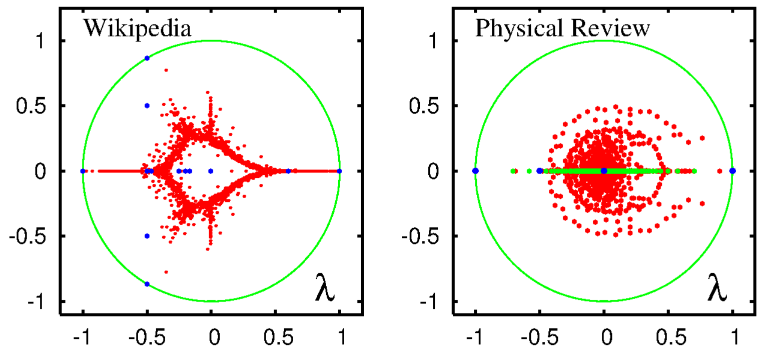


Spectral properties of Google matrix and beyond

Klaus M. Frahm[1]

Quantware MIPS Center Université Paul Sabatier

K. Jaffrès-Runser[2], **D.L. Shepelyansky**[1]

[1] Laboratoire de Physique Théorique du CNRS, IRSAMC

[2] Institut de Recherche en Informatique de Toulouse, INPT

Google Matrix: fundamentals, applications and beyond

IHES, October 15 – 18, 2018

Perron-Frobenius operators

Physical system evolving by a discrete ***Markov process***:

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

Transition probabilities G_{ij} $\Rightarrow$ ***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 complex eigenvalues λ with $|\lambda| \leq 1$.
$e^T = (1, \ldots, 1)$ 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 $\sim |\lambda_2|^t$.

Google matrix

Construct an Adjacency matrix A for a directed network with N nodes and N_ℓ links by :

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

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:

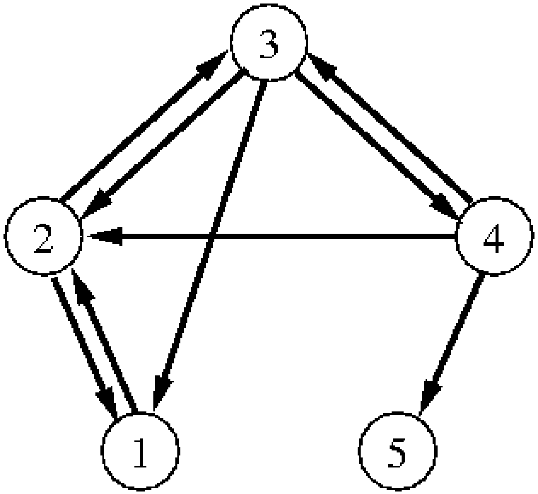
$$\boxed{G(\alpha) = \alpha S + (1-\alpha)\,\frac{1}{N}\,e e^T \quad .}$$

$\Rightarrow \quad \lambda_1$ 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***.

(*Brin and Page, Comp. Networks ISDN Syst.* **30**, *107 (1998).*)

Example:



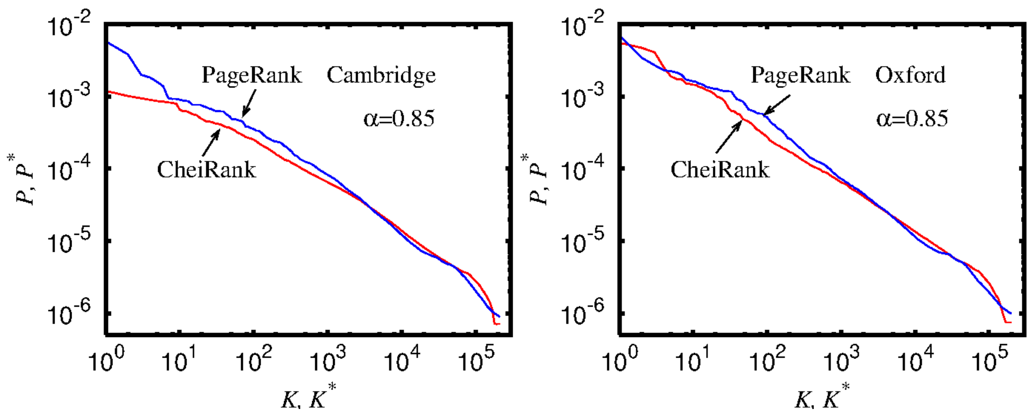
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

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

PageRank

(*KF, Georgeot and Shepelyansky, J. Phys. A* **44** *465101 (2011).*)

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 methods

- ***Power method*** to obtain P or P^*: rate of convergence for $G(\alpha) \sim \alpha^t$.
- ***2d Rank***: representation of nodes (node-density) in $K - K^*-$plane where K (K^*) is sorting index of PageRank P (CheiRank P^*).
- Complex eigenvalues: Full "exact" diagonalization for $N \lesssim 10^4$ or ***Arnoldi method*** to determine largest $n_A \sim 10^2 - 10^4$ eigenvalues.
- ***Invariant subspaces*** in realistic WWW networks $\Rightarrow$ large degeneracies of λ_1:

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

  where S_{ss} is block diagonal according to the subspaces and can be diagonalized separately. S_{cc} corresponds to the core space with $|\lambda_{\max}| < 1$.
- 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 and recently Bitcoin network).

Applications

- University WWW-networks ($N \sim 2 \times 10^5$, $N_\ell \sim 2 \times 10^6$).

- Linux Kernel network ($N \sim 3 \times 10^5$)

  (Nodes = kernel functions)

  (*Ermann, Chepelianskii and Shepelyansky, EPJB* **79**, *115 (2011)*)

- Wikipedia, different language editions ($N \sim 4 \times 10^6$, $N_\ell \sim 10^8$)

- World trade network ($N \sim 10^4$ but more complicated structure).

  (*Ermann and Shepelyansky, Acta Physica Polonica A* **120(6A)**, *A158 (2011)*)

- Twitter 2009 ($N \sim 4 \times 10^7$, $N_\ell \sim 1.5 \times 10^9$).

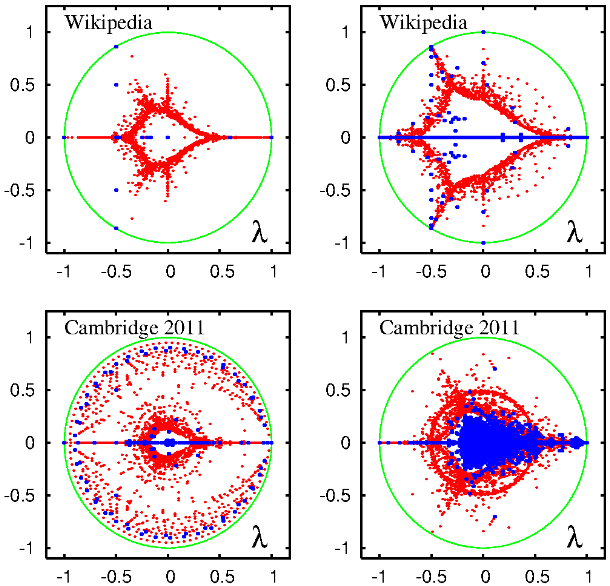
- Physical Review citation network ($N \sim 5 \times 10^5$, $N_\ell \sim 5 \times 10^6$).

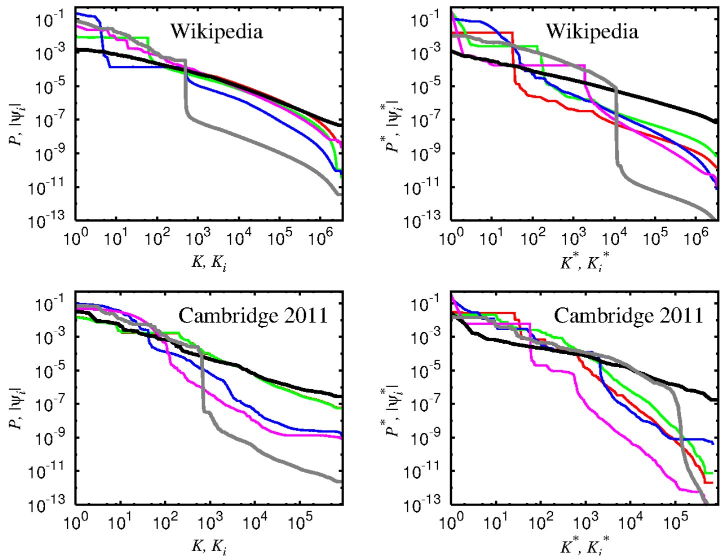
(Review: *Ermann, KF, and Shepelyansky, Rev. Mod. Phys.* **87**, *1261 (2015).*)

Example: Wikipedia 2009

(*Ermann, KF, and Shepelyansky, EPJB* **86**, *193 (2013)*)

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



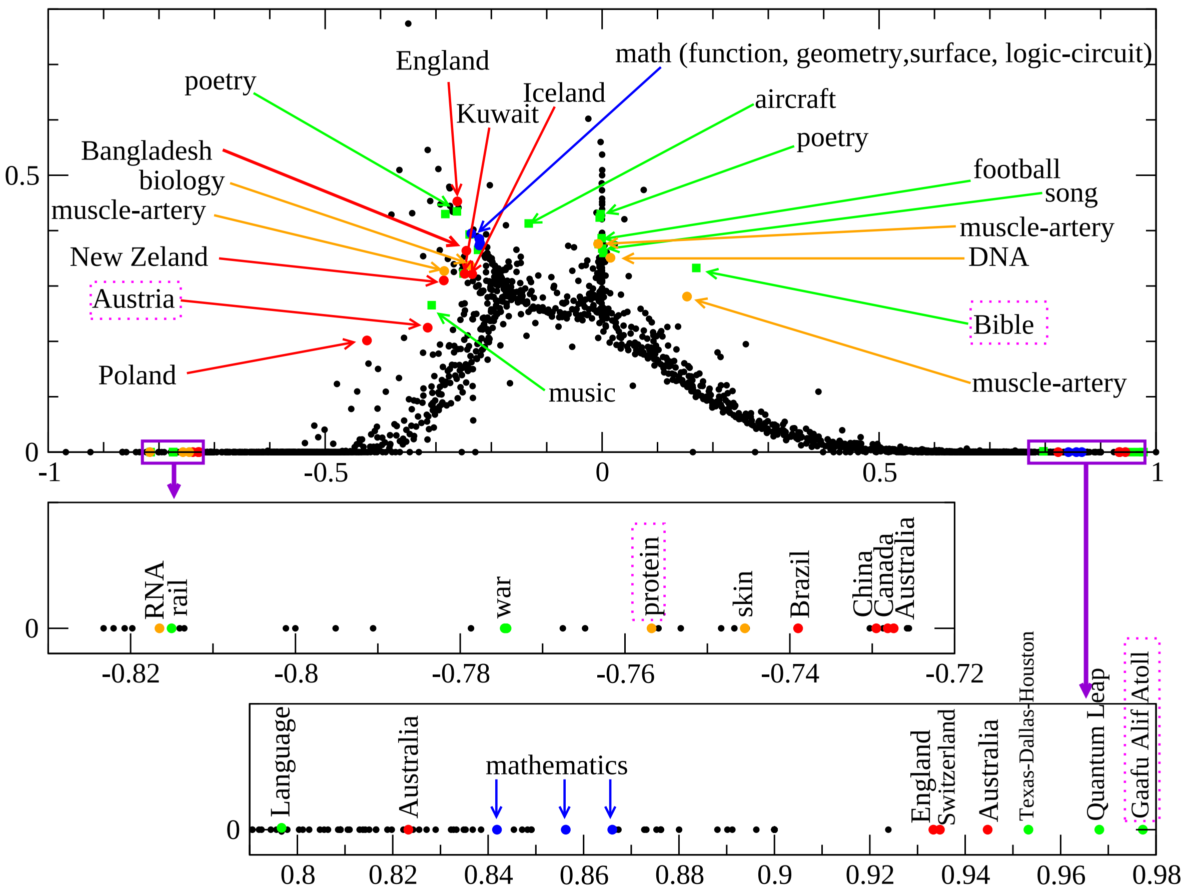


left (right): PageRank (CheiRank)

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:



Q: How to analyze network structure for such and also more general sub-groups ? (⇒ reduced Google matrix, see below)

Anderson Localization

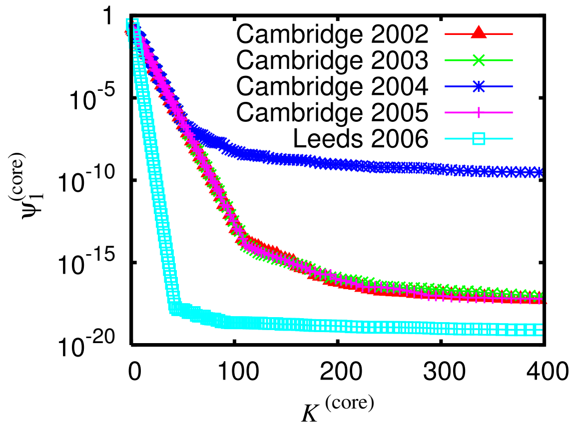
(*Anderson, Phys. Rev.* **109***, 1492 (1958); Nobel Prize 1977*).

$$H\psi_n = -t\psi_{n-1} - t\psi_{n+1} + \varepsilon_n\psi_n = E\psi_n \quad , \quad -\frac{W}{2} < \varepsilon_n < \frac{W}{2}$$

Exponential localization in $d = 1, 2$ (for $W > 0$) and $d = 3$ (for $W > W_c = 16.5t$).

For typical PageRank vectors of WWW or Wikipedia there is typically a power law localization : $P(j) \sim K(j)^{-\beta}$ with $\beta \approx 0.9$.

However, some special cases for leading core space eigenvector :



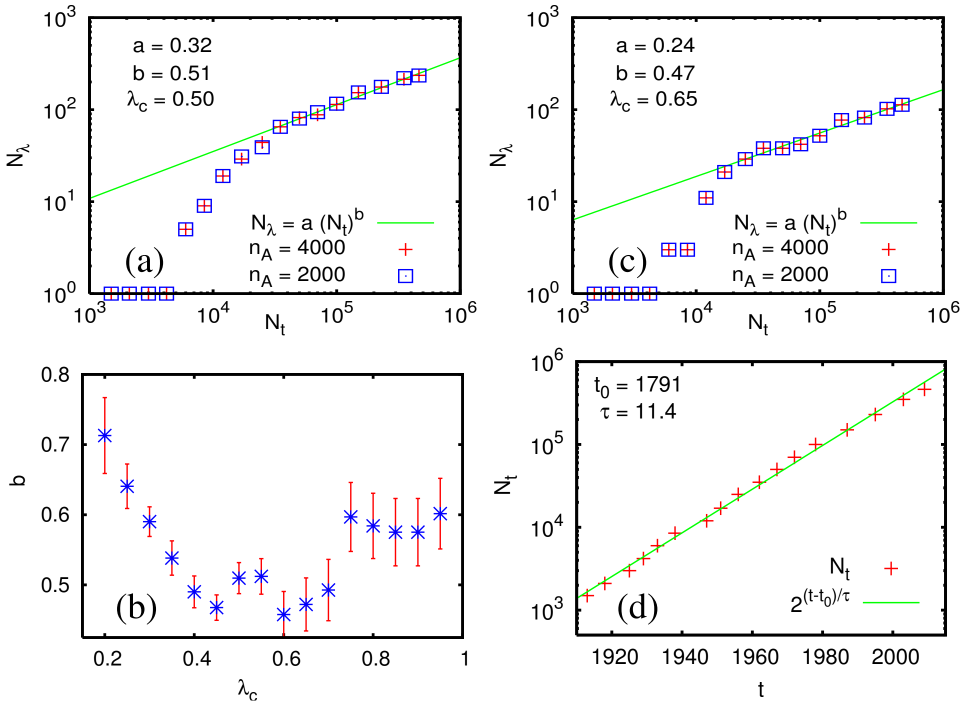
	$1 - \lambda_1^{(core)}$
Cambridge 2002	$3.996 \cdot 10^{-17}$
Cambridge 2003	$4.01 \cdot 10^{-17}$
Cambridge 2004	$2.91 \cdot 10^{-9}$
Cambridge 2005	$4.01 \cdot 10^{-17}$
Leeds 2006	$3.126 \cdot 10^{-19}$

(*KF, Georgeot and Shepelyansky, J. Phys. A* **44***, 465101 (2011).*)

Fractal Weyl law

(*KF, Eom, Shepelyansky. PRE* **89**, *052814 (2014).*)

Number of states N_λ with $|\lambda| > \lambda_c$: $N_\lambda \sim (N_t)^b$ with $N_t =$ network size of Physical Review citation network at time t:



Linux kernel function network: $N_\lambda \sim N^\nu$, $\nu \approx 0.65$ ($\lambda_c = 0.1$ or 0.25).

(*Ermann, Chepelianskii and Shepelyansky, EPJB* **79**, *115 (2011)*)

(see also talk of S. Nonnenmacher.)

Random Perron-Frobenius matrices

(*KF, Eom, Shepelyansky. PRE* **89**, *052814 (2014).*)

Random matrix ensembles for **hermitian/hamiltonian matrices** well known since 1955 (**Wigner**, *Annals of Mathematics.* **62**, *548 (1955); Nobel Prize 1963*).

Construct random matrix ensembles G_{ij} for **PF-matrices** such that:

$G_{ij} \geq 0$, G_{ij} (approximately) non-correlated, distributed with 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.

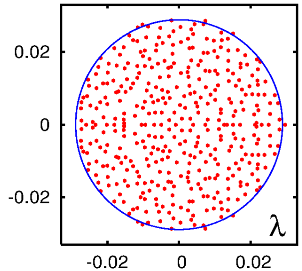
full $\quad \Rightarrow \quad R = 1/\sqrt{3N}$

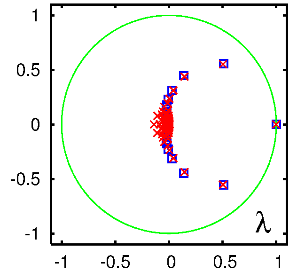
sparse with Q non-zero elements per column $\quad \Rightarrow \quad R \sim 1/\sqrt{Q}$

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

Numerical verification:

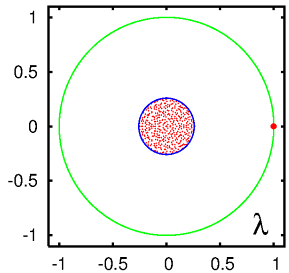
uniform full:
$N = 400$

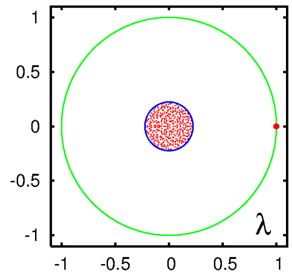




triangular random and average

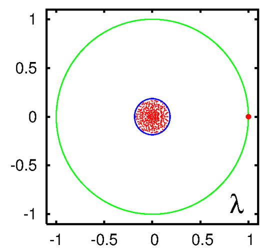
uniform sparse:
$N = 400,$
$Q = 20$

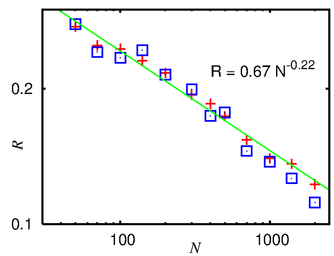




constant sparse:
$N = 400,$
$Q = 20$

power law:
$b = 2.5$





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

Reduced Google matrix

(*KF, Shepelyansky arXiv:1602.02394; KF, Jaffrès-Runser, Shepelyansky, EPJB* **89**, *269 (2016).*)

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

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

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

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

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

containing ***direct link contributions*** from G_{rr} and ***scattering contributions*** from $G_{rs}(\mathbf{1} - G_{ss})^{-1} G_{sr}$.
G_R has the same symmetries as G: $(G_R)_{ij} \ge 0$ and $\sum_i (G_R)_{ij} = 1$.

Analogy with quantum chaotic scattering: $S = \mathbf{1} - 2iW^\dagger \frac{\mathbf{1}}{E-H+iWW^\dagger} W$.

(*Mahaux and Weidenmüller, Phys. Rev.* **170**, *847 (1968).*)

Problem: practical 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: $1 - \lambda_c \ll 1$.

More efficient expression:

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

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.

$\Rightarrow$ three components of $G_{\rm R}$:

$$\boxed{G_{\rm R} = G_{rr} + G_{\rm pr} + G_{\rm qr}}$$

$$\boxed{G_{rr} = \text{rr sub-block of } G \quad \Rightarrow \quad \text{direct links}}$$

$$\boxed{G_{\rm pr} = G_{rs}\frac{\psi_R \psi_L^T}{1-\lambda_c}G_{sr} = \frac{\tilde\psi_R \tilde\psi_L^T}{1-\lambda_c}\ , \quad \text{rank } 1}$$

with

$$\tilde\psi_R = G_{rs}\psi_R \quad , \quad \tilde\psi_L^T = \psi_L^T G_{sr}$$

$$\boxed{G_{\rm qr} = G_{rs}\left[\mathcal{Q}_c \sum_{l=0}^{\infty} \bar{G}_{ss}^{\,l}\right] G_{sr} \quad \Rightarrow \quad \text{indirect links}}$$

Typically: $G_{\rm pr}$ is numerically dominant but
$G_{\rm qr}$ has a more interesting structure allowing to identify friends/followers.

http://www.quantware.ups-tlse.fr/QWLIB/wikipolitnet/

Application : Main network = **Wikipedia 2013**, different language editions.

Groups = leading 20/40 politicians of certain countries or G20 state leaders.

Node density in $\ln K - \ln K^*-$plane:

20 US, Enwiki

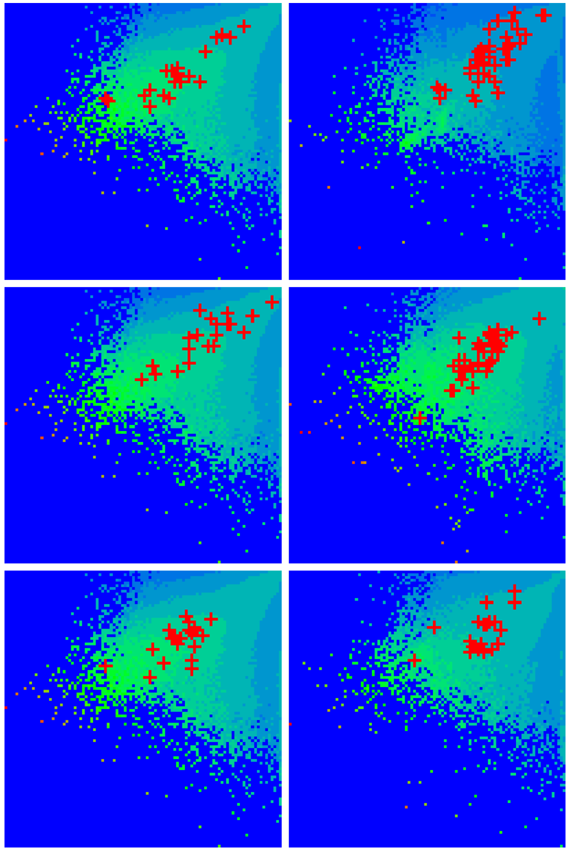
40 DE, Dewiki

20 UK, Enwiki

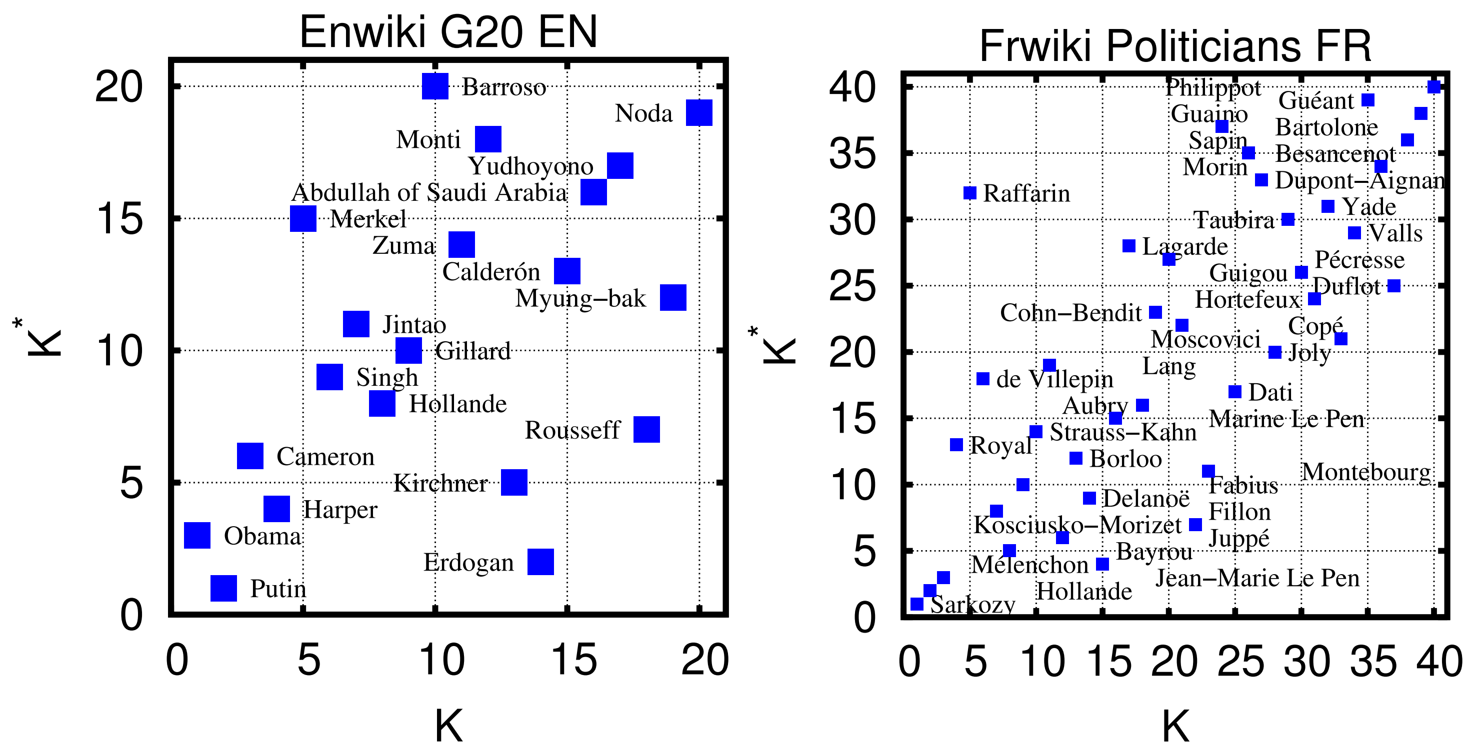
40 FR, Frwiki

G20, Enwiki

20 RU, Ruwiki

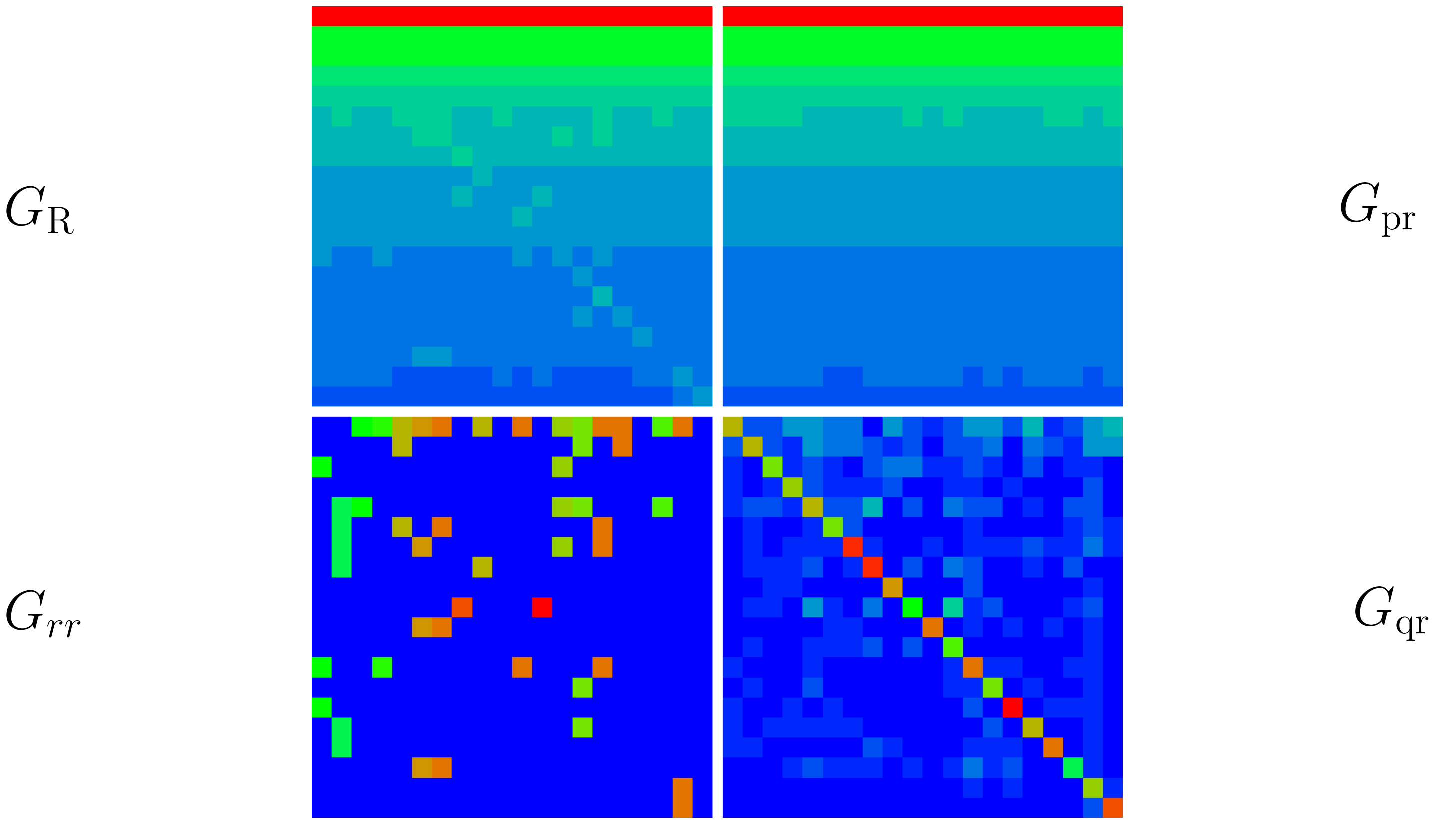


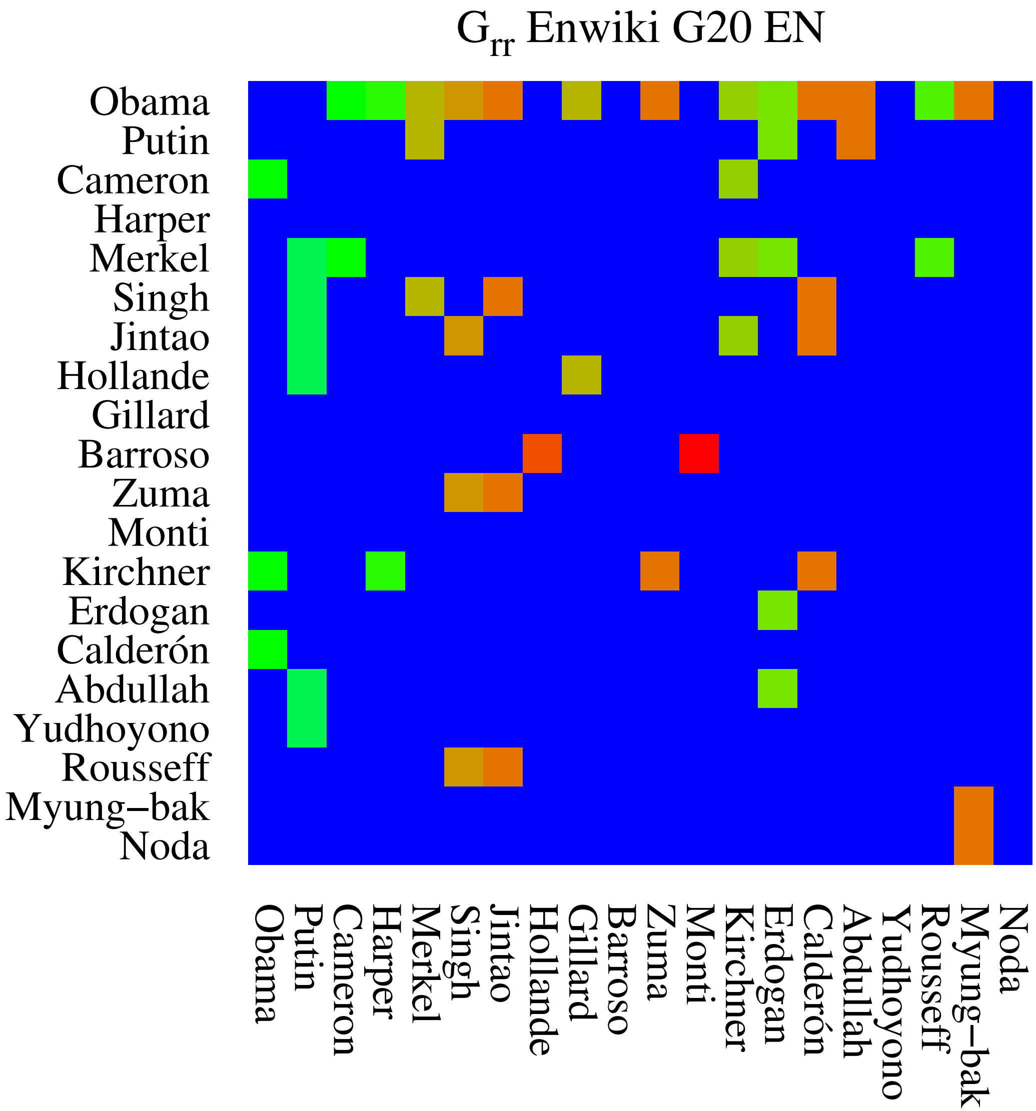
Positions in $K-K^*-$plane

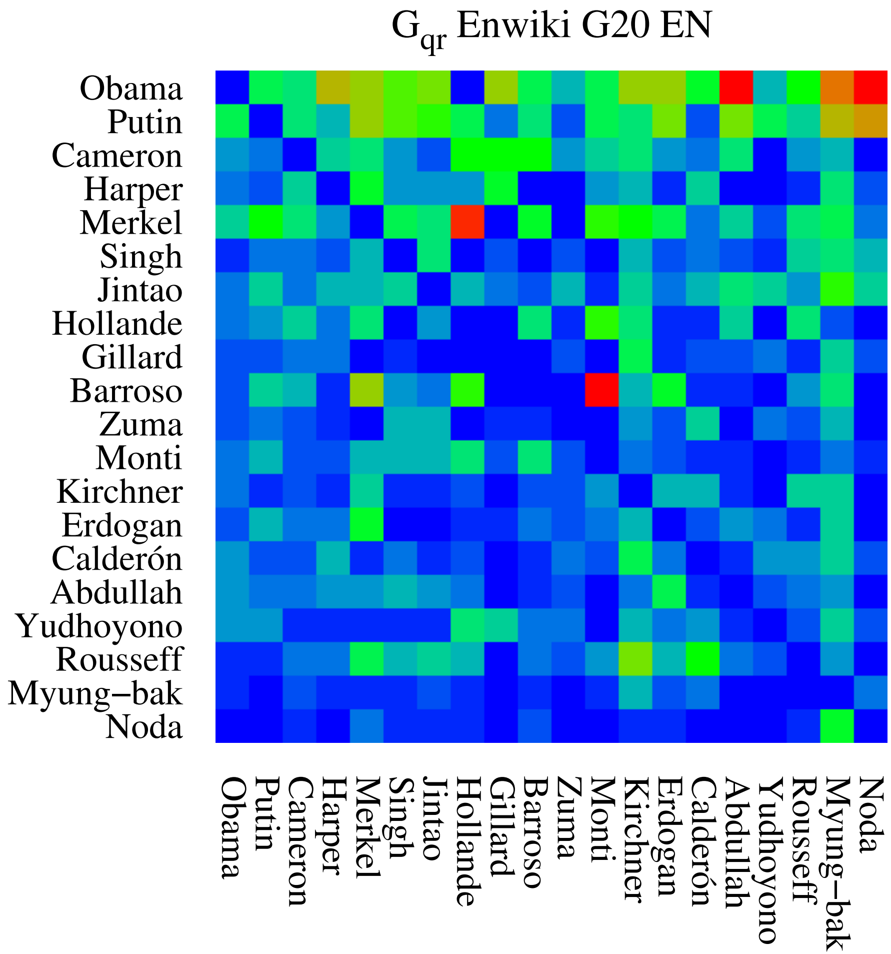


Enwiki G20 state leader

$$1 - \lambda_c = 2.465 \times 10^{-4}$$

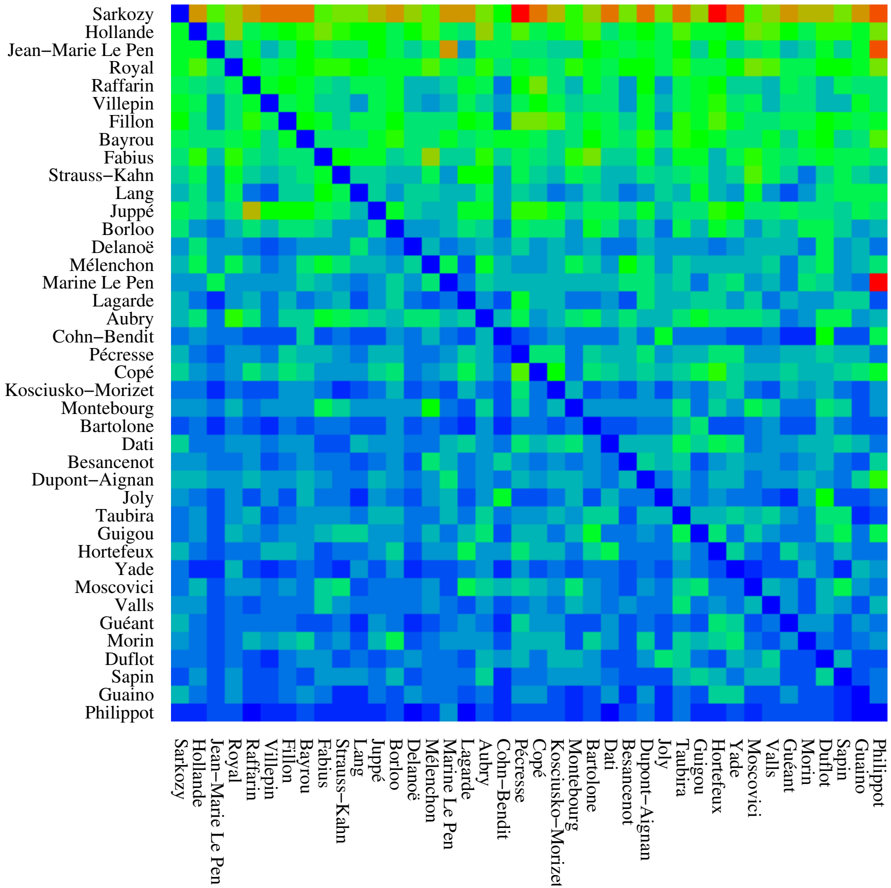



G_{rr} Enwiki G20 EN
Obama
Putin
Cameron
Harper
Merkel
Singh
Jintao
Hollande
Gillard
Barroso
Zuma
Monti
Kirchner
Erdogan
Calderón
Abdullah
Yudhoyono
Rousseff
Myung–bak
Noda

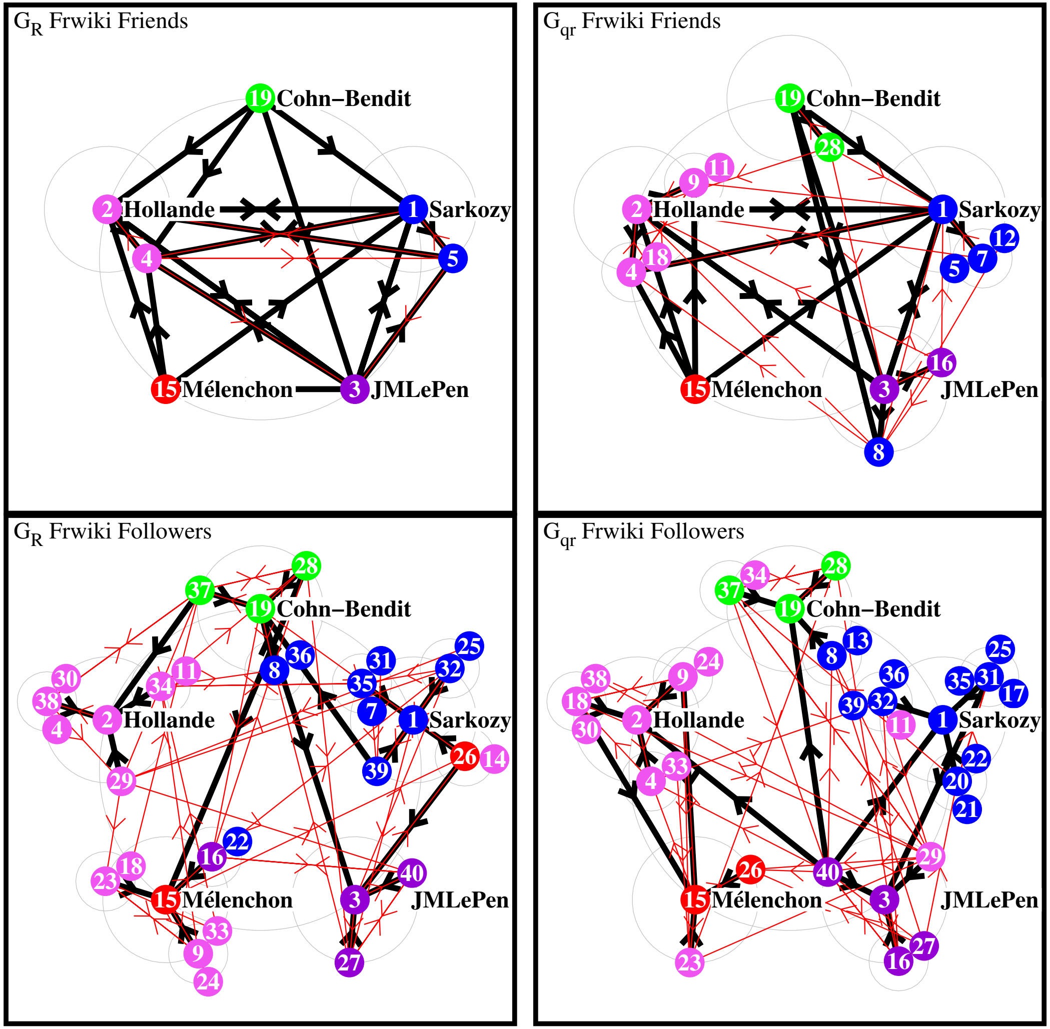


G20	EN	Enwiki
Name	Friends	Followers
Obama	Putin	Noda
	Merkel	Abdullah
	Calderón	Myung-bak
Putin	Merkel	Noda
	Obama	Myung-bak
	Barroso	Merkel
Cameron	Putin	Gillard
	Obama	Barroso
	Merkel	Hollande
Harper	Obama	Merkel
	Cameron	Gillard
	Putin	Myung-bak
Merkel	Barroso	Hollande
	Putin	Monti
	Obama	Kirchner

G_{qr} Frwiki Politicians FR



Politicians	FR	Frwiki
Name	Friends	Followers
Sarkozy	Fillon	Hortefeux
	J.-M. Le Pen	Pécresse
	Hollande	Yade
Hollande	Sarkozy	Royal
	Royal	Aubry
	Fabius	Fabius
J.-M. Le Pen	Sarkozy	Philippot
	M. Le Pen	M. Le Pen
	Bayrou	Taubira
Royal	Hollande	Moscovici
	Sarkozy	Philippot
	Fabius	Hollande
Raffarin	Sarkozy	Copé
	Juppé	Pécresse
	Fillon	Hortefeux



(Other applications of G_R $\rightarrow$ *talk of K. Jaffrès-Runser.)*

Ising-PageRank

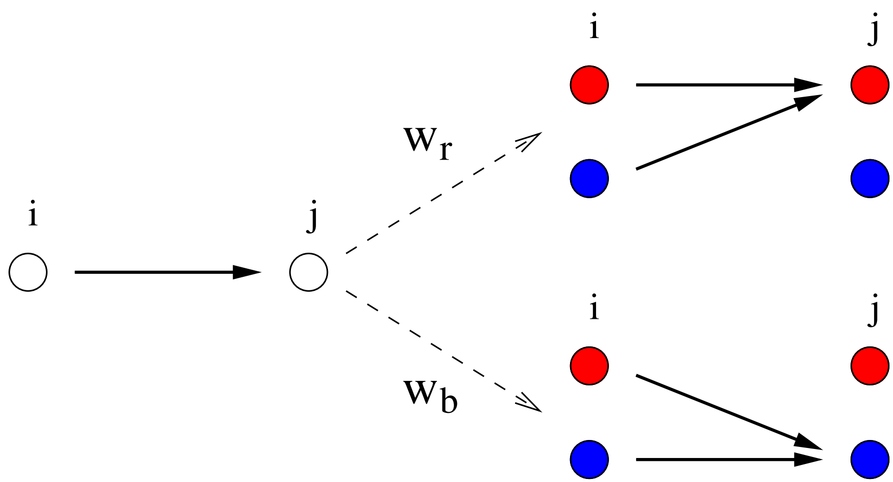
Ising model: $$H = -J \sum_{<i,j>} S_i S_j - h \sum_i S_i \quad , \quad S_i = \pm 1.$$

(*Ising, Z. Phys.*, **31**, *253 (1925)*)

Ising model of Google matrix:

(*work in progress*)

Double network size (of a given network such as Wikipedia etc.) into red and blue nodes and attribute to each node i a preference with probability w_r (or $w_b = 1 - w_r$) to link to other red (blue) nodes:

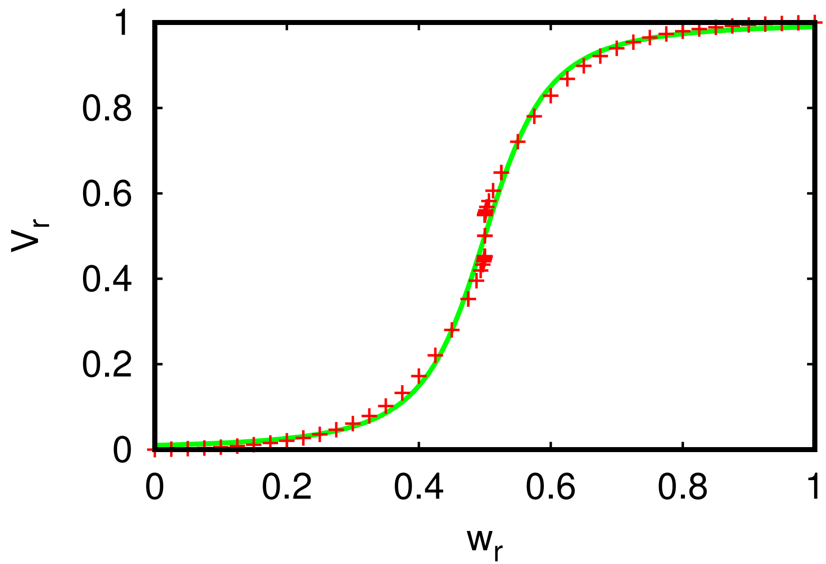


Preferential vector for dangling nodes or damping factor :

$$\frac{1}{N}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \rightarrow \quad \frac{1}{N}\begin{pmatrix} w_r \\ w_b \\ \vdots \\ w_r \\ w_b \end{pmatrix}$$

PageRank $P_r(i)$ (or $P_b(i)$) for red (blue) nodes.

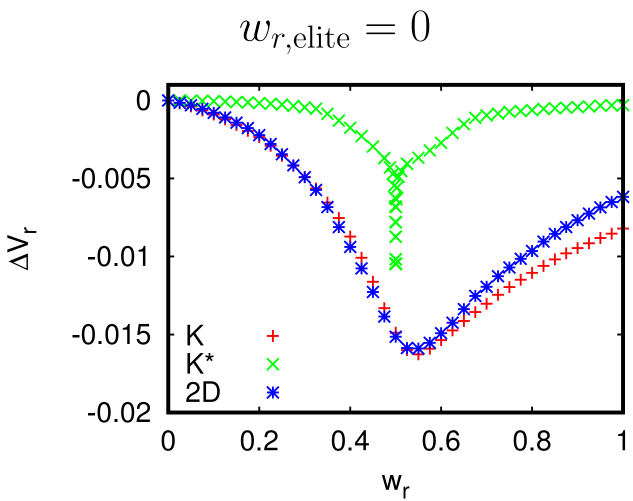
Vote: $V_r = \#\{\text{nodes } i \mid P_r(i) > P_b(i)\}/N$ (for english Wikipedia 2017):

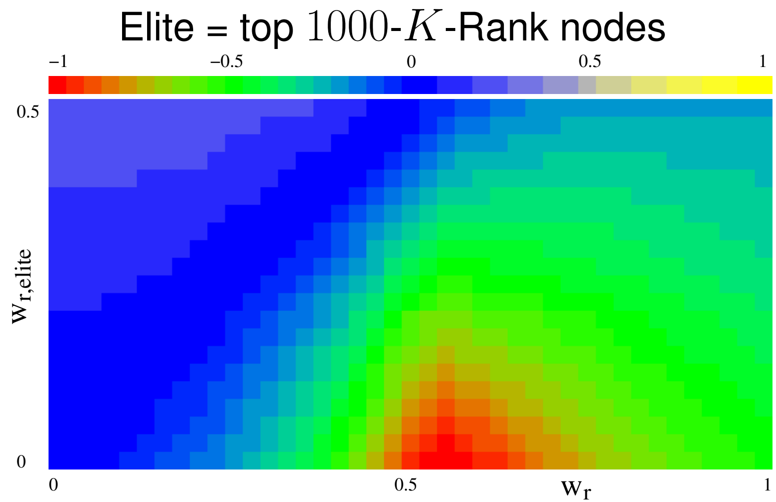


Effect of Elite Nodes

Different probabilities of red preference $w_{r,\text{elite}}$ for N_{El} elite nodes and w_r for other nodes. Elite nodes are selected as N_{El} top nodes according to K-Rank (or K^*-Rank or $2D$-Rank).

Vote modification: $\Delta V_r = V_{r,El} - V_r$
(for english Wikipedia 2017 with $N_{El} = 1000$):





Conclusions

- Google matrix constructed from directed networks (WWW, Wikipedia, Twitter, Linux kernel, PR citation network etc.) with efficient computation of PageRank, leading complex eigenvalues (also exploiting the structure of invariant subspaces) and some eigenvectors.
- Typical power law localization of PageRank but also examples of quasi exponential localization.
- Weyl fractal scaling for certain networks (Linux kernel, PR citation network, certain Ulam networks).
- Different simple models of random PF matrices do not describe the spectra of realistic Google matrices.
- Approach of reduced Google matrix $G_{\mathbf{R}}$ for sub-networks of Wikipedia etc. Decomposition of $G_{\mathbf{R}}$ in three contributions; construction of friend/follower network using $G_{\mathbf{R}}$ or $G_{\mathbf{qr}}$; different language editions of Wikipedia allow to take into account multi-cultural aspects.
- Ising-PageRank for networks with a doubled number of (red and blue) nodes. Effect of selected elite nodes on the vote.