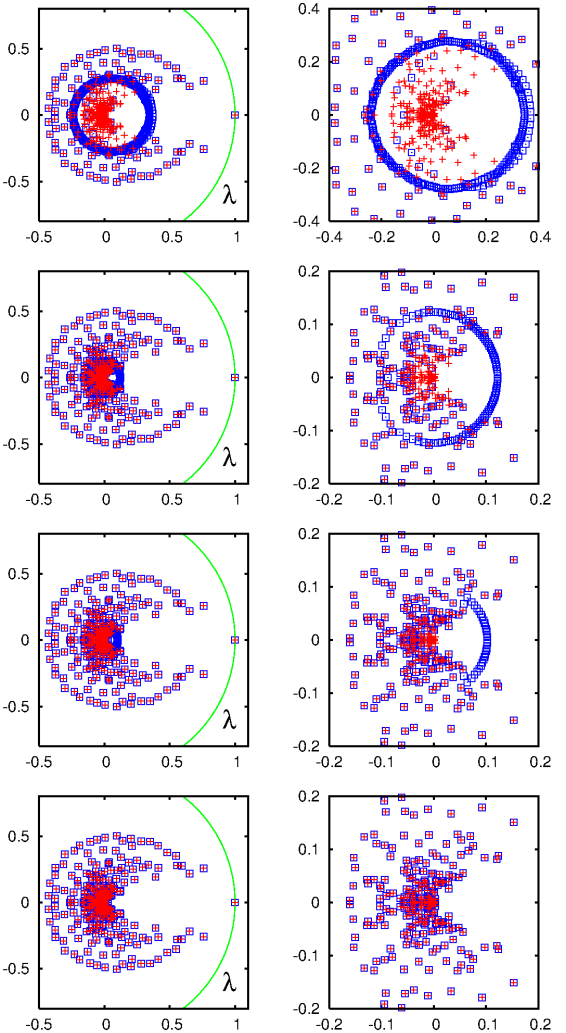
However: still numerical problems:

- $c_{l-1} \approx 3.6 \times 10^{-352}$
- 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

Complications due to small number of "future citations" which break the triangular structure $\Rightarrow$ two groups of eigenvectors ψ:

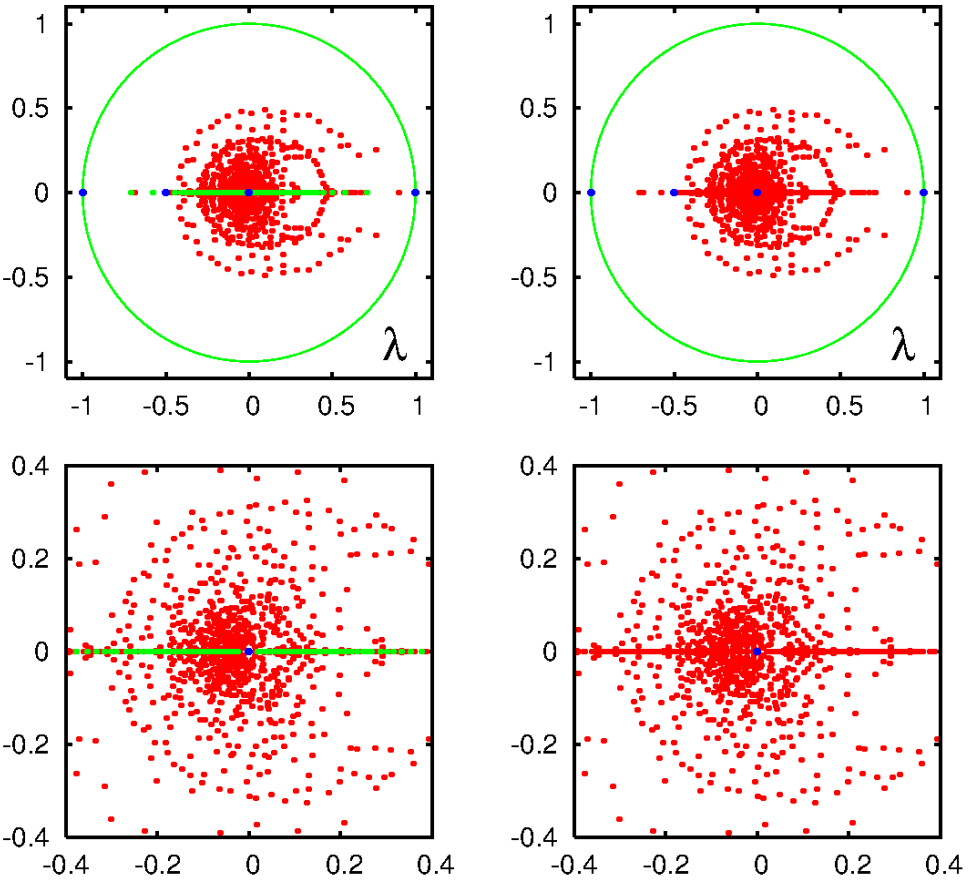
1. $d^T\psi = 0 \quad \Rightarrow \quad$ common eigenvector/eigenvalue of S and S_0, essentially : $\lambda = \pm 1/\sqrt{n}$ with $n = 1,\ 2,\ 3,\ \ldots$ and large degeneracies.

2. $d^T\psi \neq 0 \quad \Rightarrow \quad \mathcal{R}(\lambda) = 0$ with a rational function:

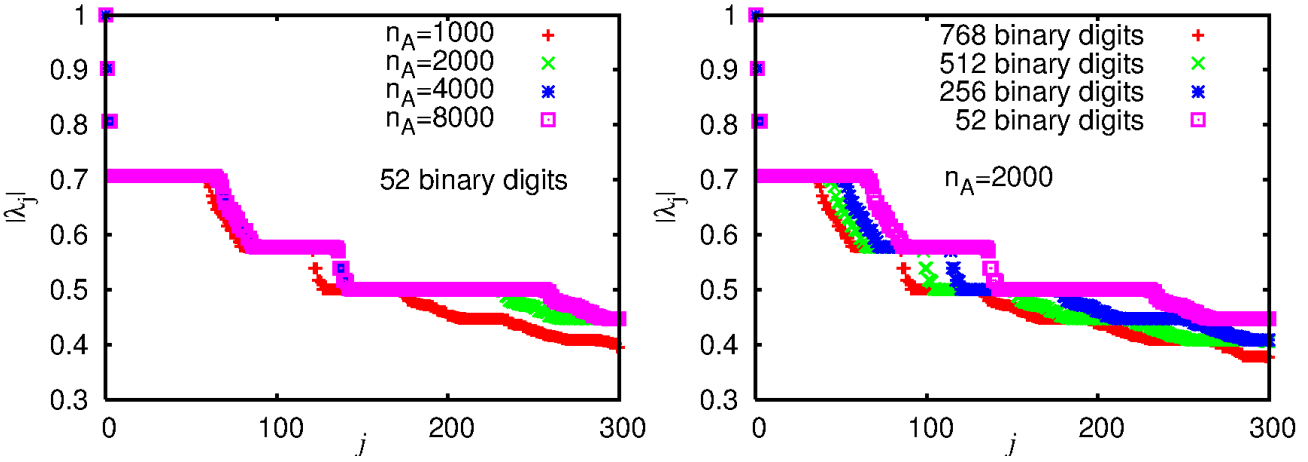
$$\mathcal{R}(\lambda) = 1 - \sum_{j=0}^{\infty} c_j \lambda^{-1-j} \quad , \quad c_j = d^T\, S_0^j\, e/N$$

with convergence for $|\lambda| > \rho_1 \approx 0.9024$. The zeros of $\mathcal{R}(\lambda)$ with $|\lambda| < \rho_1$ can be determined by a rational interpolation using many support points with $|z_j| = 1$ where the series to evaluate $\mathcal{R}(z_i)$ converges well $\quad \Rightarrow \quad$ ***rational interpolation method*** (requires also high precision computations, details in Appendix).

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



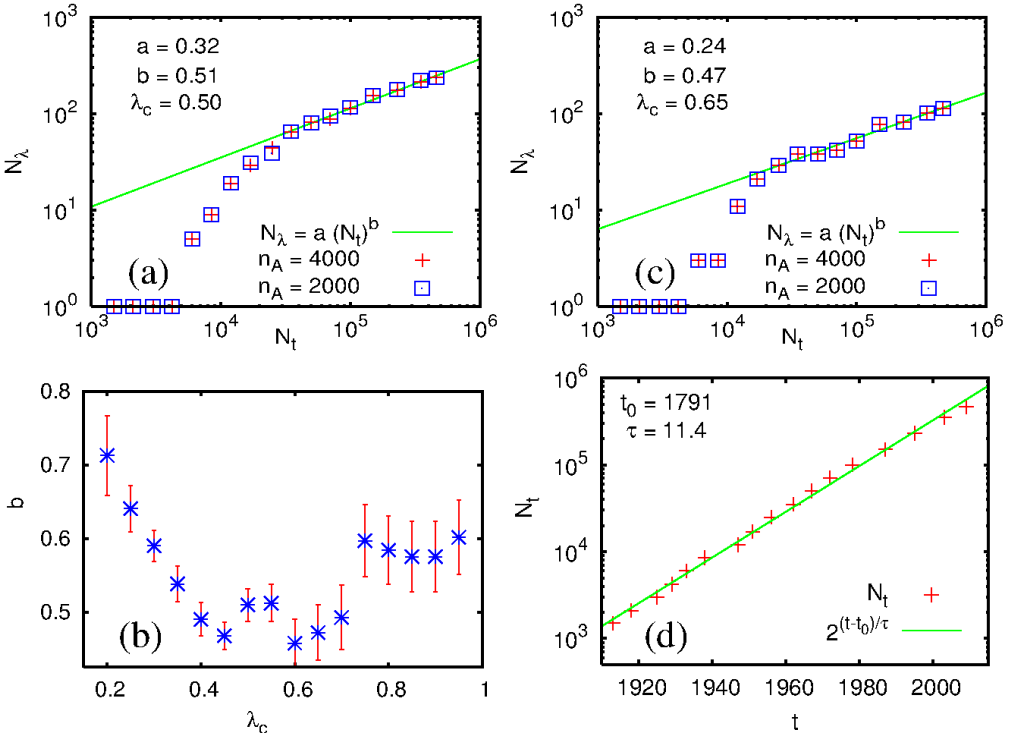
Degeneracies



High precision in Arnoldi method is “bad” to count the degeneracy of certain degenerate eigenvalues (of first group).

In theory the Arnoldi method cannot find several eigenvectors for degenerate eigenvalues, a shortcoming which is (partly) “repaired” by rounding errors.

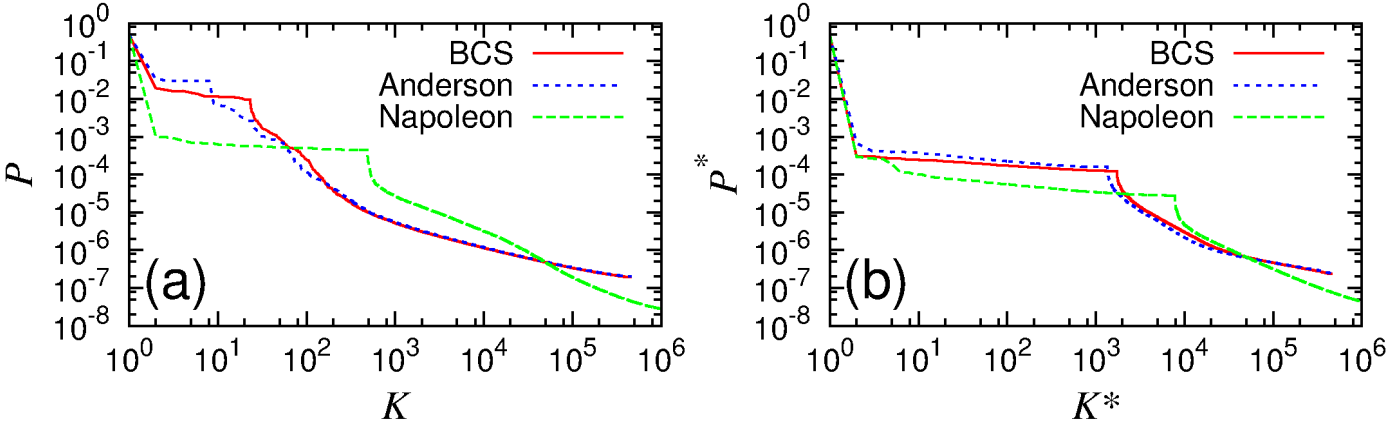
Fractal Weyl law



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

$$N_\lambda = aN_t^b$$

ImpactRank for influence propagation



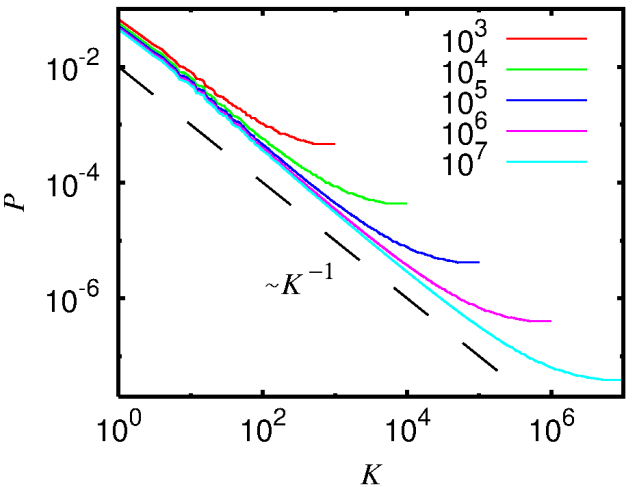
$$v_f = \frac{1-\gamma}{1-\gamma G}\, v_0 \quad , \quad v_f^* = \frac{1-\gamma}{1-\gamma G^*}\, v_0$$

v_f = "PageRank" of $\tilde{G}$ with:

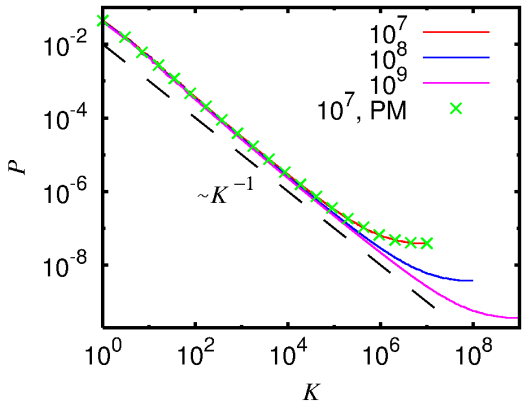
$$\tilde{G} = \gamma\, G + (1-\gamma)\, v_0\, e^T$$

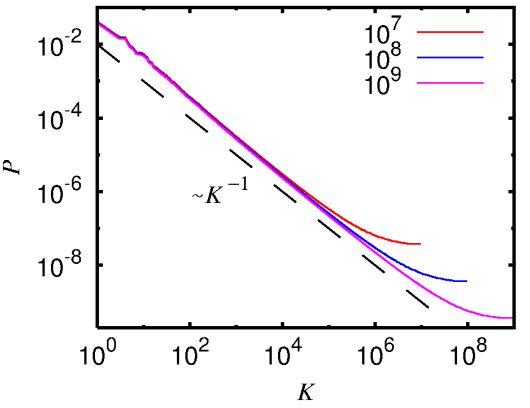
Integer network

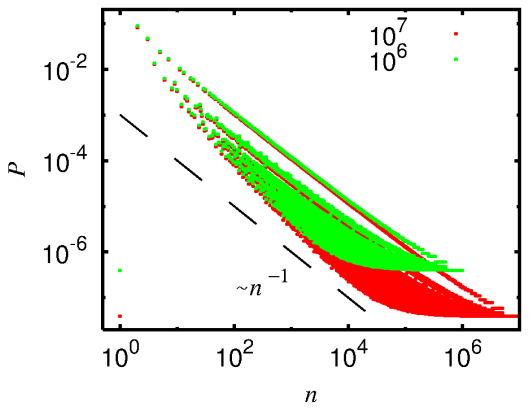
Consider the integers $n \in \{1, \ldots, N\}$ and construct an adjacency matrix by $A_{mn} = k$ where k is the largest integer such that m^k is a divisor of n if $1 < m < n$ and $A_{mn} = 0$ if $m = 1$ or $m = n$ (note $A_{mn} = k = 0$ if m is not a divisor of n). Construct S and G in the usual way from A.

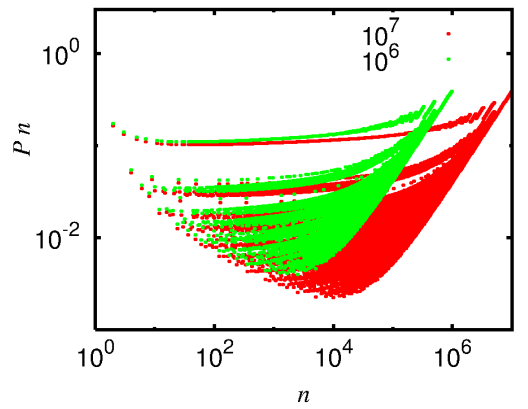


PageRank

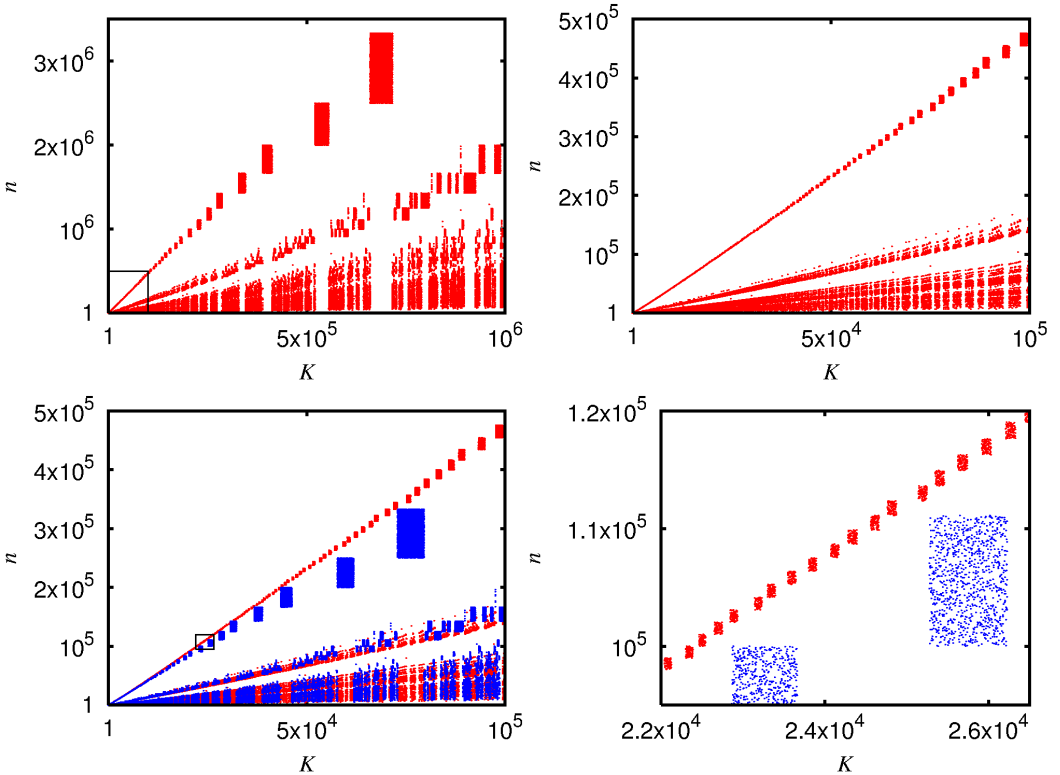

10^{7}
10^{8}
10^{9}
10^{7}, PM
$\sim K^{-1}$
P
K


10^{7}
10^{8}
10^{9}
$\sim K^{-1}$
P
K


10^{7}
10^{6}
$\sim n^{-1}$
P
n


10^{7}
10^{6}
Pn
n

Dependence of n on K-index



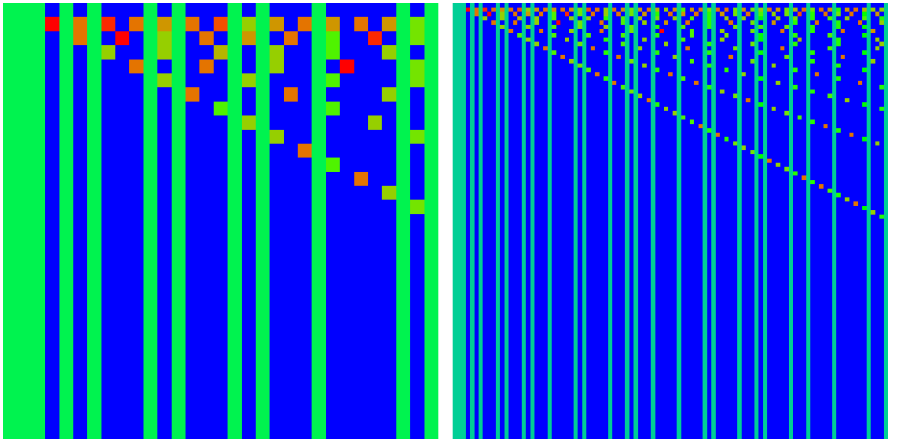
red: $N = 10^7$

blue: $N = 10^6$

"New order" of integers: $K = 1, 2, \ldots, 32 \Rightarrow n = 2, 3, 5, 7, 4, 11, 13, 17, 6, 19, 9, 23, 29, 8, 31, 10, 37, 41, 43, 14, 47, 15, 53, 59, 61, 25, 67, 12, 71, 73, 22, 21.$

Semi-analytical determination of spectrum, PageRank and eigenvectors

Matrix structure:



$$S = S_0 + v\, d^T$$

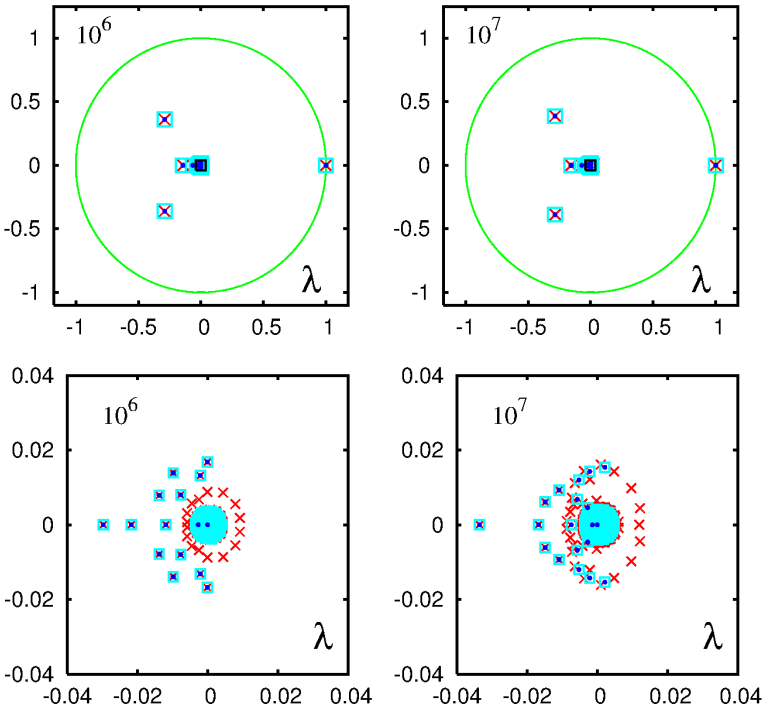
where $v = e/N$, $d_j = 1$ for dangling nodes (primes and 1) and $d_j = 0$ otherwise. S_0 is the pure link matrix which is ***nil-potent***:

$$S_0^l = 0$$

with $l = [\log_2(N)] \ll N$

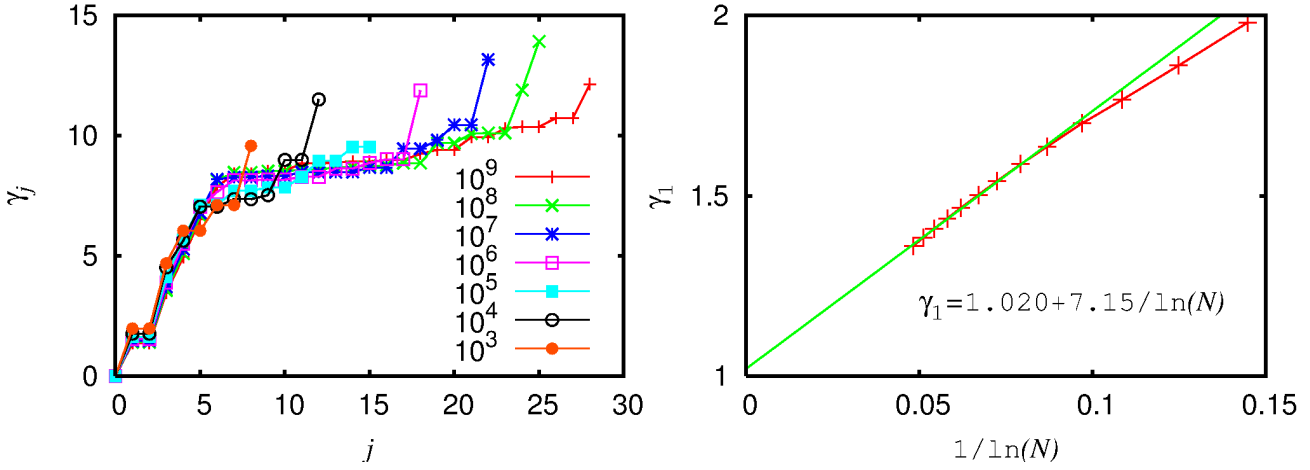
$\Rightarrow$ same theory as for the Phys.-Rev. Network.

Spectrum I



blue dots: semi-analytical eigenvalues as zeros from $\mathcal{P}_r(\lambda)$ (or eigenvalues of $\bar{S}$).
red crosses: Arnoldi method with random initial vector and $n_A = 1000$.
light blue boxes: Arnoldi method with constant initial vector $v = e/N$ and $n_A = 1000$.

Spectrum II



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

Large N limit of γ_1 with the scaling parameter: $1/\ln(N)$.

Note:

$$c_0 = d^T v = \frac{1}{N} \sum_{j=1}^{N} d_j = \frac{1 + \pi(N)}{N} \approx \frac{1}{\ln(N)}$$

where $\pi(N)$ is the number of primes below N.

References

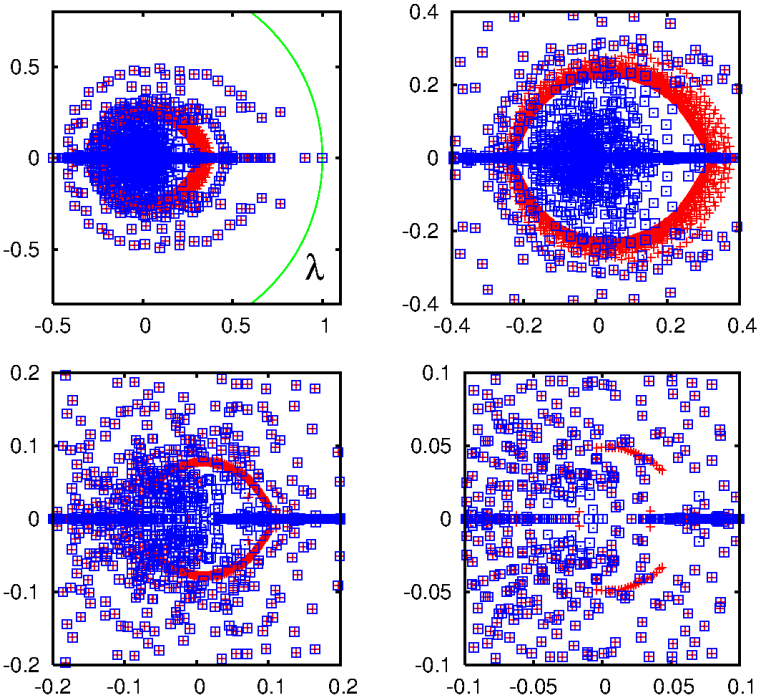
1. K. M. Frahm, A. D. Chepelianskii and D. L. Shepelyansky, ***PageRank of integers***, Phys. A: Math. Theor. **45**, 405101 (2012).

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

3. 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).

Appendix: Rational interpolation method

High precision Arnoldi method for full Physical Review network (including the "future citations") for 52, 256, 512, 768 binary digits and $n_A = 2000$:



Semi-analytical argument for the full PR network:

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

There are ***two groups of eigenvectors*** ψ with: $S\psi = \lambda\psi$

1. Those with $d^T\psi = 0 \quad \Rightarrow \quad \psi$ is also an eigenvector of S_0. Generically an arbitrary eigenvector of S_0 is **not** an eigenvector of S **unless** the eigenvalue is degenerate with degeneracy $m > 1$. Using linear combinations of different eigenvectors for the same eigenvalue one can construct $m - 1$ eigenvectors ψ respecting $d^T\psi = 0$ which are therefore eigenvectors of S.

   Pratically: determine degenerate subspace eigenvalues of S_0 (and also of S_0^T) which are of the form: $\lambda = \pm 1/\sqrt{n}$ with $n = 1,\, 2,\, 3,\, \ldots$ due to 2×2-blocks:

$$\begin{pmatrix} 0 & 1/n_1 \\ 1/n_2 & 0 \end{pmatrix} \quad \Rightarrow \quad \lambda = \pm \frac{1}{\sqrt{n_1\, n_2}}\,.$$

2. Those with $d^T\psi \neq 0 \quad \Rightarrow \quad \mathcal{R}(\lambda) = 0$ with the rational function:

$$\boxed{\mathcal{R}(\lambda) = 1 - d^T \frac{\mathbf{1}}{\lambda\,\mathbf{1} - S_0} e/N = 1 - \sum_{j,q} \frac{C_{jq}}{(\lambda - \rho_j)^q}}$$

Here C_{jq} and ρ_j are unknown, except for
$\rho_1 = 2\,\mathrm{Re}\,[(9 + i\sqrt{119})^{1/3}]/(135)^{1/3} \approx 0.9024$ and
$\rho_{2,3} = \pm 1/\sqrt{2} \approx \pm 0.7071$.

Idea: Expand the geometric matrix series $\Rightarrow$

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

which converges for $|\lambda| > \rho_1 \approx 0.9024$ since $c_j \sim \rho_1^j$ for $j \to \infty$.

Problem: How to determine the zeros of $\mathcal{R}(\lambda)$ with $|\lambda| < \rho_1$?

Analytic continuation by rational interpolation:

Use the series to evaluate $\mathcal{R}(z)$ at n_S support points $z_j = \exp(2\pi ij/n_S)$ with a given precision of p binary digits and determine the rational function $R_I(z)$ which interpolates $\mathcal{R}(z)$ at these support points. Two cases:

$$n_S = 2n_R + 1 \quad \Rightarrow \quad R_I(z) = \frac{P_{n_R}(z)}{Q_{n_R}(z)}$$

$$n_S = 2n_R + 2 \quad \Rightarrow \quad R_I(z) = \frac{P_{n_R}(z)}{Q_{n_R+1}(z)}$$

The n_R zeros of $P_{n_R}(z)$ are approximations of the eigenvalues of S (of the 2nd group).

For a given precision, e. g. $p = 1024$ binary digits one can obtain a certain number of reliable eigenvalues, e. g. $n_R = 300$. The method can be pushed up to $p = 16384$ and $n_R = 2500$ which is better than the high precision Arnoldi method with $n_A = 2000$.

Examples:

Some “artificial zeros” for $n_R = 340$ and $p = 1024$ (*left top and middle panels*) where both variants of the method differ.

For $n_R = 300$ and $p = 1024$ most zeros coincide with HP Arnoldi method (*right top and middle panels*) and both variants of the method coincide.

Lower panels: comparison for $n_R = 2000$, $p = 12288$ (left) or for $n_R = 2500$, $p = 16384$ with HP Arnoldi method.