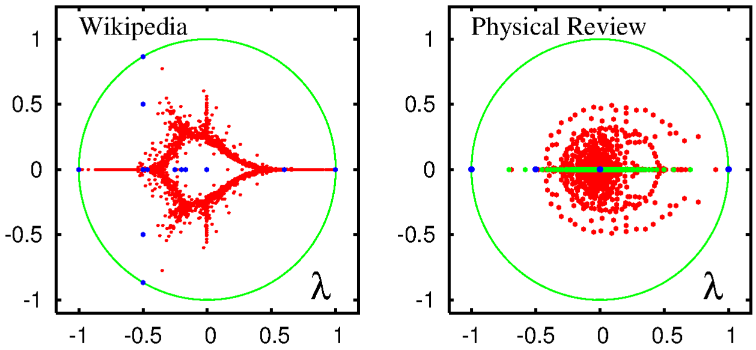


Spectrum and eigenstates of Google matrix

Klaus Frahm

Quantware MIPS Center

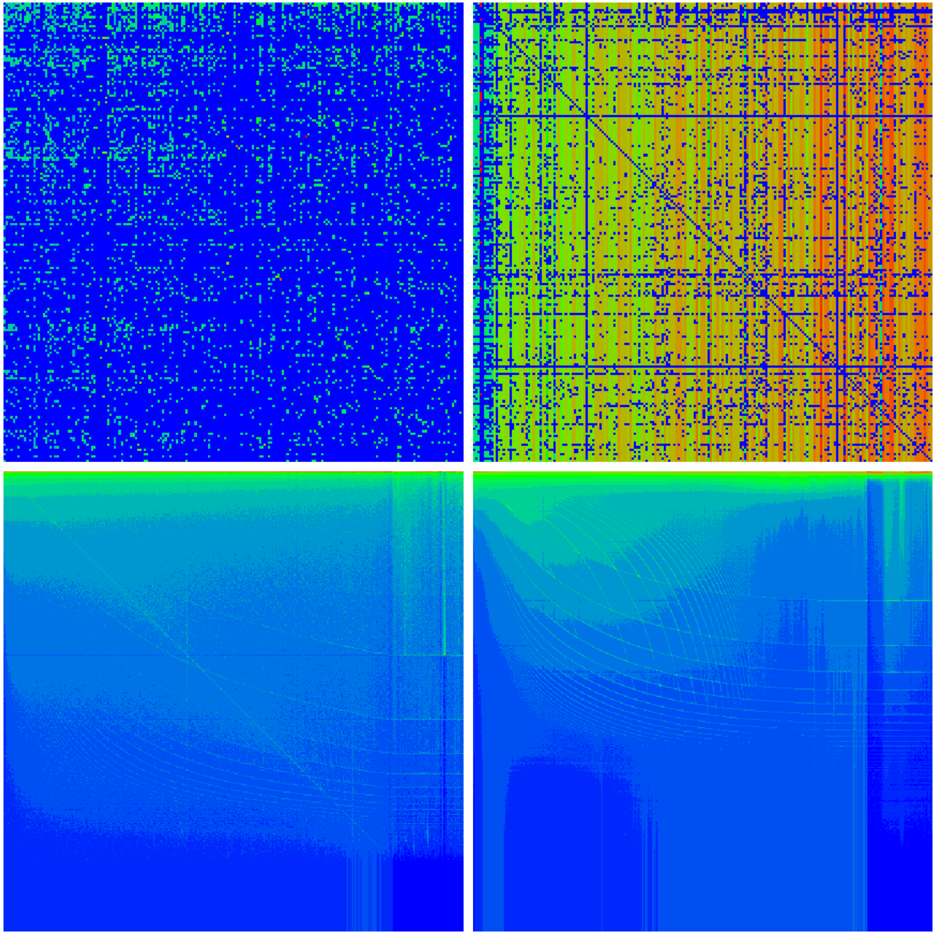
Université Paul Sabatier

Laboratoire de Physique Théorique, UMR 5152, IRSAMC, CNRS

supported by EC FET Open project NADINE

NADINE project REVIEW 2013, Toulouse, 14 November 2013

Google matrix structure



Wikipedia 2009 Twitter 2009

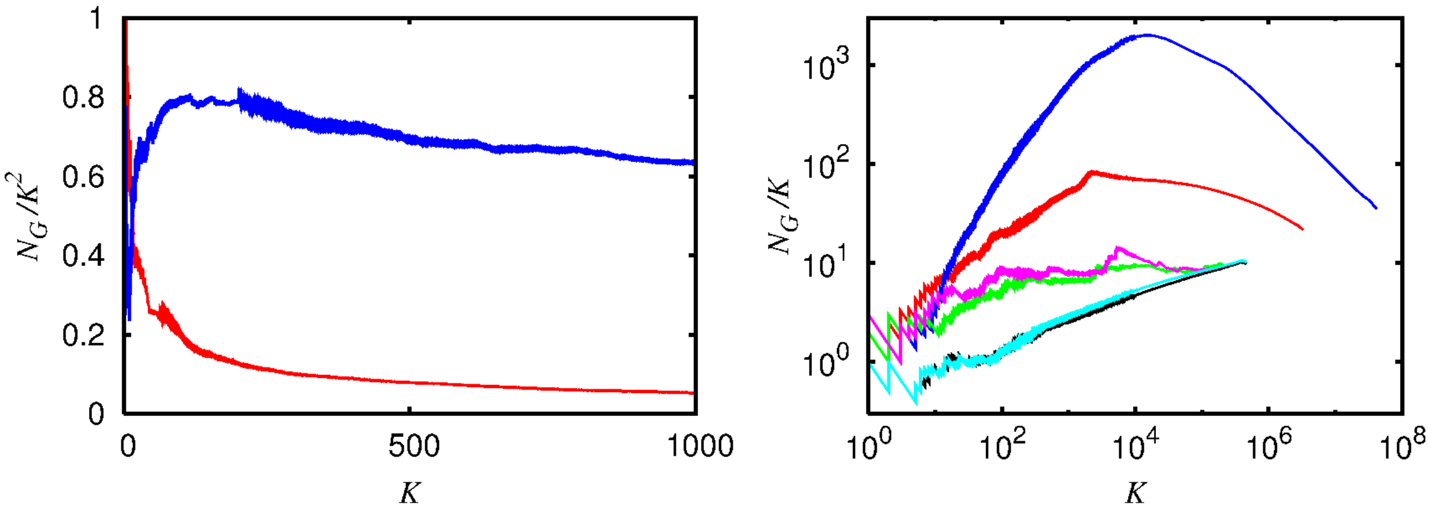
top 200×200

(in PageRank order)

coarse-grained
500×500

(in PageRank order)

Density of non-zero elements N_G of the adjacency matrix among top PageRank nodes:



Blue:	Twitter 2010
Red:	Wikipedia 2009
Magenta:	Oxford 2006
Green:	Cambridge 2006
Cyan:	Physical Review 2009, all journals
Black:	Physical Review 2009, without Rev. Mod. Phys.

Diagonalization of Google matrices

Arnoldi method to (partly) diagonalize large sparse matrices:

- choose an initial normalized vector ξ_0 (random or "otherwise")
- determine the ***Krylov space*** of dimension n_A (typically: $1 \ll n_A \ll d$) spanned by the vectors: $\xi_0,\ G\,\xi_0,\ \ldots,\ G^{n_A-1}\xi_0$
- determine by ***Gram-Schmidt*** orthogonalization an orthonormal basis $\{\xi_0,\ \ldots,\ \xi_{n_A-1}\}$ and the representation of G in this basis:

$$G\,\xi_k = \sum_{j=0}^{k+1} H_{jk}\,\xi_j$$

- diagonalize the ***Arnoldi matrix*** H which has ***Hessenberg*** form:

$$H = \left(\begin{array}{ccccc} * & * & \cdots & * & * \\ * & * & \cdots & * & * \\ 0 & * & \cdots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & * & * \\ \hline 0 & 0 & \cdots & 0 & * \end{array}\right)$$

which provides the ***Ritz eigenvalues*** that are very good aproximations to the "largest" eigenvalues of A.

Invariant subspaces

Problem: (possibly) large degeneracy of $\lambda_1 = 1$.

$\Rightarrow$ Determine the ***invariant subspaces*** defined as subsets of nodes such that for any node in a subspace each outgoing link stays in the subspace.

$\Rightarrow$ Decomposition of the network in many ***separate subspaces*** with N_s nodes and a "big" ***core space*** of the remaining $N - N_s$ nodes.

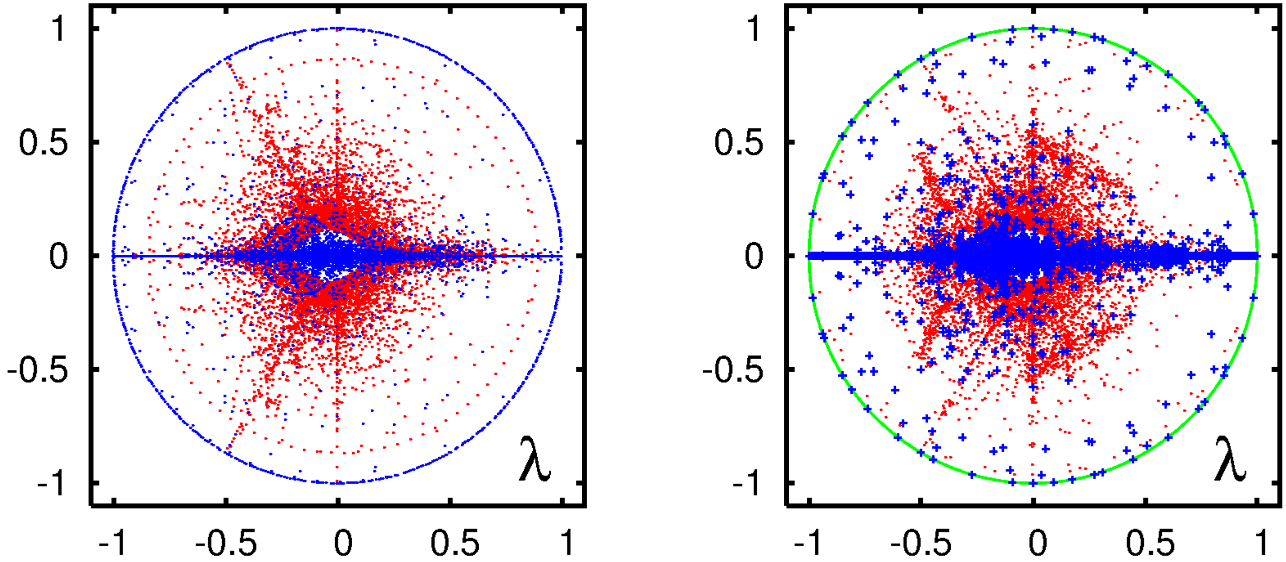
$\Rightarrow$ block structure:

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix} \quad , \quad S_{ss} = \begin{pmatrix} S_1 & 0 & \dots \\ 0 & S_2 & \\ \vdots & & \ddots \end{pmatrix}$$

- Exact (or Arnoldi) diagonalization for each subspace with at least one unit eigenvalue per subspace ($\Rightarrow$ degeneracy).
- Arnoldi method for S_{cc} to determine the largest core space eigenvalues λ_j (note: $|\lambda_j| < 1$).

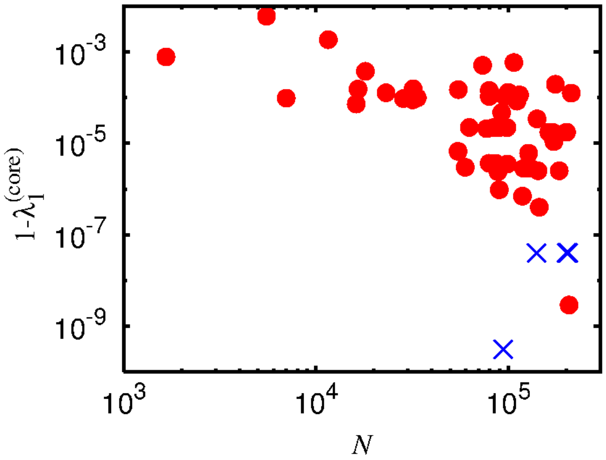
University networks

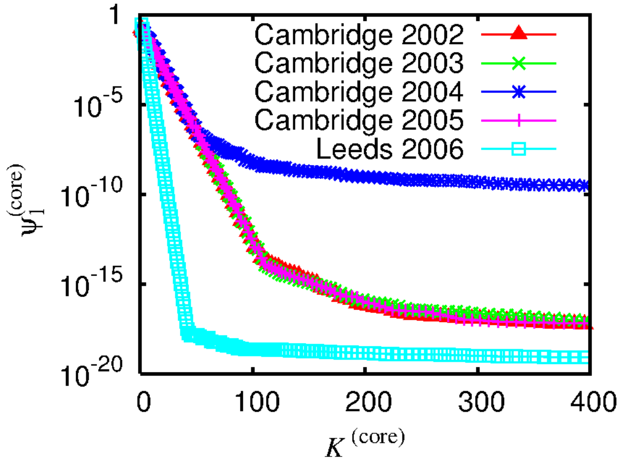
KMF, B. Georgeot and D.L. Shepelyansky, J. Phys. A: Math. Theor. **44**, 465101 (2011)



Cambridge 2006 (Oxford 2006), N = 212710 (200823), N_ℓ = 2015265 (1831542), N_s = 48239 (30579), Number of subspaces = 1543 (1889), n_A = 20000, max. dim. = 4656 (1545), degeneracy of $\lambda_1 = 1$: 1832 (2360).

Core space gap





(Blue crosses shifted up by 10^9)

	$1-\lambda_1^{(\mathrm{core})}$
Cambridge 2002	$3.996 \cdot 10^{-17}$
Cambridge 2003	$4.01 \cdot 10^{-17}$
Cambridge 2004	$2.91 \cdot 10^{-9}$
Cambridge 2005	$4.01 \cdot 10^{-17}$
Leeds 2006	$3.126 \cdot 10^{-19}$

Small gap:

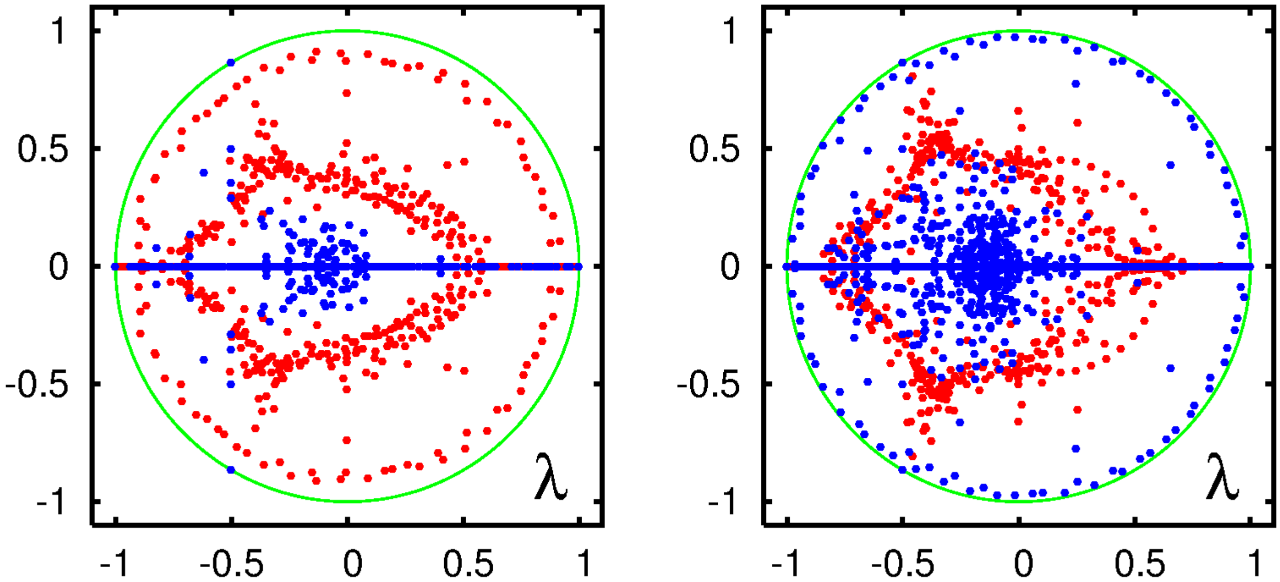
$\Rightarrow$ exponential localization of eigenvectors

$\Rightarrow$ quasi-subspace

Spectrum of Twitter

KMF and D.L. Shepelyansky, Eur. Phys. J. B **85**, 355 (2012)

Twitter 2010 : $N = 41652230$ nodes, $N_\ell = 1468365182$ network links.



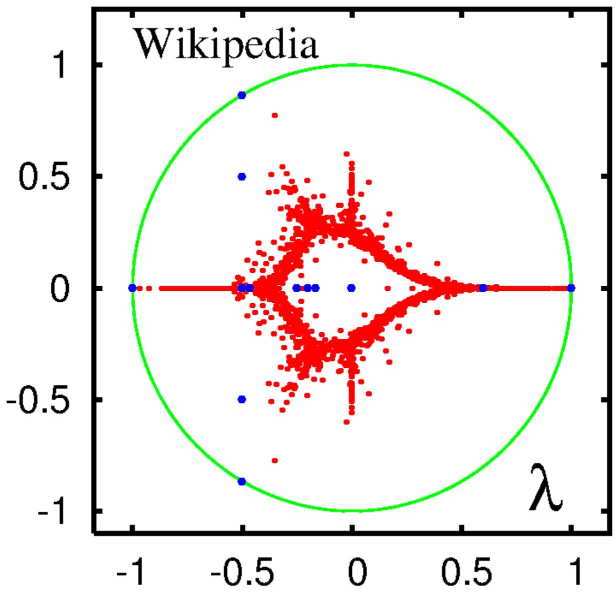
spectrum of S, $N_s = 40307$ spectrum of S^*, $N_s = 180414$

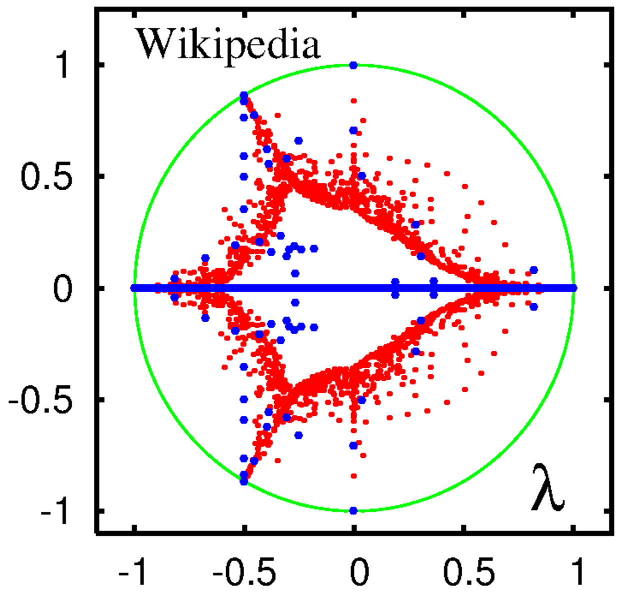
$n_A = 640$ for both cases

Spectrum of Wikipedia

L. Ermann, KMF and D.L. Shepelyansky, Eur. Phys. J. B **86**, 193 (2013)

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



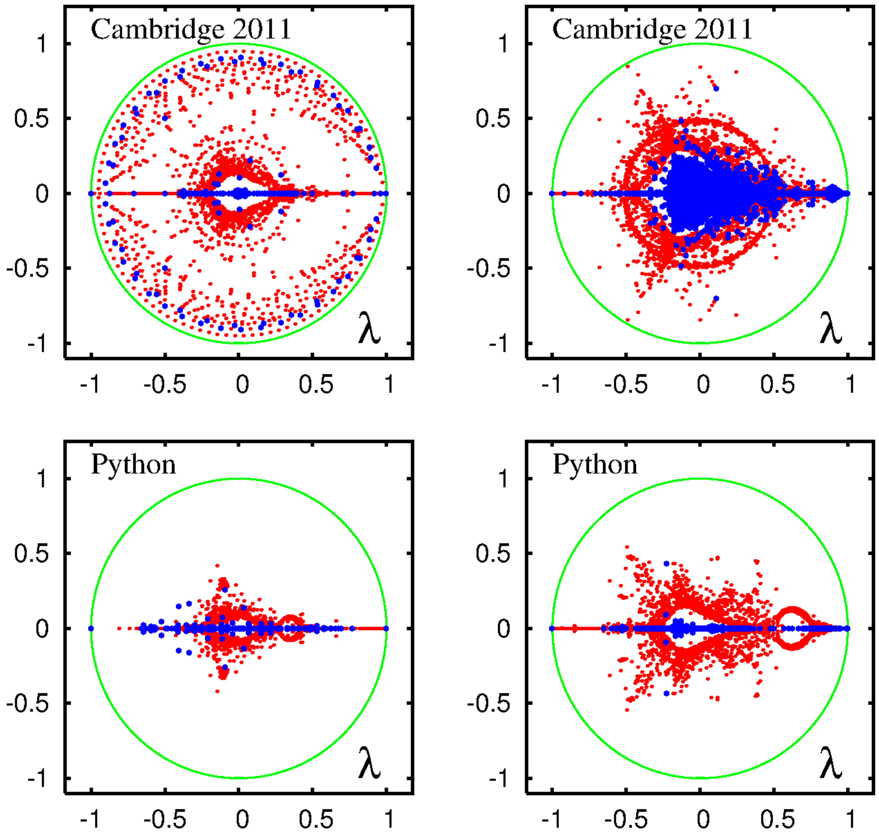


spectrum of S, $N_s = 515$

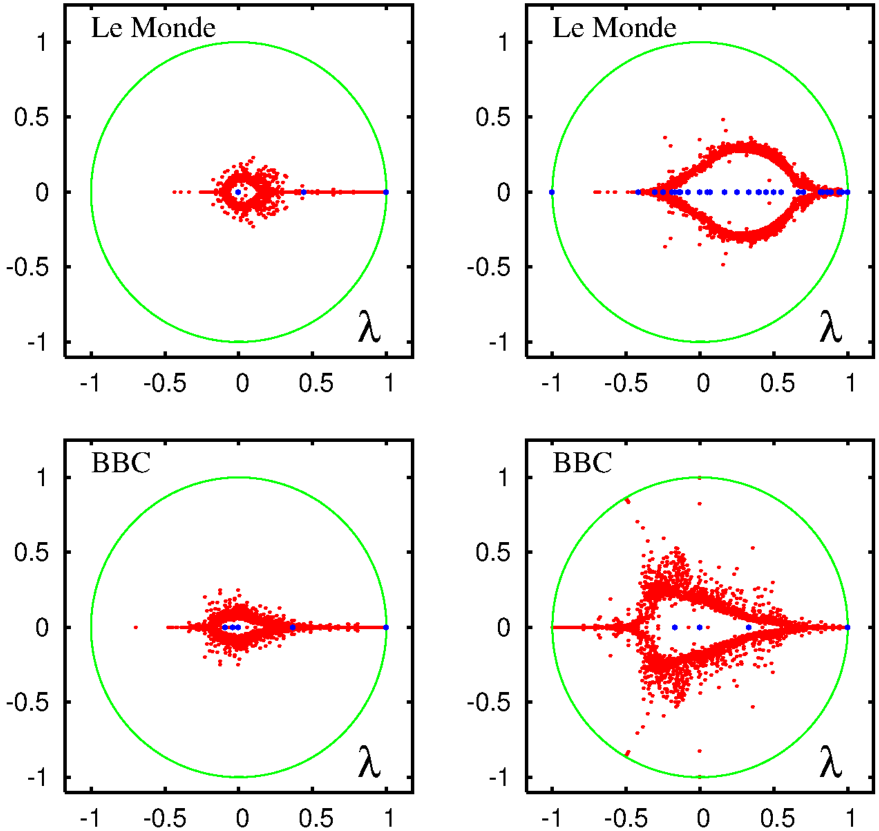
spectrum of S^*, $N_s = 21198$

$n_A = 6000$ for both cases

Spectra of other networks

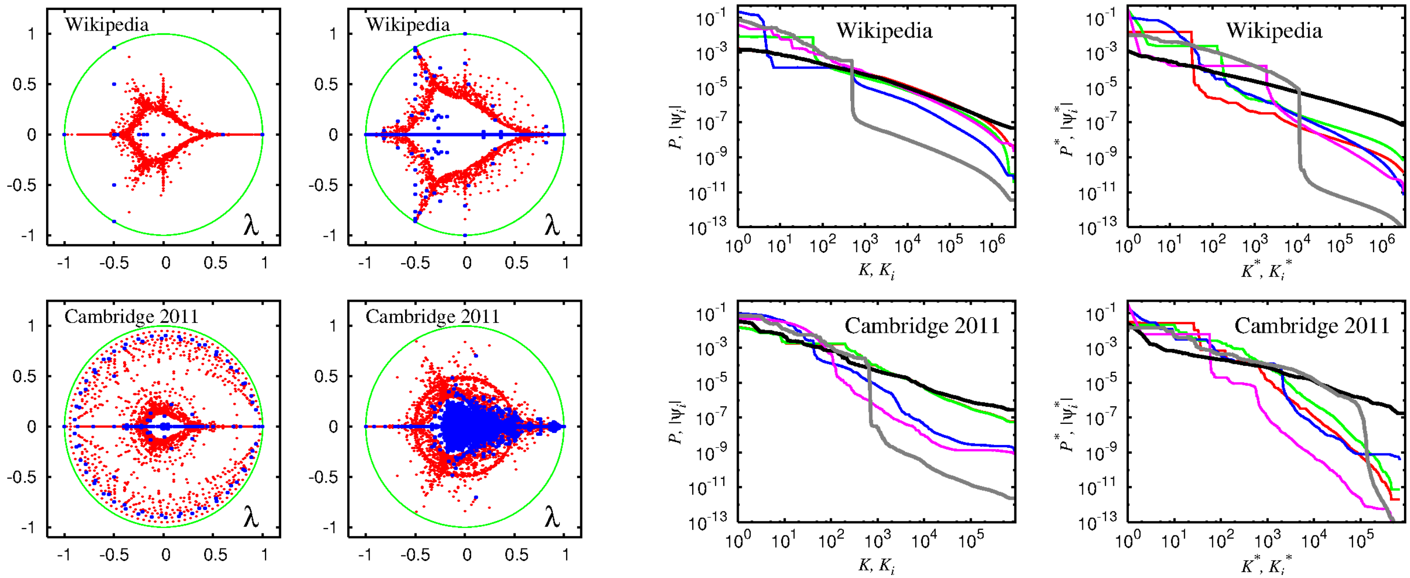


spectrum of S spectrum of S^*



spectrum of S spectrum of S^*

Some Eigenvectors:



left (right): PageRank (CheiRank)

$$\boxed{\xi_{\mathrm{IPR}} \sim 10 - 100 \ll N}$$

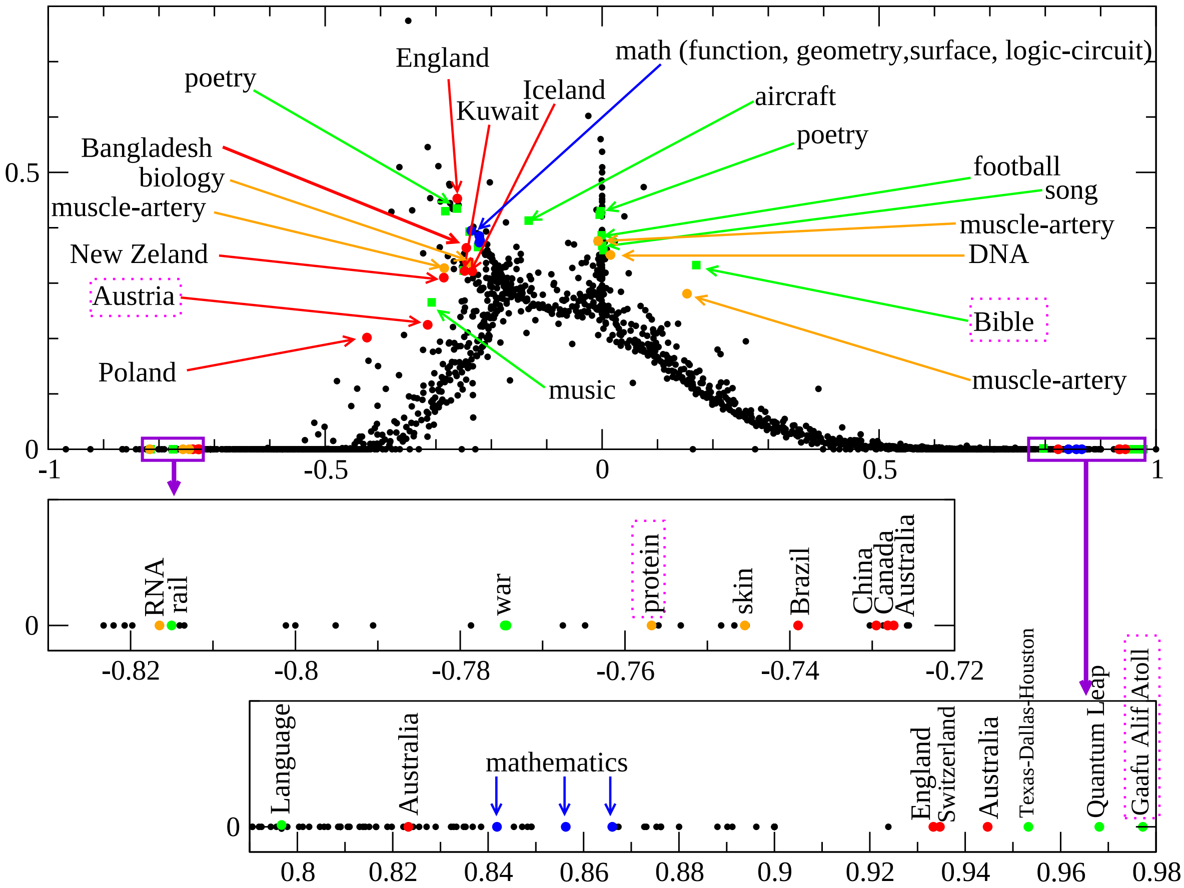
black: PageRank (CheiRank) at $\alpha = 0.85$

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 eigenvectors (Wikipedia 2009):



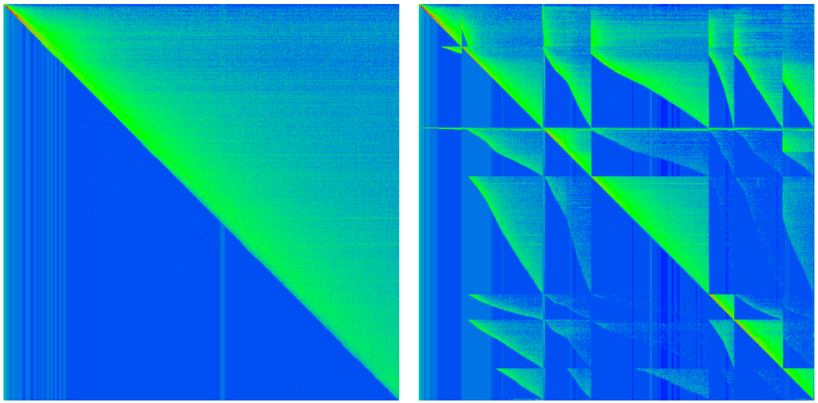
	$\lambda_{1481} = 0.1699 + i0.3325$ ("Bible")	$\lvert\psi_i\rvert$
1	Portal:Bible	0.02311
2	Portal:Bible/Featured chapter/archives	0.02201
3	Portal:Bible/Featured article	0.02063
4	Bible	0.01684
5	Portal:Bible/Featured chapter	0.01644
6	Books of Samuel	0.00852
7	Books of Kings	0.00849
8	Books of Chronicles	0.00840
9	Book of Leviticus	0.00426
10	Book of Ezra	0.00425
11	Book of Ruth	0.00420
12	Book of Deuteronomy	0.00417
13	Book of Joshua	0.00400
14	Book of Exodus	0.00397
15	Book of Judges	0.00395
16	Book of Genesis	0.00394
17	Book of Numbers	0.00389
18	Portal:Bible/Featured chapter/1 Kings	0.00347
19	Portal:Bible/Featured chapter/Numbers	0.00347
20	Portal:Bible/Featured chapter/2 Samuel	0.00347

Physical Review network

KMF, Young-Ho Eom, D. Shepelyansky, arXiv:1310.5624

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

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



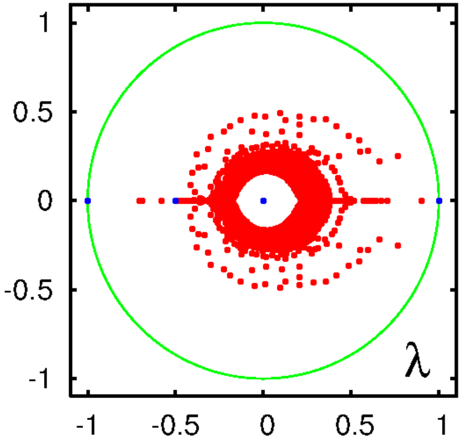
left: time ordered

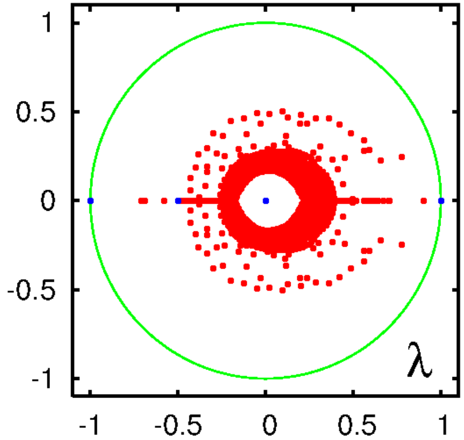
right: journal and then time ordered

“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 adjancy 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!

<u>Reason:</u> 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 $\varepsilon = 10^{-16} \quad \Rightarrow \quad$ “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***:

$\boxed{S_0^l = 0}$ 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 $l = 352$:

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

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

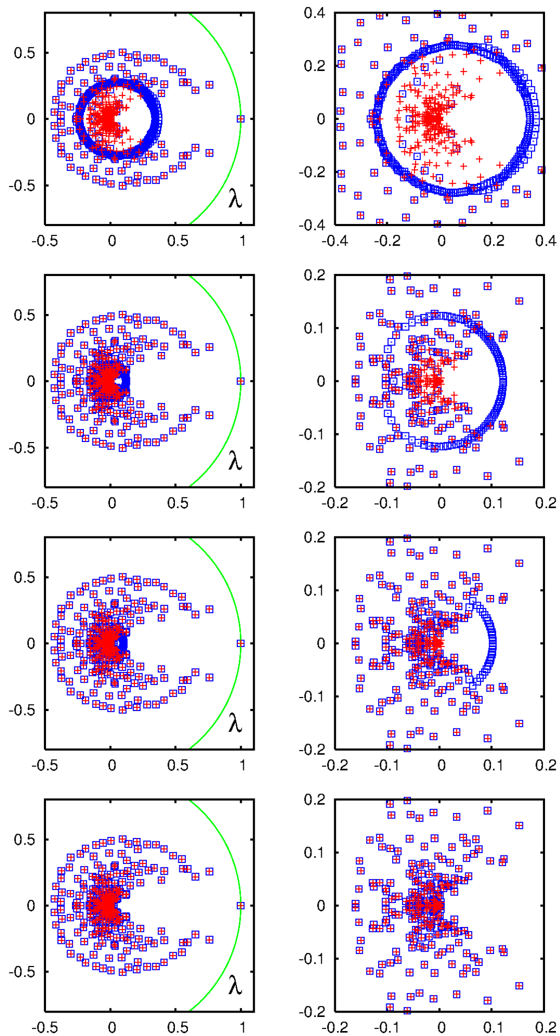
<u>However: still numerical problems:</u>

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

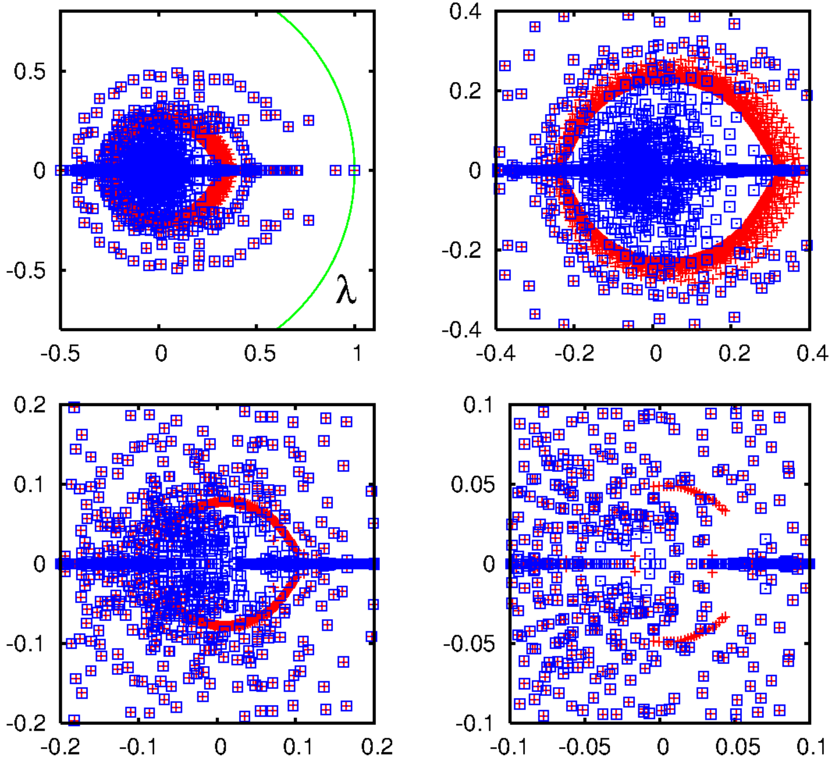


red crosses: zeros of $\mathcal{P}_r(\lambda)$ from 256 binary digits calculation

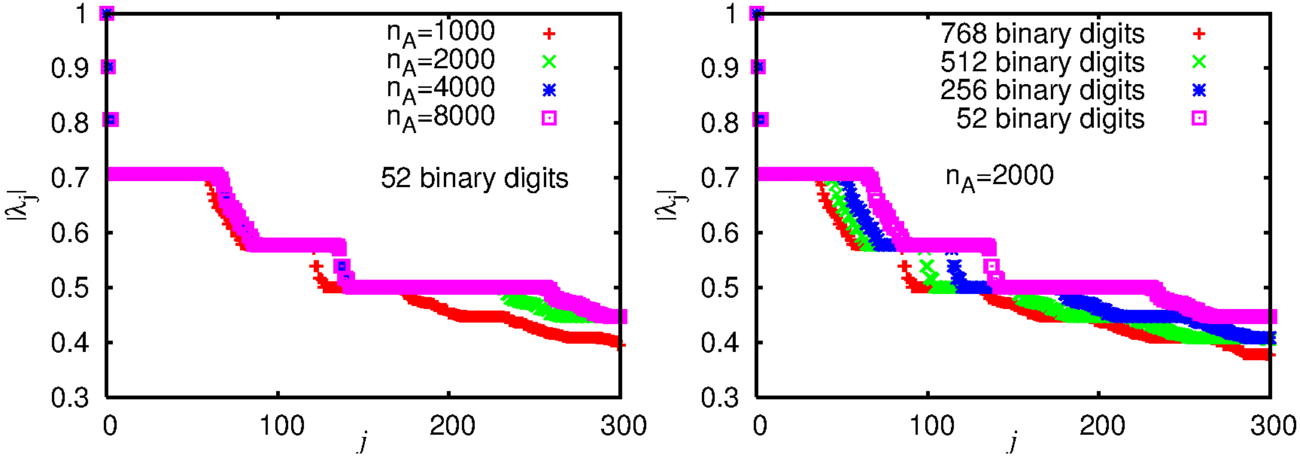
blue squares: eigenvalues from Arnoldi method with 52, 256, 512, 1280 binary digits. In the last case: $\Rightarrow$ break off at $n_A = 352$ with vanishing coupling element.

Full Physical Review network

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



Degeneracies



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

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

Q: How are highly degenerate core space eigenvalues possible ?

Semi-analytical argument for the full PR network:

$\boxed{S = S_0 + e\, d^T/N}$ $\Rightarrow$ ***two groups of eigenvectors*** ψ

1. Those with $d^T\psi = 0 \quad \Rightarrow \quad \psi$ is also an eigenvector of S_0.

   Determine ***degenerate*** subspace eigenvalues of S_0 of the form:
   $\lambda = \pm 1/\sqrt{n}$ with $n = 1,\, 2,\, 3,\, \ldots$

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

$$\boxed{\mathcal{R}(\lambda) = 1 - d^T \frac{\mathbf{1}}{\lambda\,\mathbf{1} - S_0} e/N = 1 - \sum_{j=0}^{\infty} c_j \lambda^{-1-j} \approx \frac{P_{n_R}(\lambda)}{Q_{n_R}(\lambda)}}$$

Determine $\mathcal{R}(\lambda)$ for $2n_R + 1$ values with $|\lambda| = 1$ where the series converges: $\Rightarrow$ ***Rational interpolation method***

The zeros of $P_{n_R}(\lambda)$ are approximations of the eigenvalues of S.

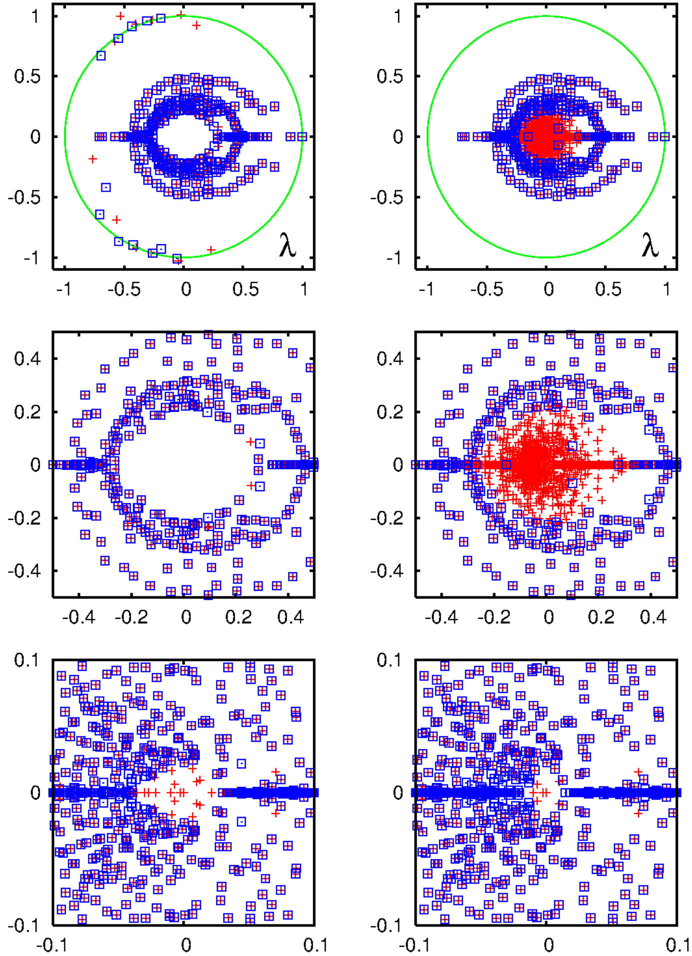
The maximal value of n_R for reliable eigenvalues depends on the precision p of binary digits: e. g. $p = 1024 \ \Rightarrow\ n_R = 300$.

Examples:

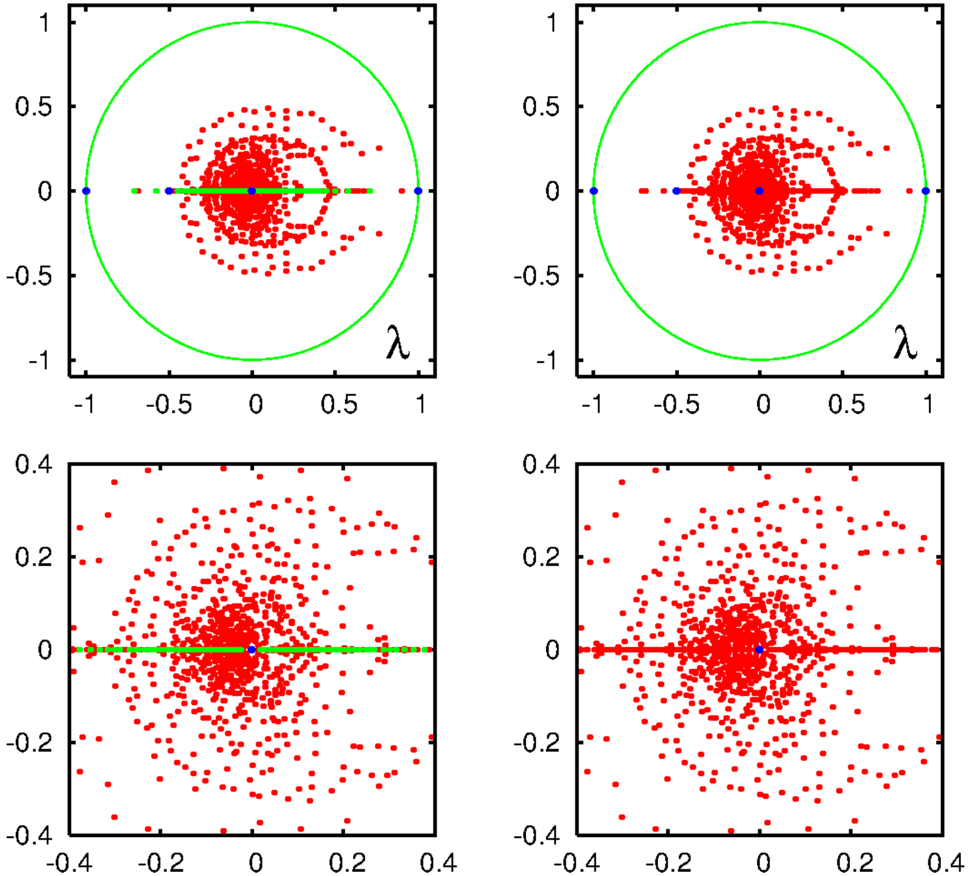
Some “artificial zeros” for $n_R = 340$ and $p = 1024$ (*left top and middle panels*) where both variants of the method differ.

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.



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



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 $0 \le G \le 2/N$
  $\Rightarrow \quad R = 1/\sqrt{3N}$

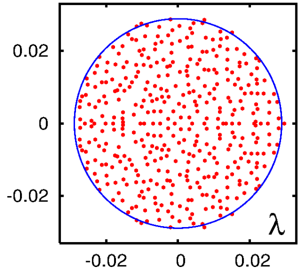
- ***uniform sparse*** with Q non-zero elements per column:
  $P(G) = Q/2$ for $0 \le G \le 2/Q$ with probability Q/N
  and $G = 0$ with probability $1 - Q/N$
  $\Rightarrow \quad R = 2/\sqrt{3Q}$

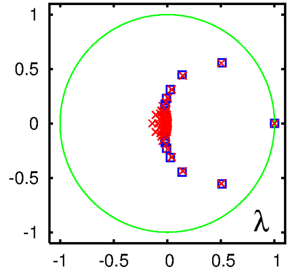
- ***constant sparse*** with Q non-zero elements per column:
  $G = 1/Q$ with probability Q/N
  and $G = 0$ with probability $1 - Q/N$
  $\Rightarrow \quad R = 1/\sqrt{Q}$

- ***powerlaw*** with $p(G) = D(1 + aG)^{-b}$ for $0 \le G \le 1$ 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:

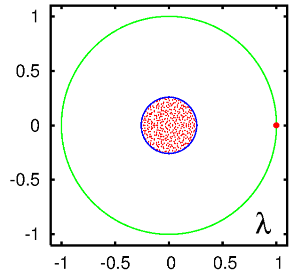
uniform full:
$N = 400$

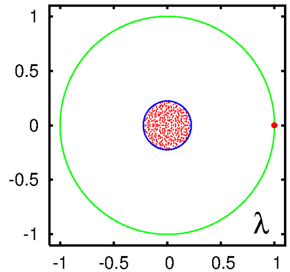




triangular random and average

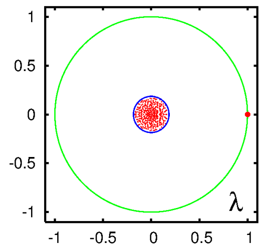
uniform sparse:
$N = 400$,
$Q = 20$

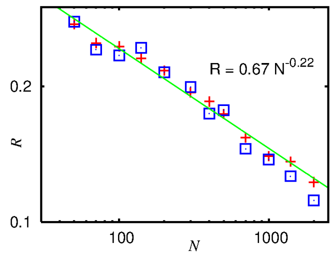




constant sparse:
$N = 400$,
$Q = 20$

power law:
$b = 2.5$





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

Conclusion

- Accurate eigenvalue computation requires determination of ***invariant subspaces***.
- ***Eigenvalue spectra*** for many different network examples.
- Mainly localized eigenvectors for the ***Wikipedia network***: identification of ***themes*** or ***communities***.
- Subtle numerical problems for the eigenvalue problem of the ***Physical Review citation network*** which can be solved by a semi-analytical method and a high precision implementation of the Arnoldi method.
- ***Random Perron-Frobenius matrices*** with nearly uniform circular eigenvalue density: $R \sim 1/\sqrt{Q}$ for Q non-zero elements per column.

- Understanding of the ***degeneracies of core space eigenvalues*** and a decompostion of the core space eigenvalues in two groups. Important role of subspaces of S_0 (very different from the subspaces of S !).
- New ***rational interpolation method*** to determine accurately the eigenvalues of a network matrix. Well suited for nearly triangular matrices but works in principle also for other case (e. g. Wikipedia but less efficient here).
- Drastic effect of the ***triangular approximation*** on the eigenvalue spectrum. Strong reduction of non-vanishing eigenvalues, from about $\sim 8000 - 10000$ to 352 and only very few eigenvalues on the real axis. This implies a very strong effect of the few ***future citations*** on the spectrum.
- Very useful applications of the ***GNU high precision library GMP: http://gmplib.org/*** for different numerical methods: determination of zeros of the reduced polynomial, rational interpolation method, Arnoldi method.

Appendix:

The subspace of $\lambda \neq 0$ is represented by the vectors $v^{(j)} = S_0^{j-1} e/N$ for $j = 1, \ldots, l$

$$\Rightarrow \quad S\, v^{(j)} = c_{j-1}\, v^{(1)} + v^{(j+1)} = \sum_{k=0}^{l-1} \bar{S}_{k,j}\, v^{(k)}$$

“Small” $l \times l$-representation matrix :

$$\bar{S} = \begin{pmatrix} c_0 & c_1 & \cdots & c_{l-2} & c_{l-1} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \quad , \quad \bar{P} = C \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

with $P = \sum_j \bar{P}_j\, v^{(j)} = C \sum_j v^{(j)}$ and due to sum rule: $\sum_j c_j = 1$.