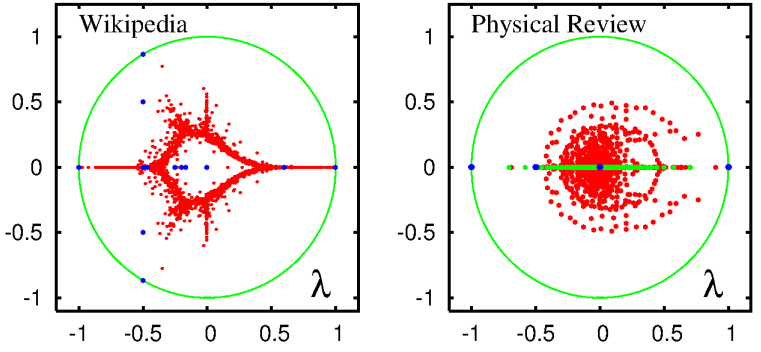


Google matrix analysis of directed networks

Lecture 1

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

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

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***:

$$p_i(t+1) = \sum_j G_{ij}\, p_j(t)$$

The transition probabilities G_{ij} provide a ***Perron-Frobenius*** matrix G such that:

$$\sum_i G_{ij} = 1 \quad , \quad G_{ij} \geq 0\,.$$

Conservation of probability:

$\|G\,v\|_1 = \|v\|_1$ if $v_i \in \mathbb{R}$ and $v_i \geq 0 \;\Rightarrow\; \|p(t+1)\|_1 = \|p(t)\|_1 = 1.$

$\|G\,v\|_1 \leq \|v\|_1$ for any other (complex) vector

where $\|v\|_1 = \sum_i |v_i|$ is the usual 1-norm.

In general $G^T \neq G$ and eigenvalues λ may be complex.

If v is a (right) eigenvector of G: $\quad G\,v = \lambda\,v \quad \Rightarrow \quad |\lambda| \leq 1.$

The vector $e^T = (1,\ \ldots,\ 1)$ is left eigenvector with $\lambda = 1$:

$$e^T\,G = 1\;e^T$$

$\Rightarrow$ existence of (at least) one right eigenvector P for $\lambda = 1$ also called ***PageRank*** in the context of Google matrices:

$$\boxed{G\,P = 1\;P}$$

Biorthogonality between left and right eigenvectors:

$G\,v = \lambda\,v$ and $w^T\,G = \tilde{\lambda}\,w^T \quad \Rightarrow \quad \boxed{w^T\,v = 0 \quad \text{if} \quad \lambda \neq \tilde{\lambda}\;.}$

Expansion in terms of eigenvectors:

$$p(0) = \sum_j C_j \, v^{(j)} \quad \Rightarrow \quad p(t) = \sum_j C_j \, \lambda_j^t \, v^{(j)}$$

with $\lambda_1 = 1$ and $v^{(1)} = P$. If $C_1 \neq 0$ and $|\lambda_j| < 1$ for $j \geq 2$

$$\Rightarrow \qquad \lim_{t \to \infty} p(t) = P \; .$$

$\Rightarrow$ ***Powermethod*** to compute P

Rate of convergence:

$$\sim |\lambda_2|^t = e^{t \, \ln(1-(1-|\lambda_2|))} \approx e^{-t(1-|\lambda_2|)}$$

$\Rightarrow$ Problem if $1 - |\lambda_2| \ll 1$ of even if $|\lambda_2| = 1$.

Complications if G is not diagonalizable

The eigenvectors do not constitute a full basis and further ***generalized eigenvectors*** are required:

$$\begin{aligned}(\lambda_j \mathbf{1} - G)\, v^{(j,0)} &= 0 \\ (\lambda_j \mathbf{1} - G)\, v^{(j,1)} &= v^{(j,0)} \\ (\lambda_j \mathbf{1} - G)\, v^{(j,2)} &= v^{(j,1)} \\ &\vdots\end{aligned}$$

$\Rightarrow$ Contributions $\sim t^l\, \lambda_j^t$ with $l = 0,\, 1,\, \ldots$ in $p(t)$ expansion.

However, for $\lambda_1 = 1$ only $l = 0$ is possible since otherwise:

$$\|p(t)\|_1 \approx \text{const.}\, t^l \quad \to \quad \infty\ .$$

"Analogy" with hamiltonian quantum systems

$$i\hbar \frac{\partial}{\partial t}\,\psi(t) = H\,\psi(t)$$

where $\psi(t)$ quantum state and $H = H^\dagger$ is a hermitian (or real symmetric) operator.

Expansion in terms of eigenvectors: $H\,\varphi^{(j)} = E_j\,\varphi^{(j)}$

$$\boxed{\psi(t) = \sum_j C_j\, e^{-i\,E_j\,t/\hbar}\,\varphi^{(j)}}$$

- H is always diagonalizable with $E_j \in \mathbb{R}$ and $(\varphi^{(k)})^T\,\varphi^{(j)} = \delta_{kj}$.
- Eigenvectors $\varphi^{(j)}$ are valid *physical states* while for PF operators only real vectors with positive entries are physical states and most eigenvectors are complex.

Example hamilontian operators:

- Disorder Anderson model in 1 dimension:

$$H_{jk} = -(\delta_{j,k+1} + \delta_{j,k-1}) + \varepsilon_j \, \delta_{j,k}$$

with random on-site energies $\varepsilon_j \in [-W/2, W/2] \quad \Rightarrow$ localized eigenvectors $\varphi_l \sim e^{-|l-l_0|/\xi}$ with localization length $\xi \sim W^{-2}$. General mesure of localization length by ***inverse participation ratio*** :

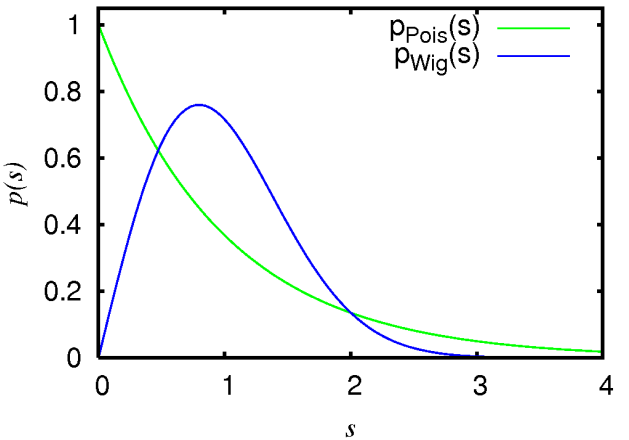
$$\frac{1}{\xi_{\mathrm{IPR}}} = \frac{\sum_l \varphi_l^4}{(\sum_l \varphi_l^2)^2} \sim \frac{1}{\xi}$$

- Gaussian Orthogonal Ensemble (GOE): $H_{jk} = H_{kj} \in \mathbb{R}$ and H_{jk} independent random gaussian variables with:

$$\langle H_{jk} \rangle = 0 \quad , \quad \langle H_{jk}^2 \rangle = (1 + \delta_{jk})\sigma^2.$$

Universal level statistics

Distribution of rescaled nearest level spacing $s = (E_{j+1} - E_j)/\Delta$ with average level spacing Δ:



- ***Poisson statistics:*** $P_{\rm Pois}(s) = \exp(-s)$

  Anderson model with $\xi \ll L$ ($L =$ system size), integrable systems, . . .

- ***Wigner surmise:*** $P_{\rm Wig} = (\pi s/2)\exp(-\pi s^2/4)$

  GOE, Anderson model with $\xi \gtrsim L$, generic (classically) chaotic systems, . . .

PF Operators for directed networks

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

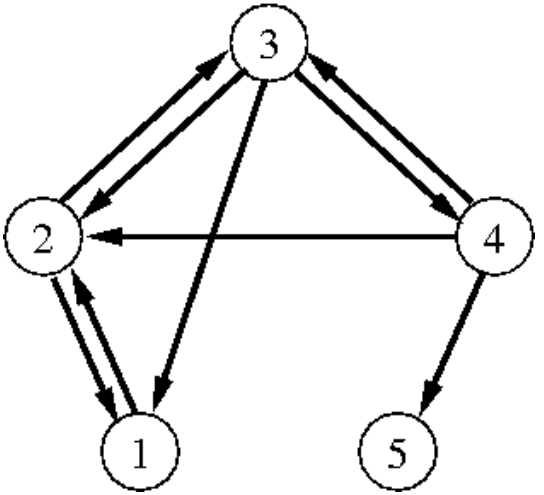
- Define the adjacency matrix by $A_{jk} = 1$ if there is a link $k \to j$ and $A_{jk} = 0$ otherwise. In certain cases, when explicitely considering multiple links, one may have $A_{jk} = m$ where $m =$ multiplicity of a a link (e. g. Network for integer numbers).

- Define a matrix S_0 from A by sum-normalizing each non-zero column to one and keeping zero columns.

- Define a matrix S from S_0 by replacing each zero column with $1/N$ entries.

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

Example:



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

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

The nodes with no out-going links, associated to zero columns in A, are called ***dangling nodes***. On can formally write:

$$S = S_0 + \frac{1}{N}\, e\, d^T$$

with $d =$ dangling vector with $d_j = 1$ for dangling nodes and $d_j = 0$ for other nodes and $e =$ uniform unit vector with $e_j = 1$ for all nodes.

Damping factor

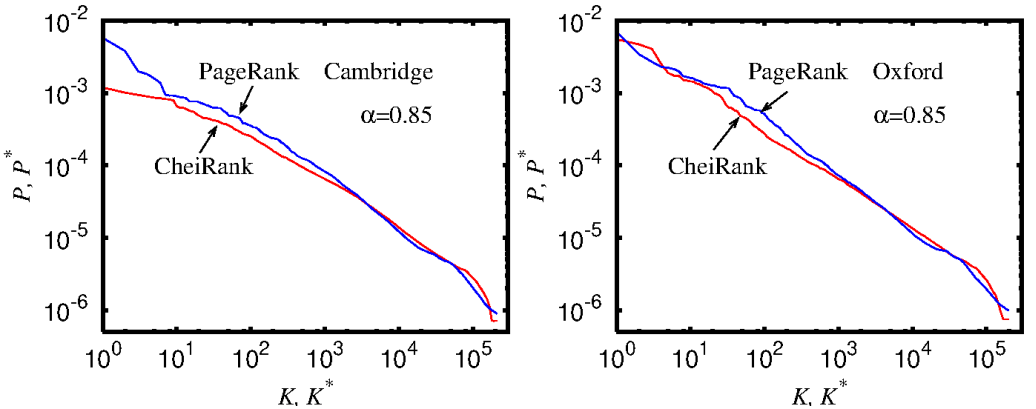
Define for $0 < \alpha < 1$, typically $\alpha = 0.85$, the matrix:

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

- G is also PF operator with columns sum normalized.
- G has the eigenvalue $\lambda_1 = 1$ with multiplicity $m_1 = 1$ and other eigenvalues are $\alpha\lambda_j$ (for $j \geq 2$) with $\lambda_j =$ eigenvalues of S. The right eigenvectors for $\lambda_j \neq 1$ are not modified (since they are orthogonal to the left eigenvector e^T for $\lambda_1 = 1$).
- Similar expression for $G^*(\alpha)$ using S^*.

PageRank

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

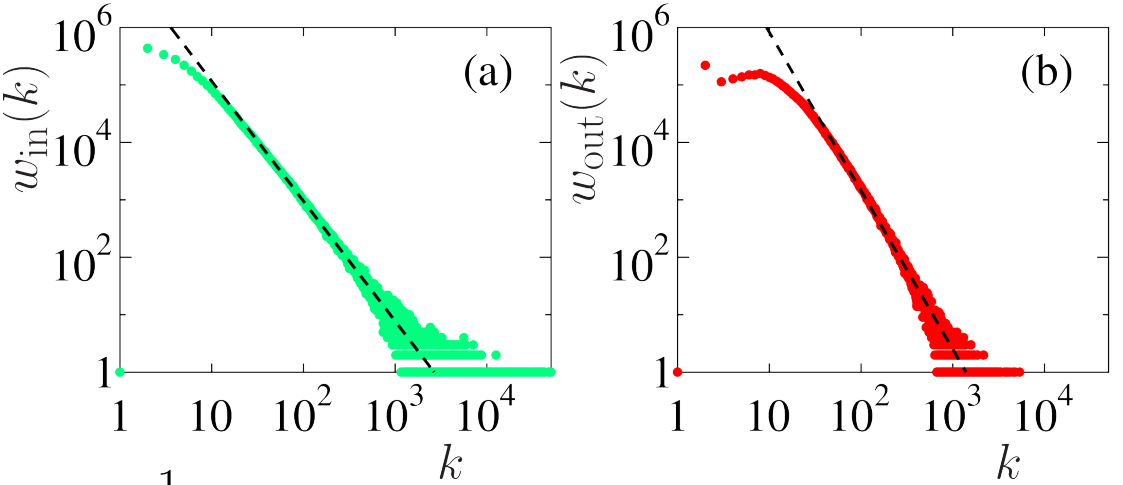


$$P(i) = \sum_j G_{ij}\, P(j)$$

$P(i)$ 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.

Scale Free properties

Distribution of number of in- and outgoing links for Wikipedia:



$$w_{\rm in,out}(k) \sim \frac{1}{k^{\,\mu_{\rm in,out}}} \quad , \quad \mu_{\rm in} = 2.09 \pm 0.04 \quad , \quad \mu_{\rm out} = 2.76 \pm 0.06 \; .$$

(Zhirov et al. EPJ B **77**, 523)

Small world properties: “Six degrees of separation”

(cf. *Milgram’s* ”small world experiment” 1967)

Numerical diagonalization

- Powermethod to obtain P: rate of convergence for $G(\alpha)$ is better than $\sim \alpha^t$.

- Full "exact" diagonalization: possible for $N \lesssim 10^4$:
  memory usage $\sim N^2$ and computation time $\sim N^3$.

- Arnoldi method to determine largest $n_A \sim 10^2 - 10^4$ eigenvalues:
  memory usage $\sim N\, n_A + C_1\, N_\ell + C_2\, n_A^2$ and
  computation time $\sim N\, n_A^2 + C_3\, N_\ell\, n_A + C_4\, n_A^3$.

- Strange numerical problems to determine accurately "small" eigenvalues, in particular for (nearly) triangular network structure due to large Jordan-blocks ($\Rightarrow$ $3^{\rm rd}$ *lecture*).

Arnoldi method

to (partly) diagonalize large sparse non-symmetric $N \times N$ matrices G such that the product "$G\times$vector" can be computed efficiently (G may contain some constant columns $\sim e$):

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

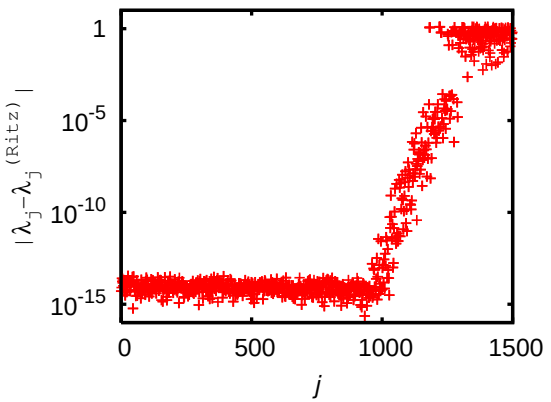
$$G\,\xi_k = \sum_{j=0}^{k+1} H_{jk}\,\xi_j$$

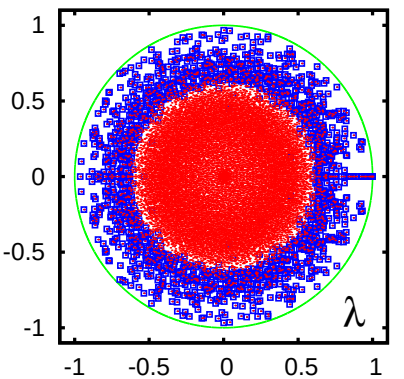
Note: if $G = G^T \ \Rightarrow\ H =$ tridiagonal symmetric and the ***Arnoldi method*** is identical to the ***Lanczos method***.

- 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 G.





Example: PF Operator for Ulam-Map ($\Rightarrow$ 2^{nd} *lecture*)

$N = 16609,\ N_\ell = 76058,\ n_A = 1500$

Invariant subspaces

In realistic WWW networks invariant subspaces of nodes create large degeneracies of λ_1 (or λ_2 if $\alpha < 1$) which is very problematic for the Arnoldi method.

Therefore determine the ***invariant subspaces*** as follows:

Let $N_c = bN$ a certain fraction of the network size N (e.g. $b = 0.1$).

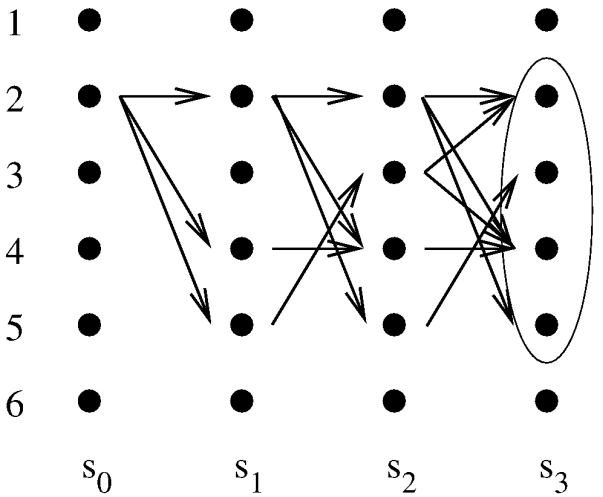
- For a given initial node i_0 determine a sequence of node sets s_n by $s_0 = \{i_0\}$ and s_{n+1} is the set containing all nodes of s_n and those which can be reached by a link from a node in s_n.

- If $s_n = s_{n+1}$ with at most N_c elements for some $n \Rightarrow s_n$ is an ***invariant subspace***.

- If for some n the set s_n contains a dangling node (connected by construction to any other node) or if s_n contains more than N_c elements $\Rightarrow i_0$ is identified as a node belonging to the ***core space*** (space of nodes not belonging to an invariant subspace).
- Repeat the procedure for every network node as potential initial node except for those nodes which are already identified as subspace nodes. If for some n the set s_n contains a previously found core space node $\Rightarrow i_0$ also belongs to the core space.
- Merge all subspaces with common members. In this way one obtains a decomposition of the network in many ***separate subspaces*** with N_s nodes and a "big" ***core space***.

This procedure can be efficiently implemented as a computer program. It turns out that for most networks the exact choice of b is not important (e.g. $b = 0.1$ or $b = 0.9$) as long as $b = \mathcal{O}(1)$. Note that a core space node may have a link to an invariant subspace but a subspace node may not have a link to another subspace or the core space.

Example:



$s_0 = \{2\}$

$s_1 = \{2, 4, 5\}$

$s_2 = \{2, 3, 4, 5\} = s_3 =$ invariant subspace

The decomposition in subspaces and a core space implies a block structure of the matrix S:

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

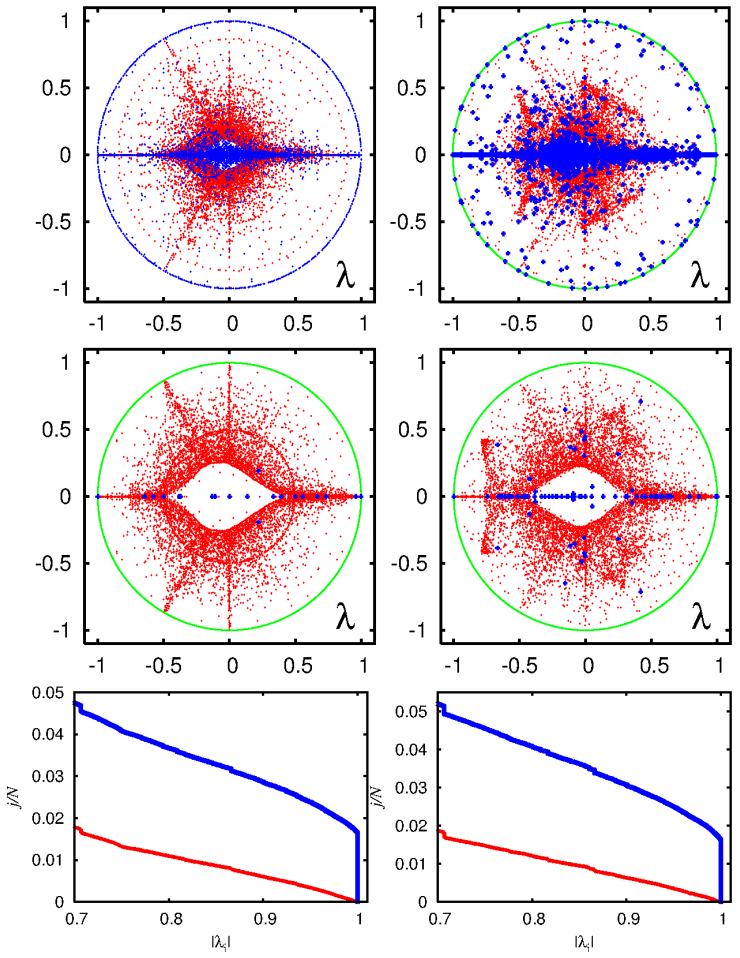
where S_{ss} is block diagonal according to the subspaces. The subspace blocks of S_{ss} are all matrices of PF type with at least one eigenvalue $\lambda_1 = 1$ explaining the high degeneracies.

To determine the spectrum of S apply:

- Exact (or Arnoldi) diagonalization on each subspace.

- The Arnoldi method to S_{cc} to determine the largest core space eigenvalues λ_j (note: $|\lambda_j| < 1$). The largest eigenvalues of S_{cc} are no longer degenerate but other degeneracies are possible (e.g. $\lambda_j = 0.9$ for Wikipedia).

University Networks



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

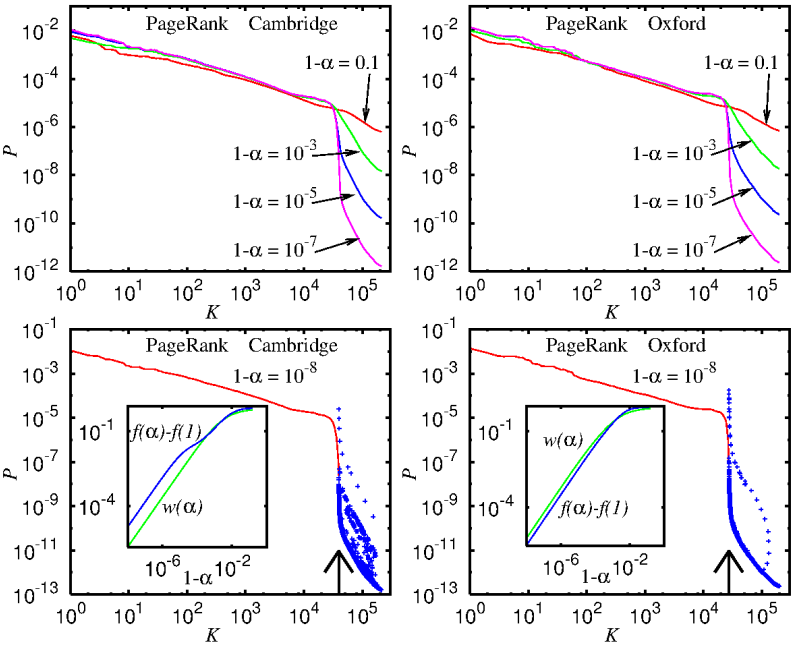
Oxford 2006 (right),
$N = 200823$, $N_s = 30579$

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

Blue: subspace eigenvalues

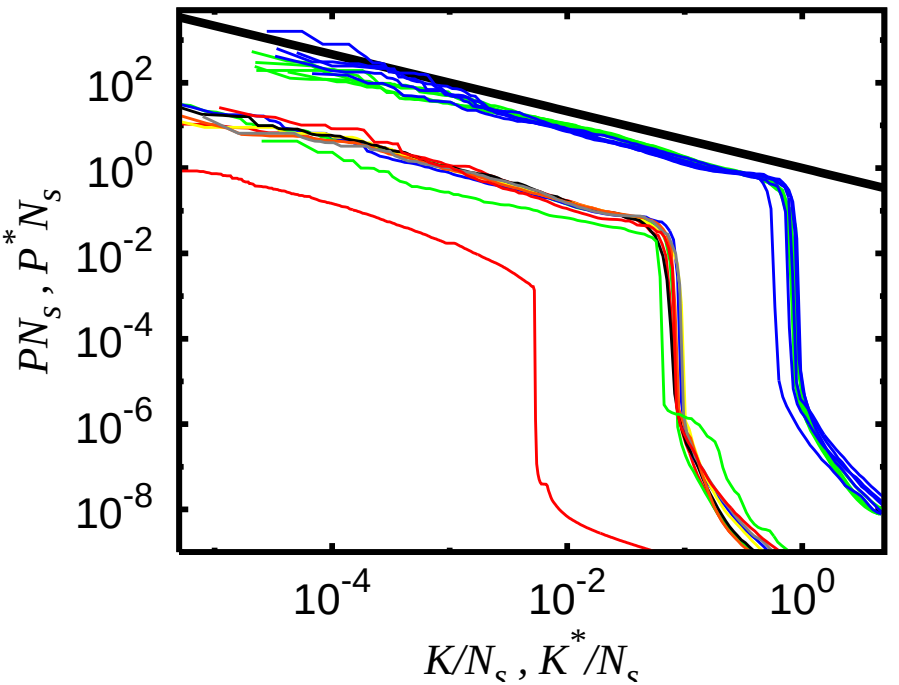
Red: core space eigenvalues (with Arnoldi dimension $n_A = 20000$)

PageRank for $\alpha \to 1$:



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

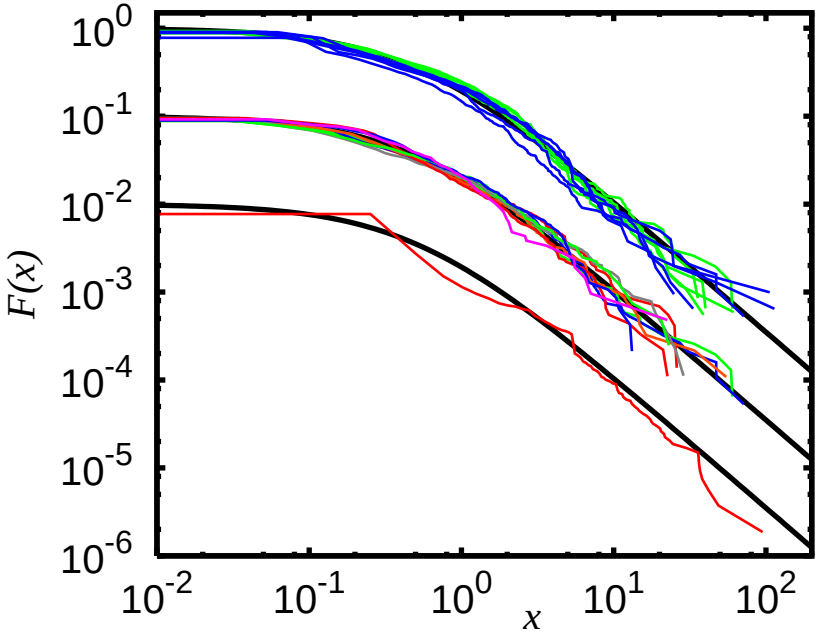
Rescaled PageRank at $\alpha = 1 - 10^{-8}$:



Top: Cambridge, Oxford 2002-2006; middle: other universities; bottom: Wikipedia*; black line $\propto K^{-2/3}$; $N_s =$ sum of all subspace dimensions.

Distribution of dimensions of invariant subspaces

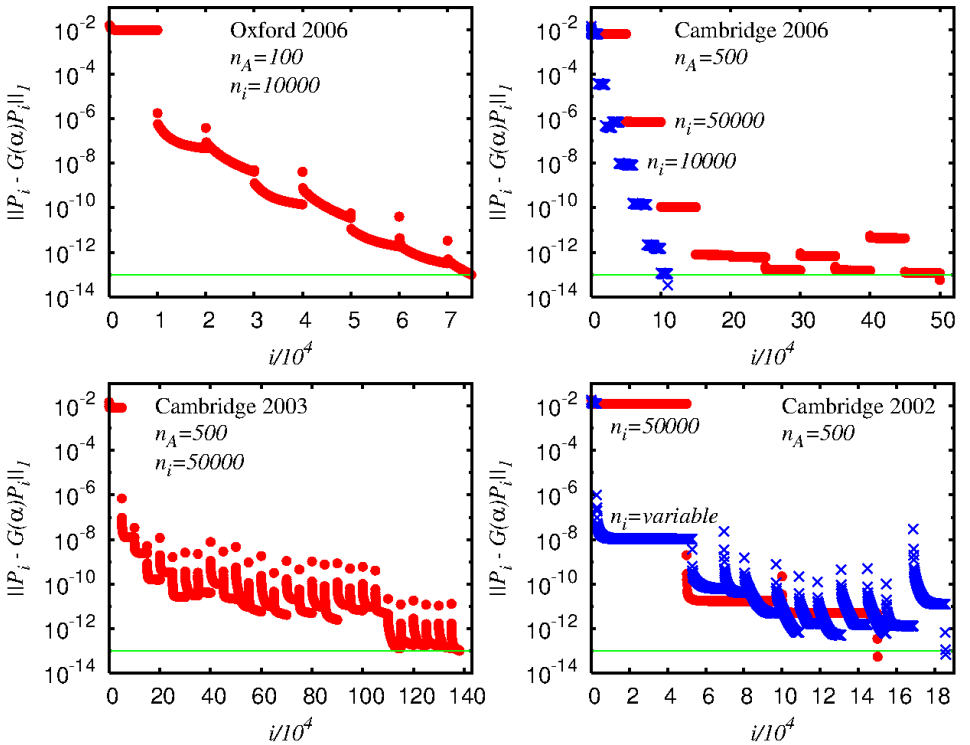
$F(x) =$ fraction of invariant subspaces with dimension larger than $x\langle d\rangle$ where $\langle d\rangle =$ average subspace dimension.



Top: Cambridge, Oxford 2002-2006; middle: other universities; bottom: Wikipedia*; black line: $F(x) = 1/(1+2x)^{3/2}$.

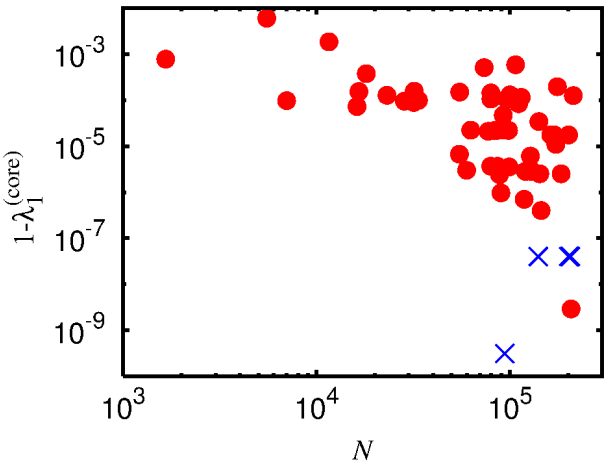
Numerical PageRank method for $\alpha \to 1$

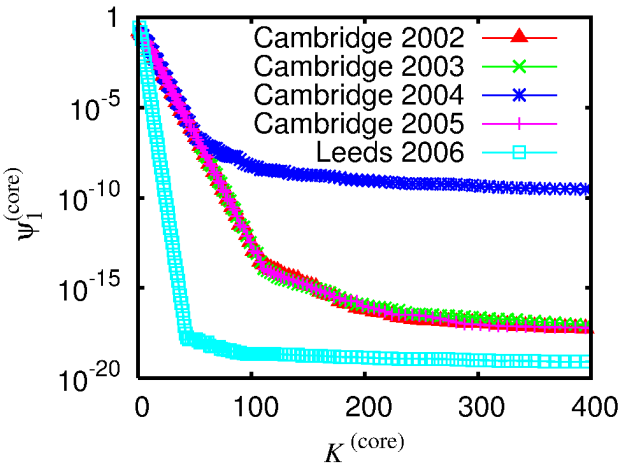
Combination of power method and Arnoldi diagonalization :



Here: $\alpha = 1 - 10^{-8}$

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 $< 10^{-16}$.

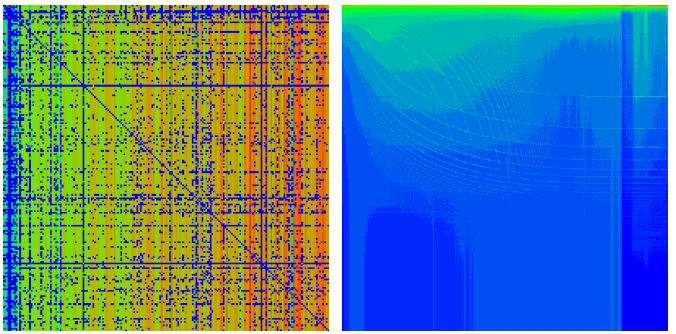
Right: first core space eigenvecteur for universities with gap $< 10^{-16}$ or gap $= 2.91 \times 10^{-9}$ for Cambridge 2004.

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

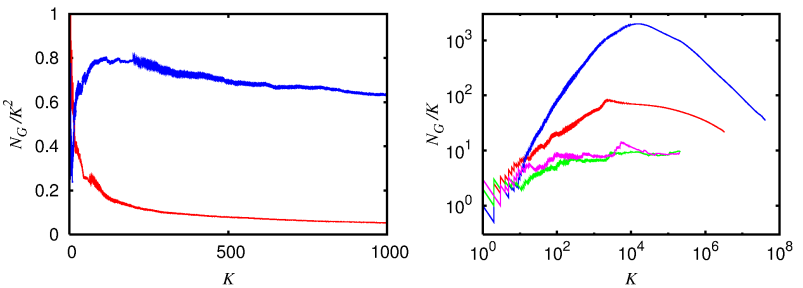
Twitter network

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

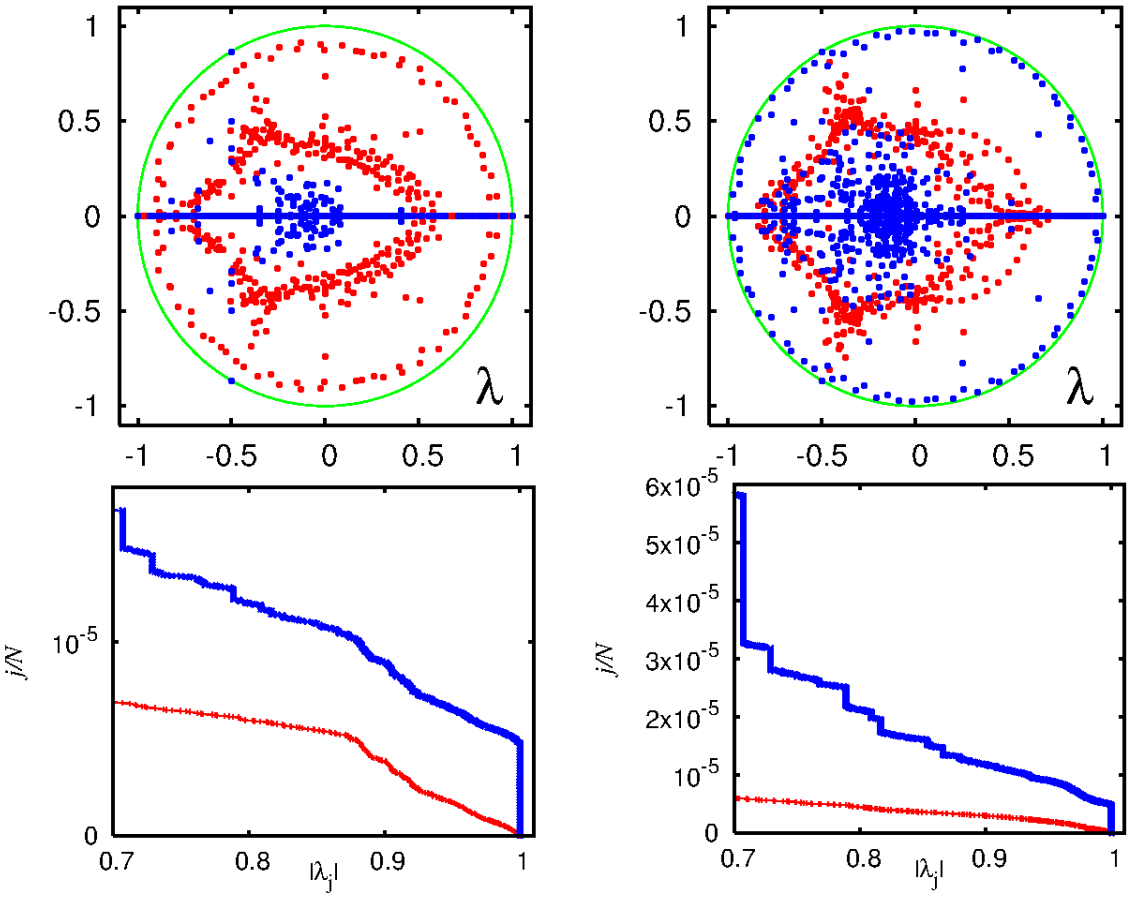
Matrix structure in K-rank order:



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

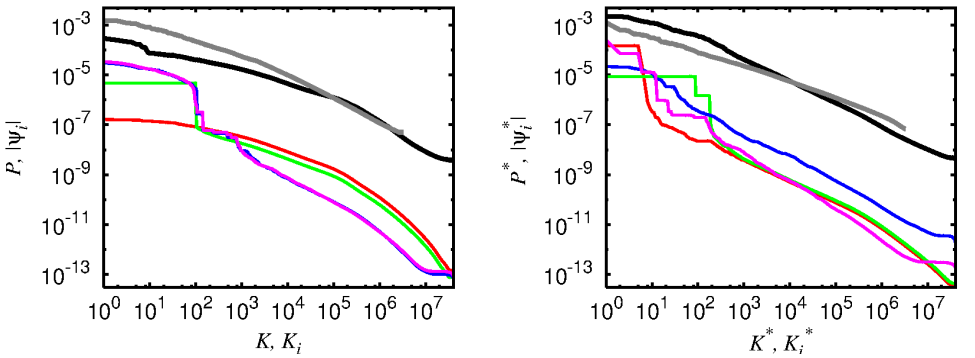


Spectrum

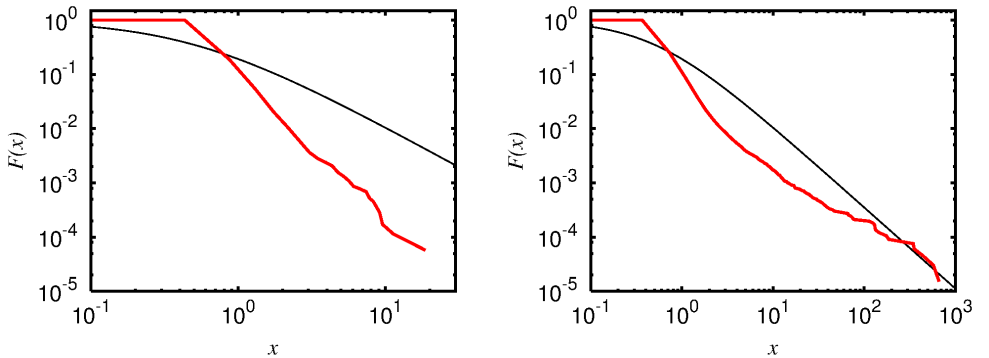


$n_A = 640 \quad \Rightarrow \quad 250$ GB of RAM memory.

PageRank, CheiRank, eigenvectors



Subspace distribution



Black line: $F(x) = 1/(1+2x)^{3/2}$.

References

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

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

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

4. 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).