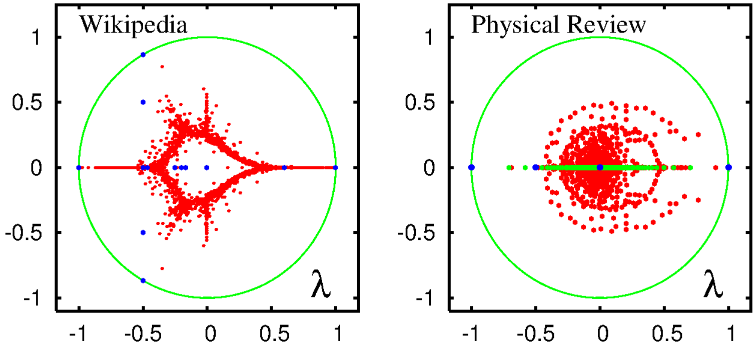


Wikipedia mining of hidden links between political leaders using reduced Google matrix

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

XIII Brunel-Bielefeld Workshop on Random Matrix Theory

Bielefeld, December 14 - 16, 2017

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 $\alpha = 0.85$):

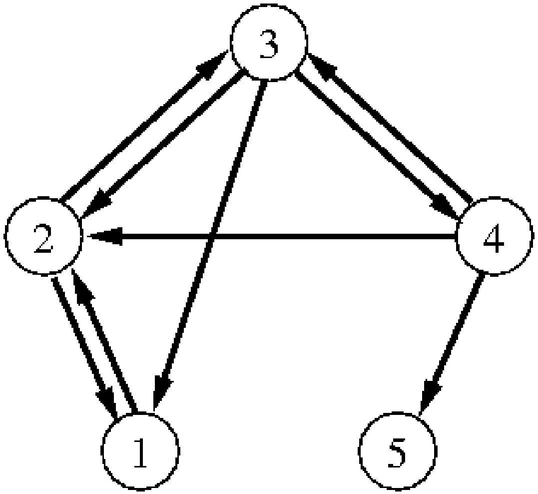
Google matrix:

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

$\Rightarrow \quad \lambda_1$ 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:



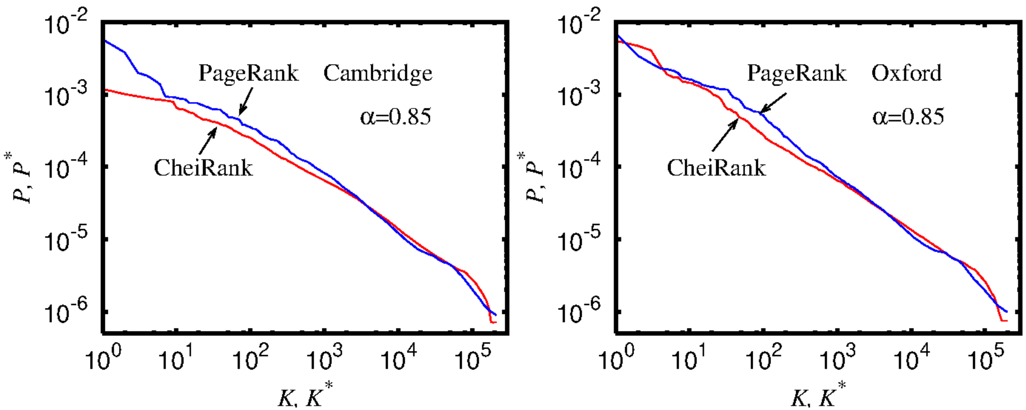
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

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

PageRank

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

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 ($N \sim 3 \times 10^5$)

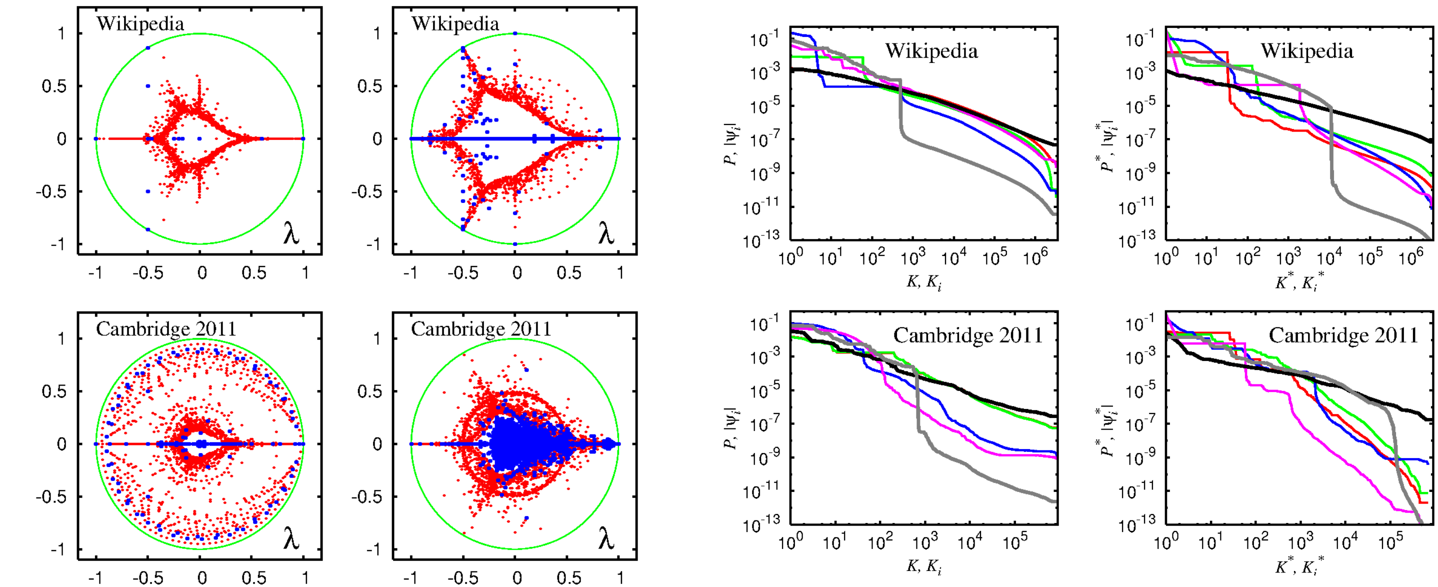
  (Nodes = kernel functions)
- Wikipedia, different language editions ($N \sim 4 \times 10^6$, $N_\ell \sim 10^8$)
- World trade network ($N \sim 10^2$ but more complicated structure).
- Twitter 2009 ($N \sim 4 \times 10^7$, $N_\ell \sim 1.5 \times 10^9$).
- Physical Review citation network ($N \sim 5 \times 10^5$, $N_\ell \sim 5 \times 10^6$).

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

Example: Wikipedia 2009

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

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

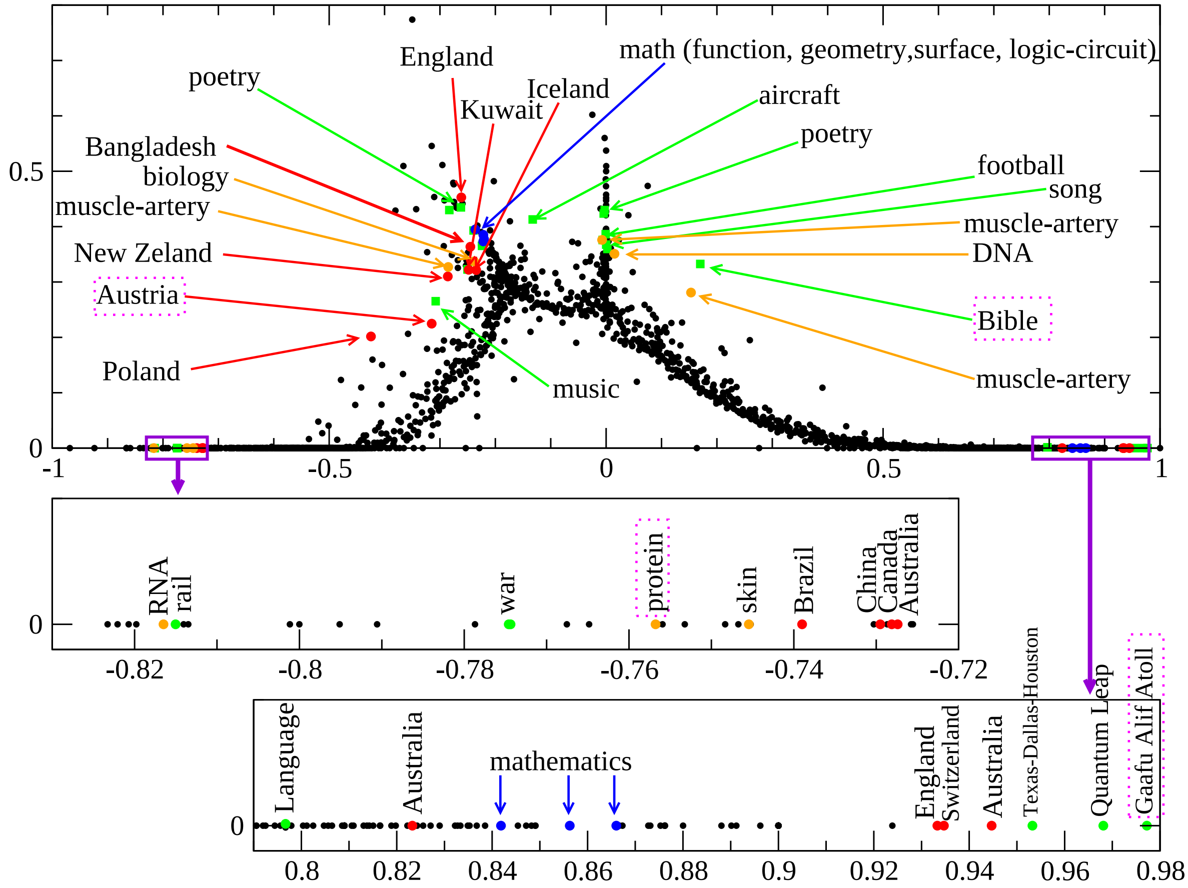


left (right): PageRank (CheiRank)

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

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

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

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

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

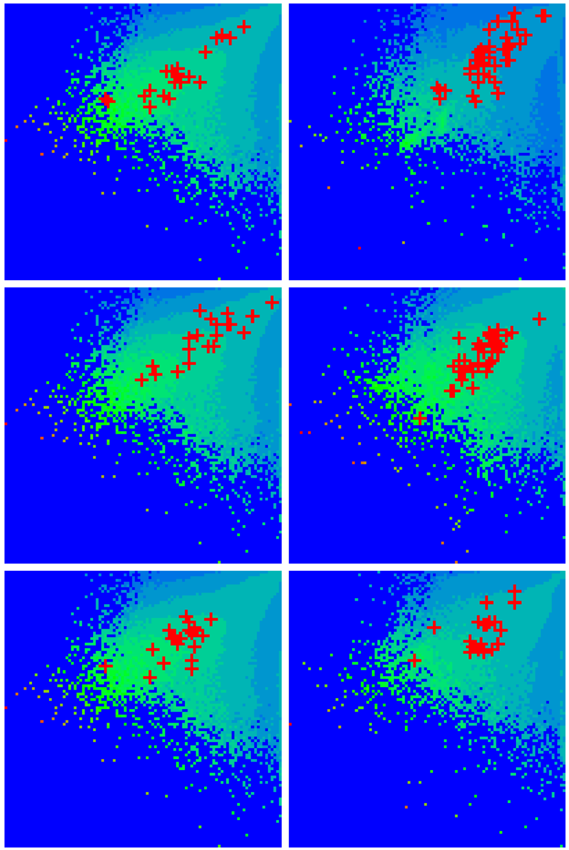
$$\tilde{\psi}_R = G_{rs}\psi_R \quad , \quad \tilde{\psi}_L^T = \psi_L^T G_{sr}$$

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

Typically: $G_{\rm pr}$ is numerically dominant but

$G_{\rm qr}$ has a more interesting structure allowing to identify friends/followers.

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

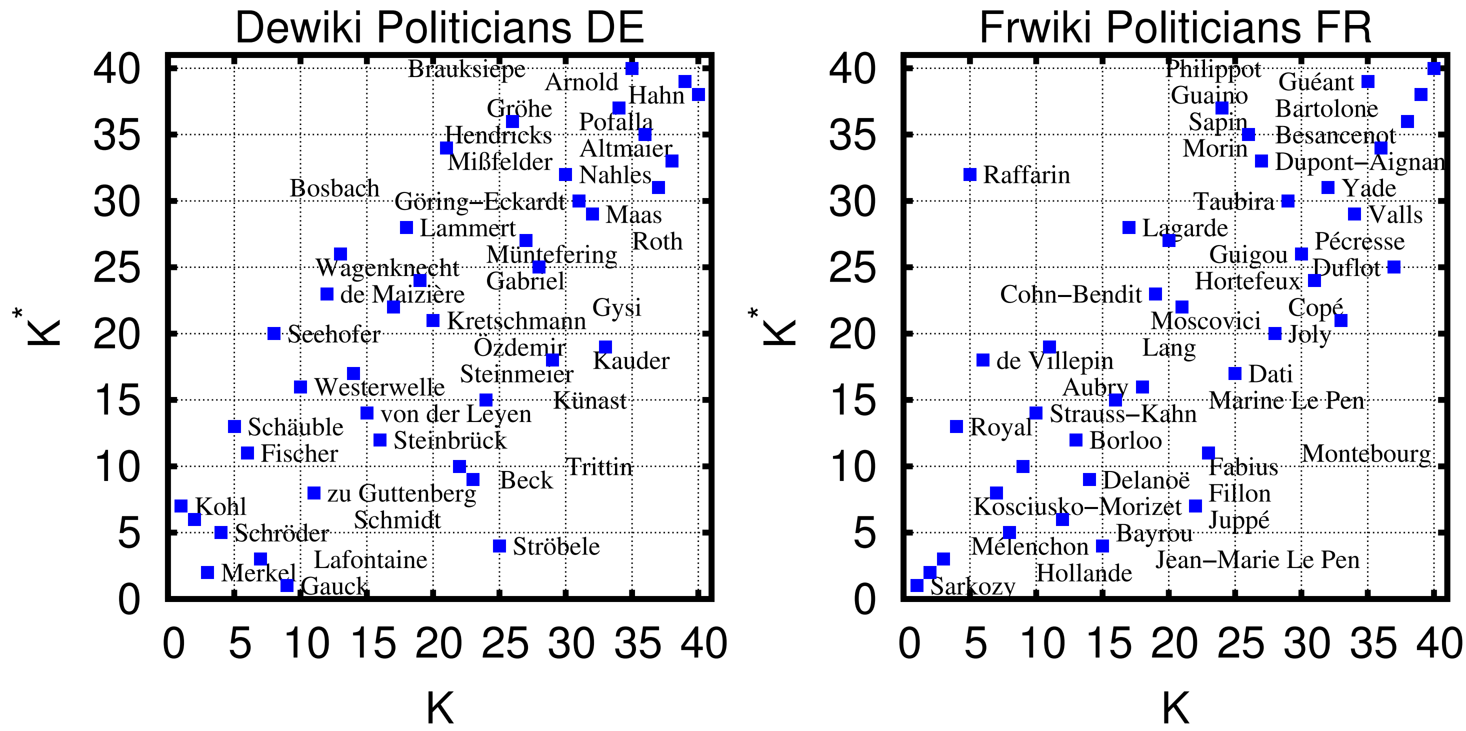
20 UK, Enwiki

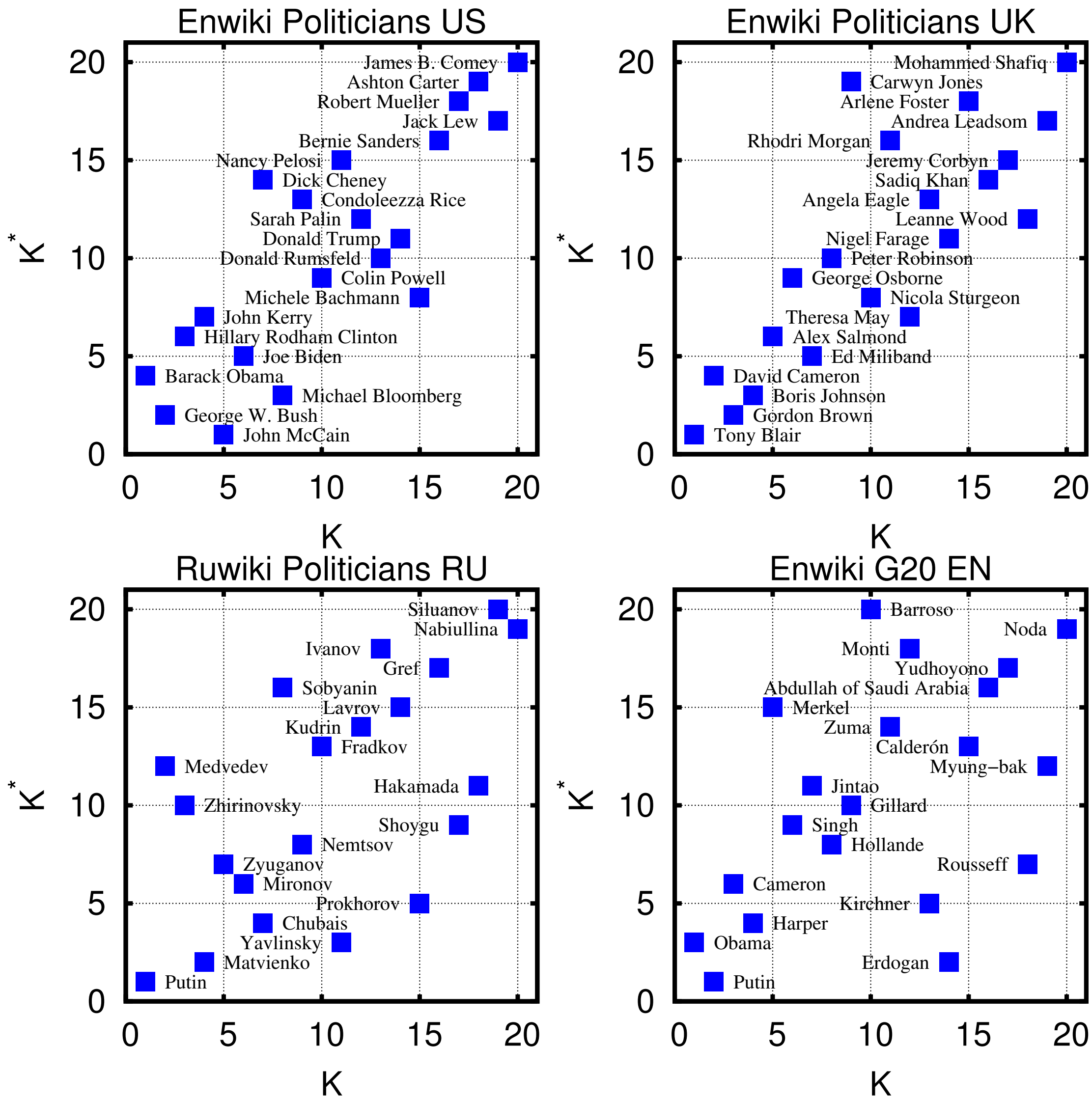
40 FR, Frwiki

G20, Enwiki

20 RU, Ruwiki

Positions in $K - K^*$-plane

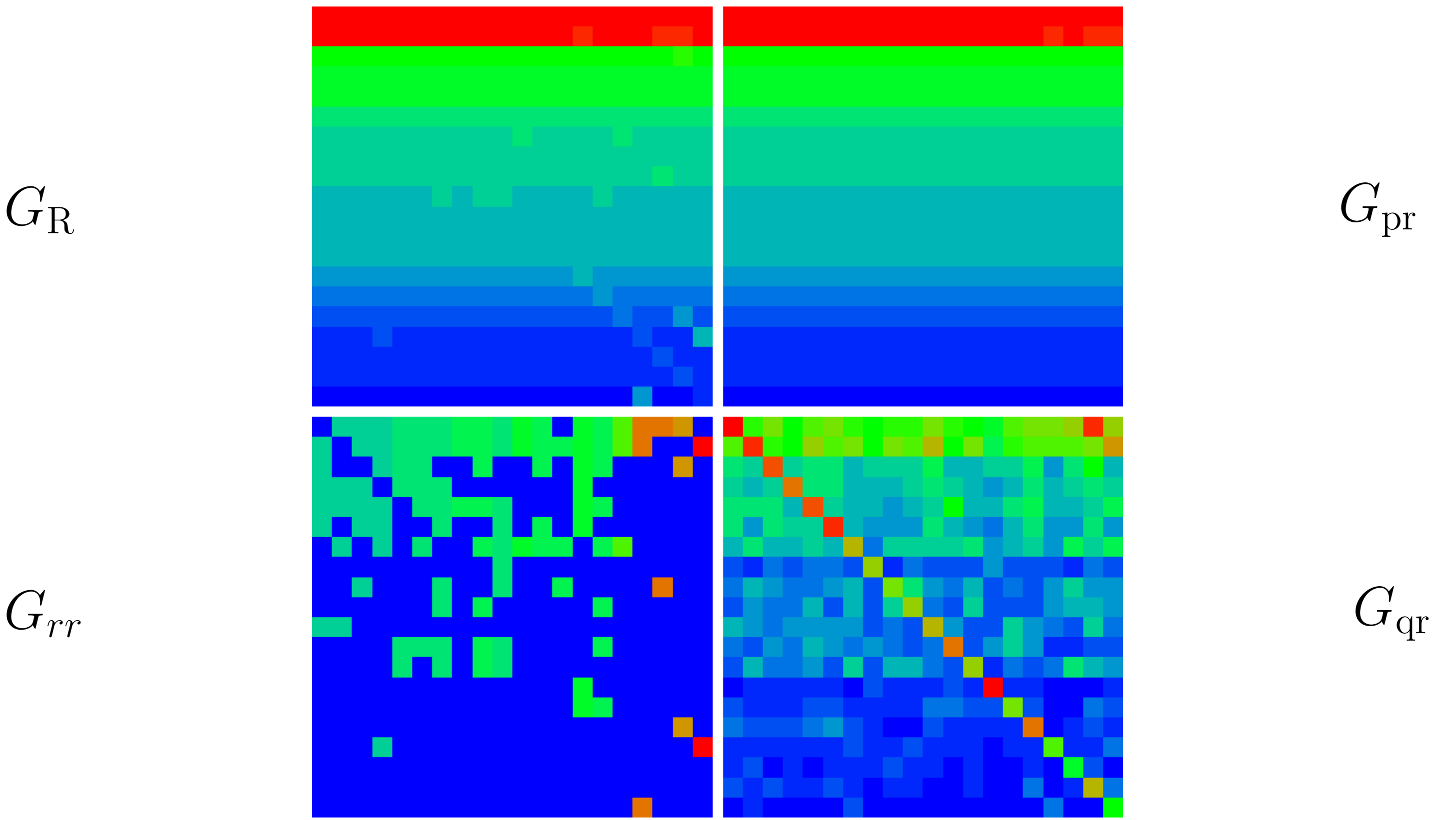


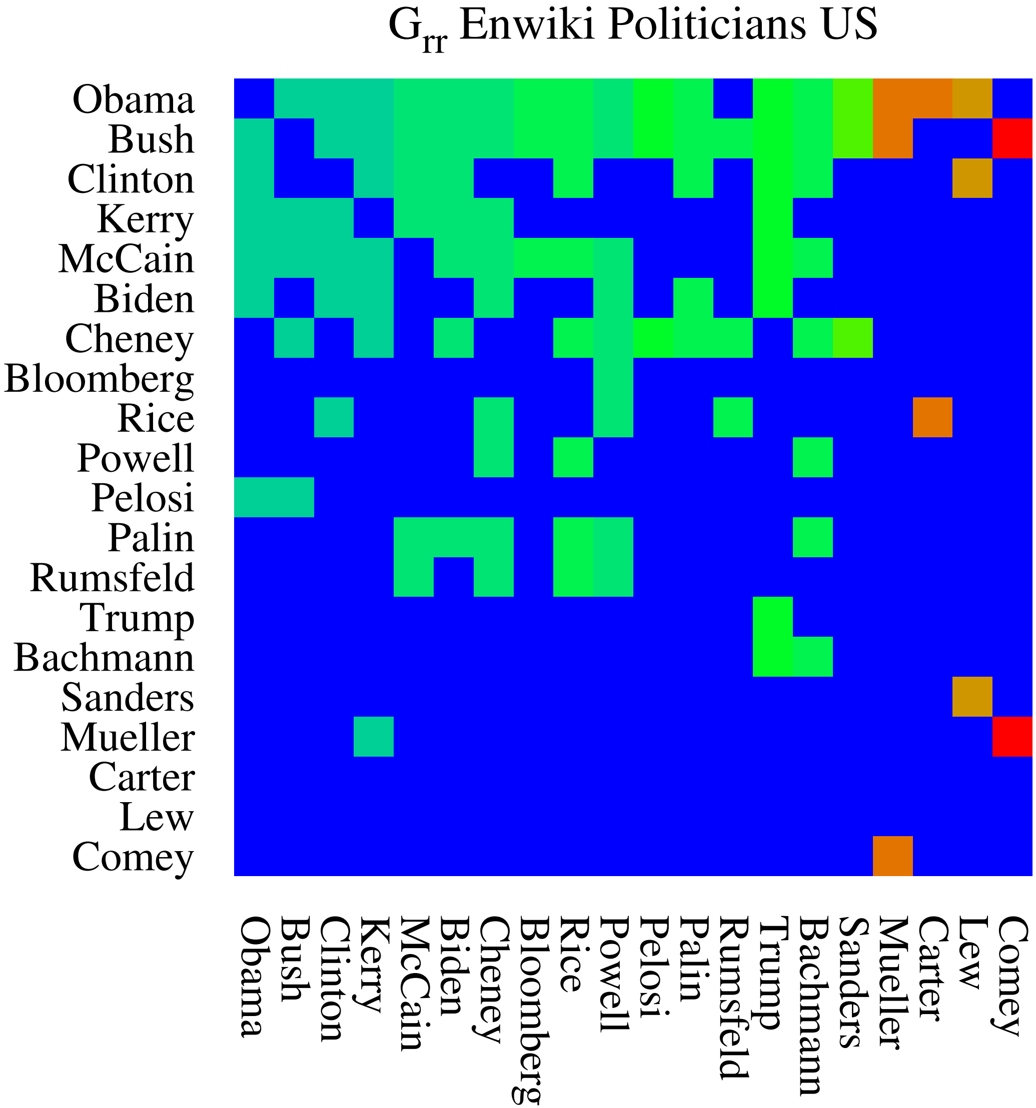


Enwiki Politicians US
James B. Comey
Ashton Carter
Robert Mueller
Jack Lew
Bernie Sanders
Nancy Pelosi
Dick Cheney
Condoleezza Rice
Sarah Palin
Donald Trump
Donald Rumsfeld
Colin Powell
Michele Bachmann
John Kerry
Hillary Rodham Clinton
Joe Biden
Barack Obama
Michael Bloomberg
George W. Bush
John McCain
Enwiki Politicians UK
Mohammed Shafiq
Carwyn Jones
Arlene Foster
Andrea Leadsom
Rhodri Morgan
Jeremy Corbyn
Sadiq Khan
Angela Eagle
Leanne Wood
Nigel Farage
Peter Robinson
George Osborne
Nicola Sturgeon
Theresa May
Alex Salmond
Ed Miliband
David Cameron
Boris Johnson
Gordon Brown
Tony Blair
Ruwiki Politicians RU
Siluanov
Nabiullina
Ivanov
Gref
Sobyanin
Lavrov
Kudrin
Fradkov
Medvedev
Hakamada
Zhirinovsky
Shoygu
Nemtsov
Zyuganov
Mironov
Prokhorov
Chubais
Yavlinsky
Matvienko
Putin
Enwiki G20 EN
Barroso
Noda
Monti
Yudhoyono
Abdullah of Saudi Arabia
Merkel
Zuma
Calderón
Myung−bak
Jintao
Gillard
Singh
Hollande
Rousseff
Cameron
Kirchner
Harper
Obama
Erdogan
Putin
K*
K
0
5
10
15
20

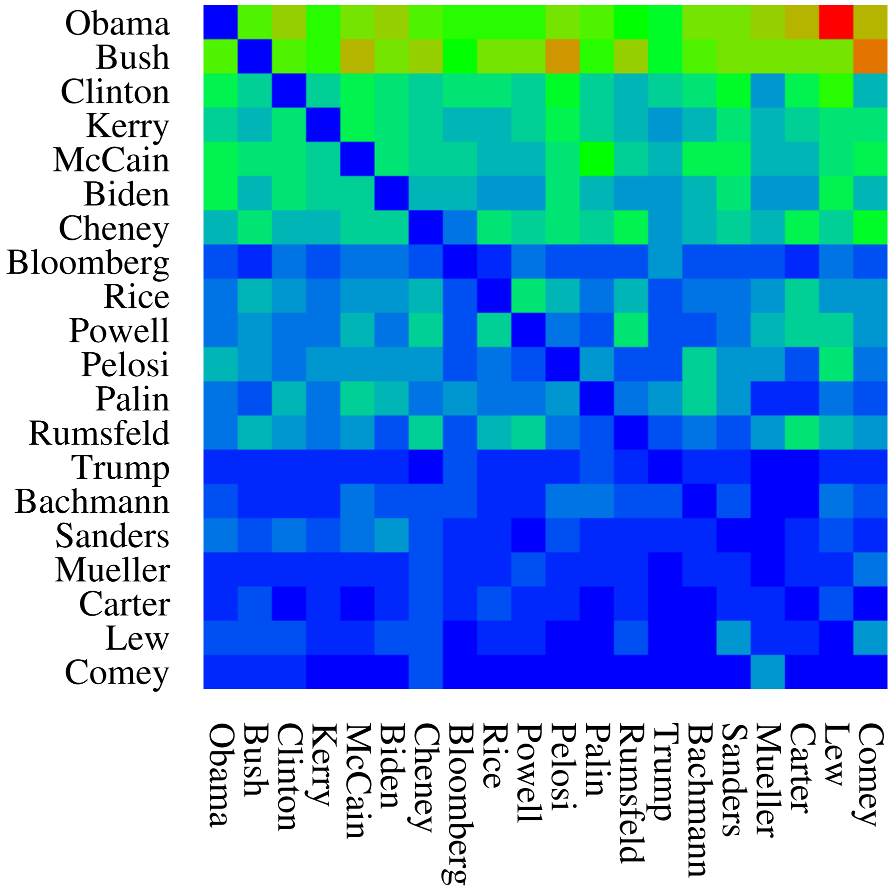
Enwiki Politicians US

$$1 - \lambda_c = 3.68 \times 10^{-4}$$

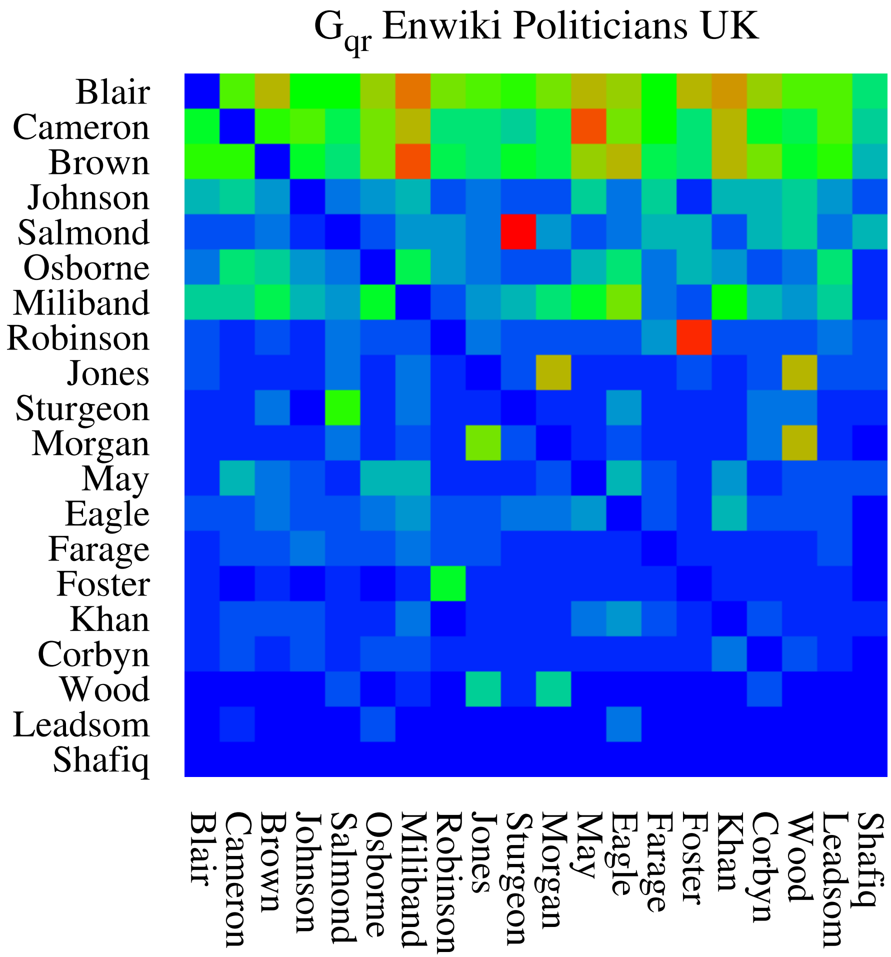



G_{rr} Enwiki Politicians US
Obama
Bush
Clinton
Kerry
McCain
Biden
Cheney
Bloomberg
Rice
Powell
Pelosi
Palin
Rumsfeld
Trump
Bachmann
Sanders
Mueller
Carter
Lew
Comey

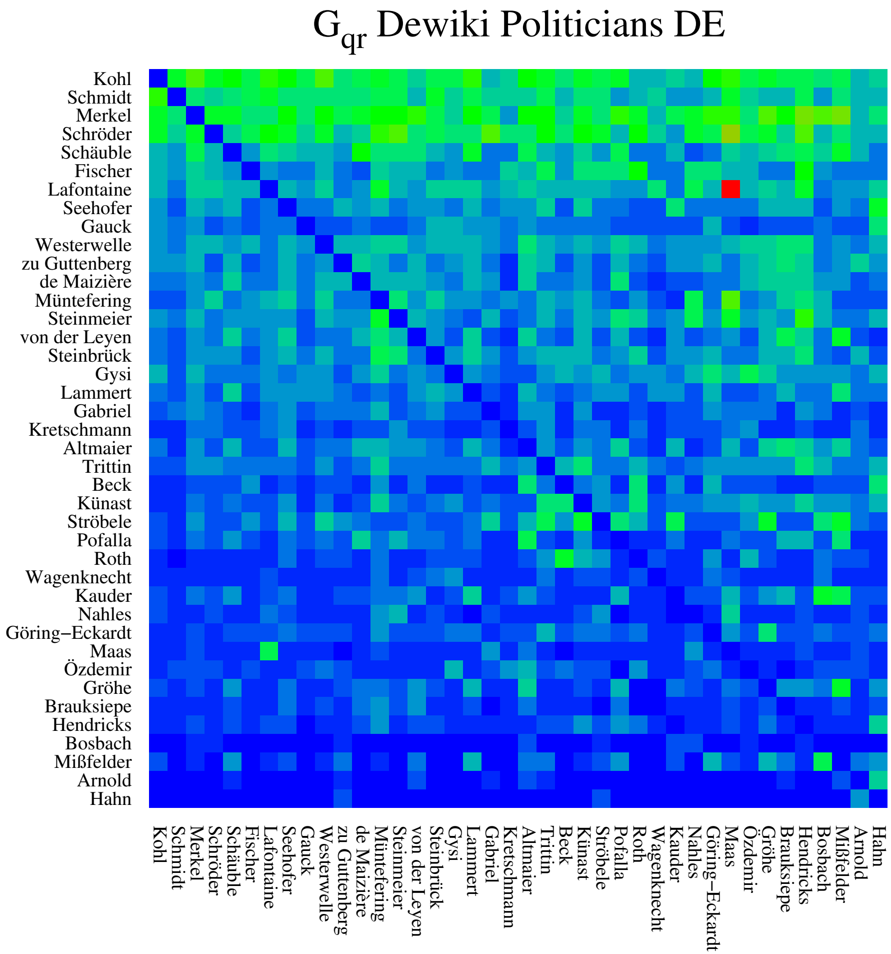
G_{qr} Enwiki Politicians US



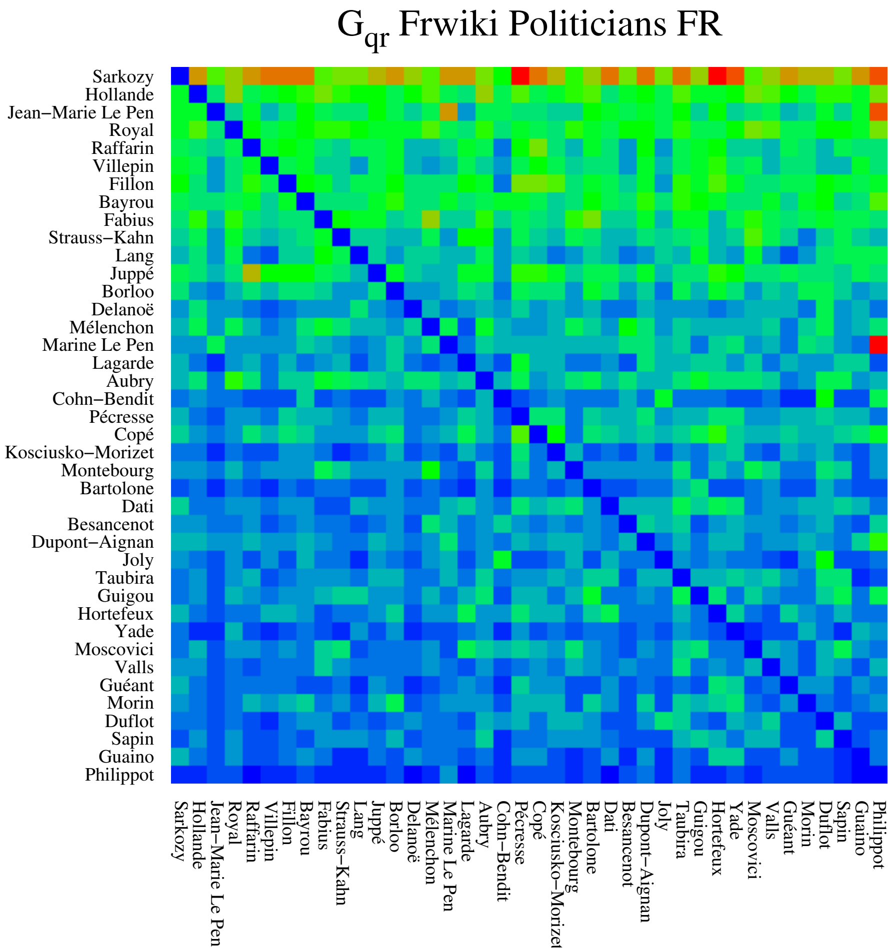
Politicians	US	Enwiki
Name	Friends	Followers
Obama	Bush	Lew
	Clinton	Comey
	Biden	Carter
Bush	Obama	Comey
	Cheney	Pelosi
	McCain	McCain
Clinton	Obama	Lew
	Bush	Sanders
	McCain	Pelosi
Kerry	Obama	Pelosi
	Bush	McCain
	Clinton	Biden
McCain	Bush	Palin
	Obama	Comey
	Clinton	Sanders



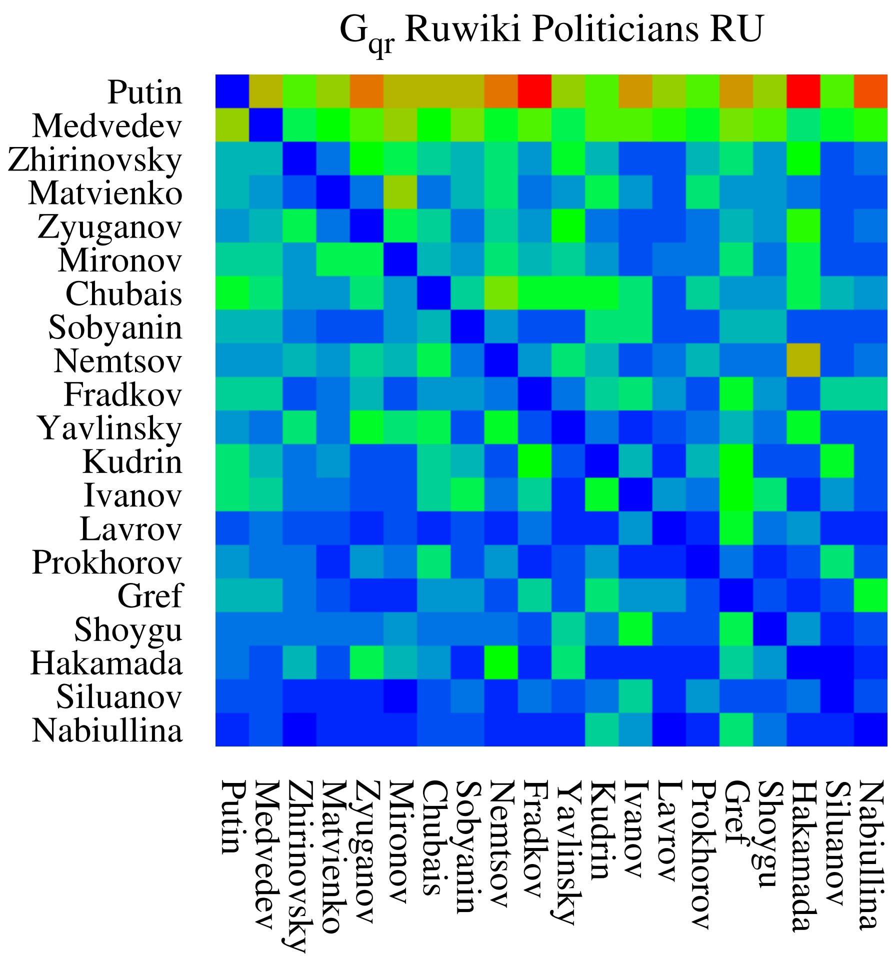
Politicians	UK	Enwiki
Name	Friends	Followers
Blair	Brown	Miliband
	Cameron	Khan
	Miliband	Foster
Cameron	Blair	May
	Brown	Khan
	Osborne	Miliband
Brown	Blair	Miliband
	Cameron	Khan
	Miliband	Eagle
Johnson	Cameron	May
	Blair	Wood
	Brown	Cameron
Salmond	Sturgeon	Sturgeon
	Blair	Wood
	Cameron	Foster



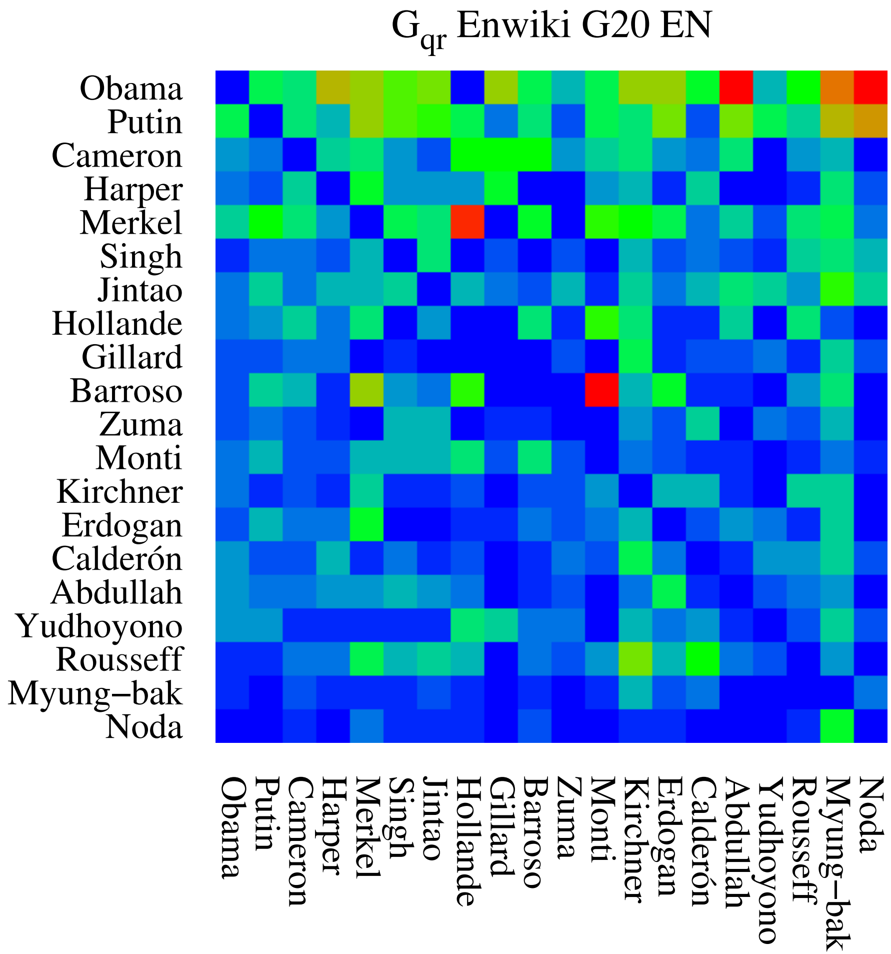
Politicians	DE	Dewiki
Name	Friends	Followers
Kohl	Schmidt Merkel Schröder	Merkel Westerwelle Lafontaine
Schmidt	Kohl Merkel Schröder	Kohl Lafontaine Steinbrück
Merkel	Kohl Schröder Schäuble	Hendricks Mißfelder Gröhe
Schröder	Kohl Merkel Schmidt	Maas Steinmeier Hendricks
Schäuble	Kohl Merkel Schmidt	Maizière Mißfelder Lammert



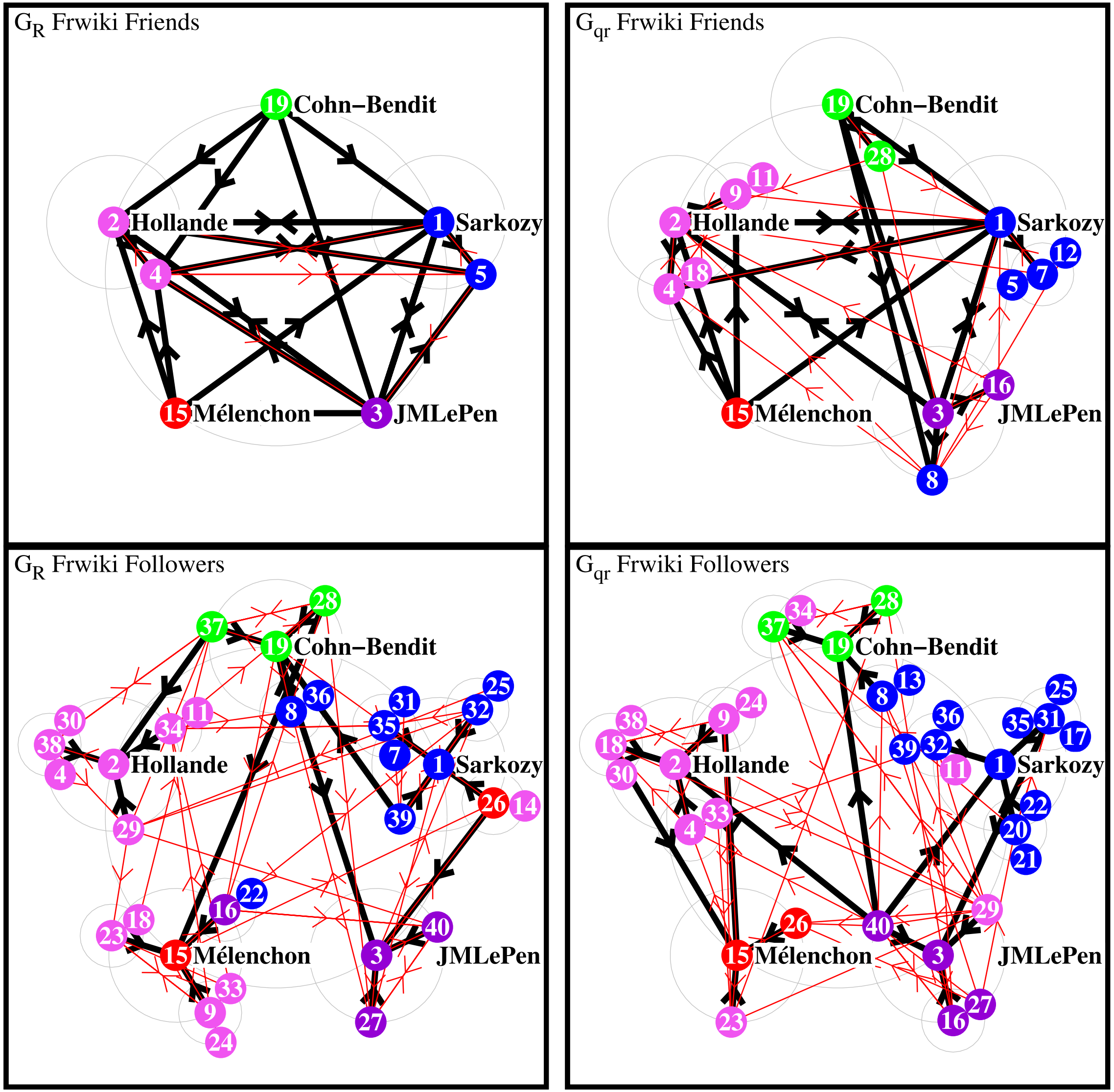
Politicians	FR	Frwiki
Name	Friends	Followers
Sarkozy	Fillon J.-M. Le Pen Hollande	Hortefeux Pécresse Yade
Hollande	Sarkozy Royal Fabius	Royal Aubry Fabius
J.-M. Le Pen	Sarkozy M. Le Pen Bayrou	Philippot M. Le Pen Taubira
Royal	Hollande Sarkozy Fabius	Moscovici Philippot Hollande
Raffarin	Sarkozy Juppé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



Politicians	RU	Ruwiki
Name	Friends	Followers
Putin	Medvedev Chubais Ivanov	Hakamada Fradkov Nabiullina
Medvedev	Putin Chubais Ivanov	Putin Mironov Gref
Zhirinovsky	Putin Medvedev Zyuganov	Zyuganov Hakamada Yavlinsky
Matvienko	Putin Medvedev Mironov	Mironov Kudrin Nemtsov
Zyuganov	Putin Medvedev Zhirinovsky	Hakamada Yavlinsky Zhirinovsky



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



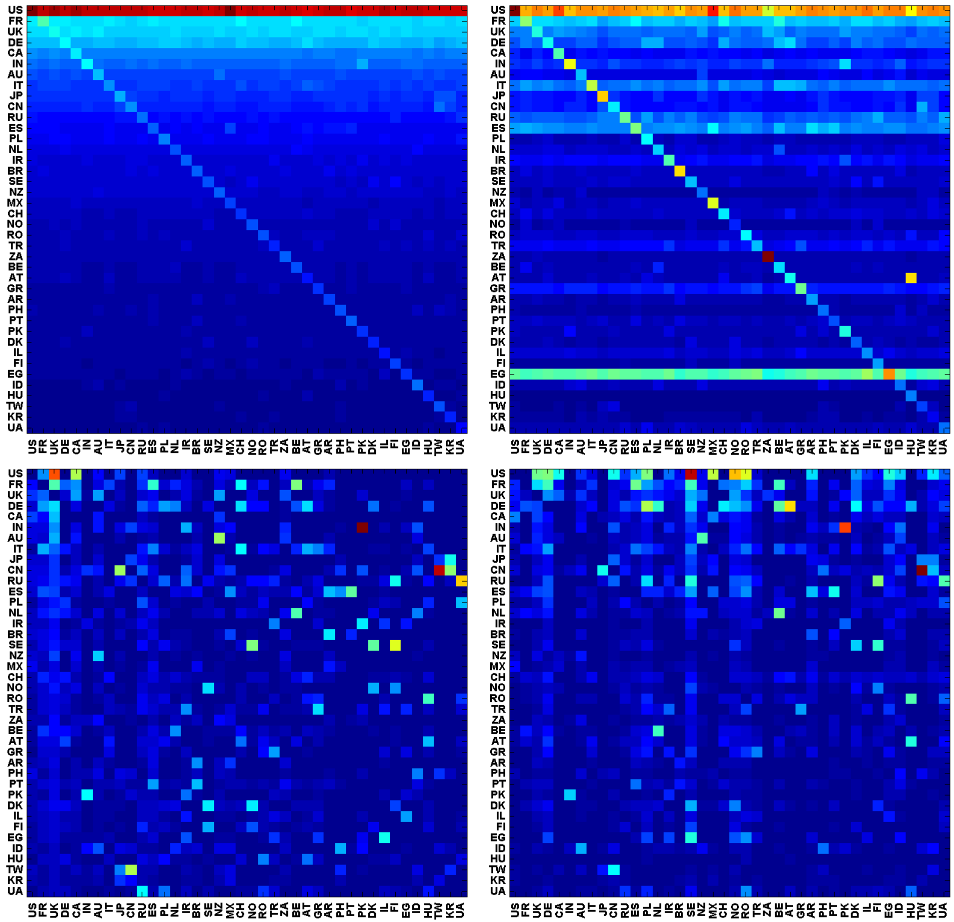
G_R Frwiki Friends
Cohn–Bendit
Hollande
Sarkozy
Mélenchon
JMLePen
G_{qr} Frwiki Friends
Cohn–Bendit
Hollande
Sarkozy
Mélenchon
JMLePen
G_R Frwiki Followers
Cohn–Bendit
Hollande
Sarkozy
Mélenchon
JMLePen
G_{qr} Frwiki Followers
Cohn–Bendit
Hollande
Sarkozy
Mélenchon
JMLePen

Geopolitics interactions between countries

(*KF, Zant, Jaffrès-Runser, Shepelyansky, PLA 381 (2017) 2677.*)

G_{R} Enwiki

G_{R} Arwiki

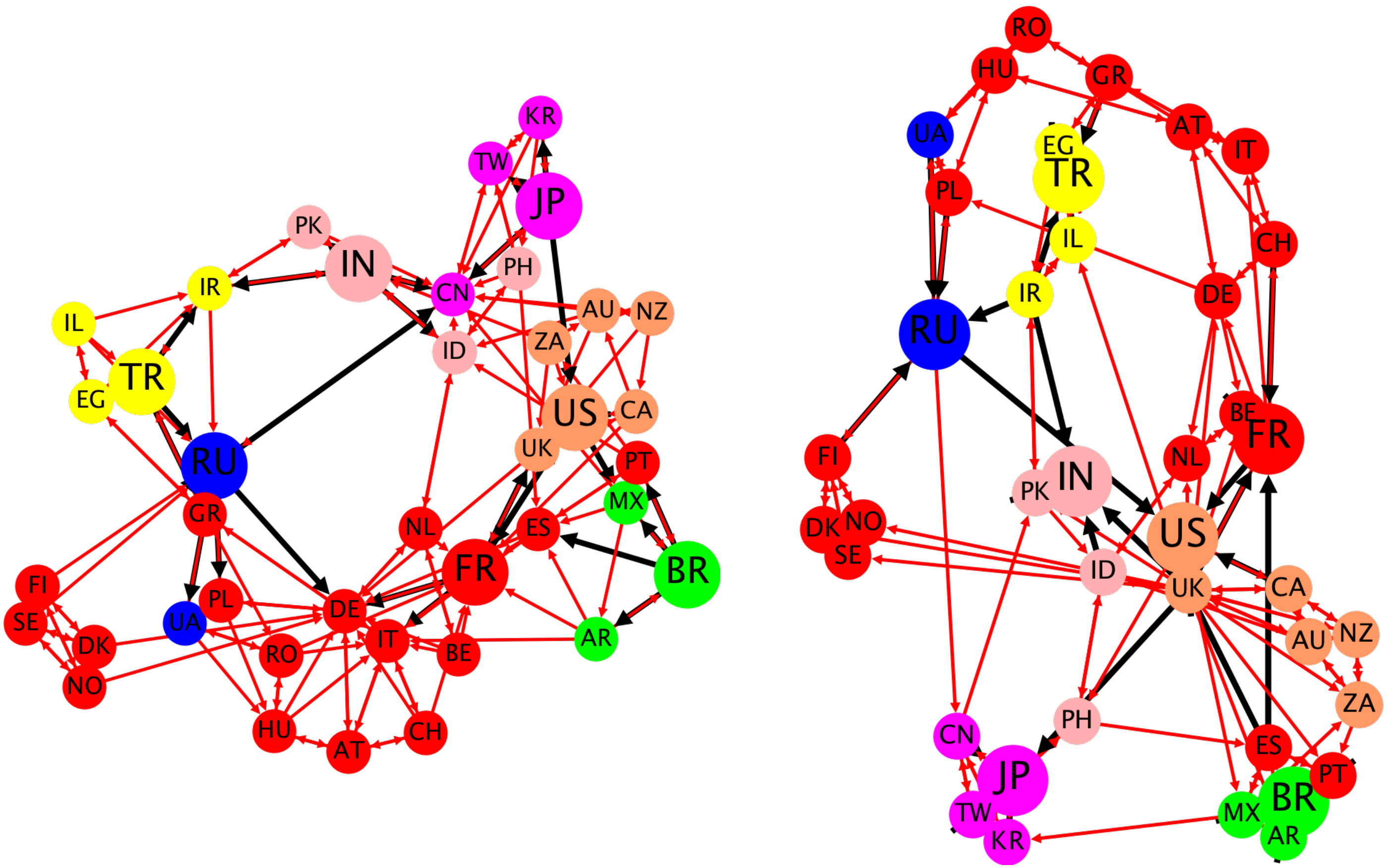


G_{qrnd} Enwiki

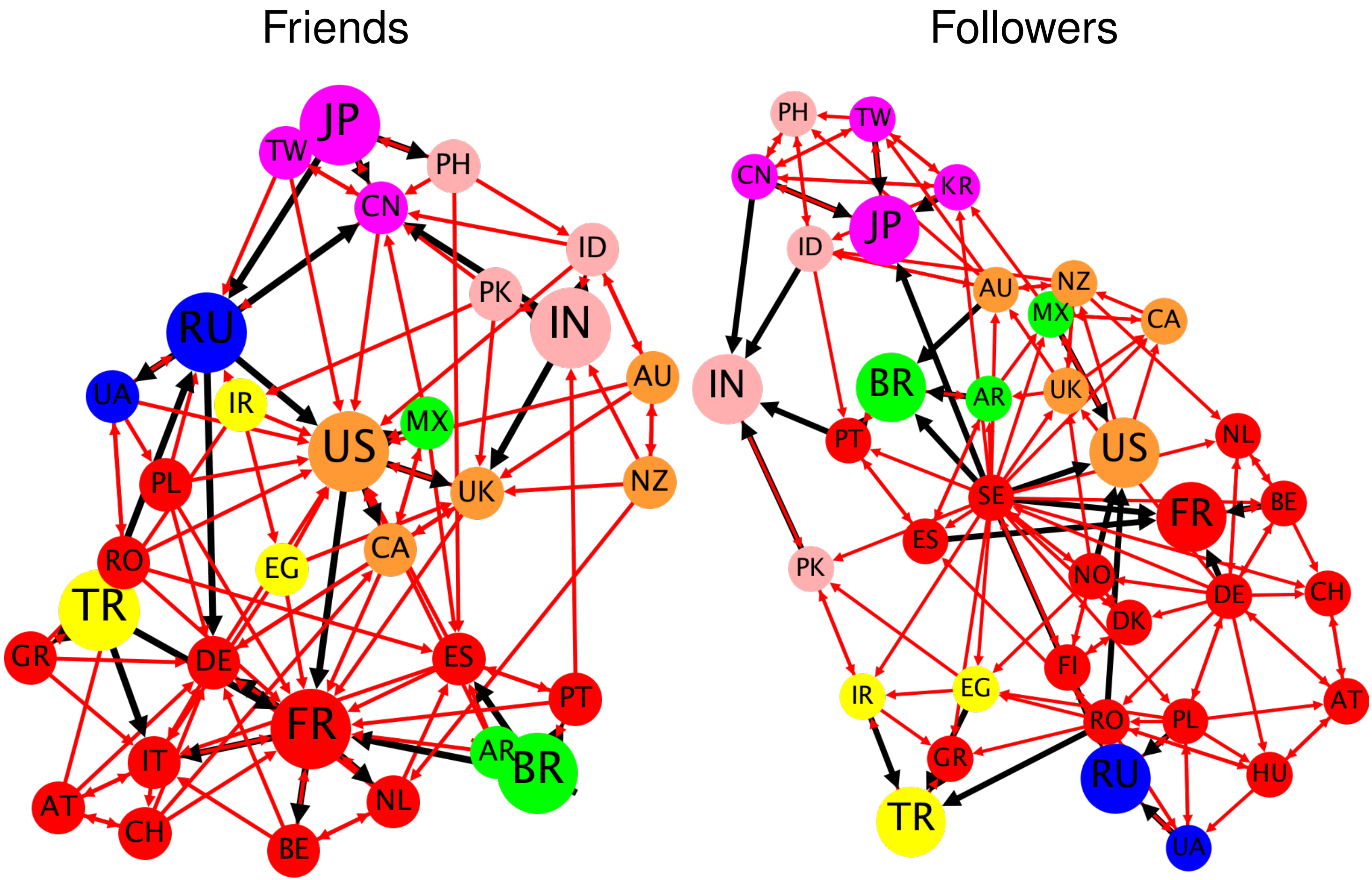
G_{qrnd} Arwiki

G_{qr} Enwiki

Friends | Followers



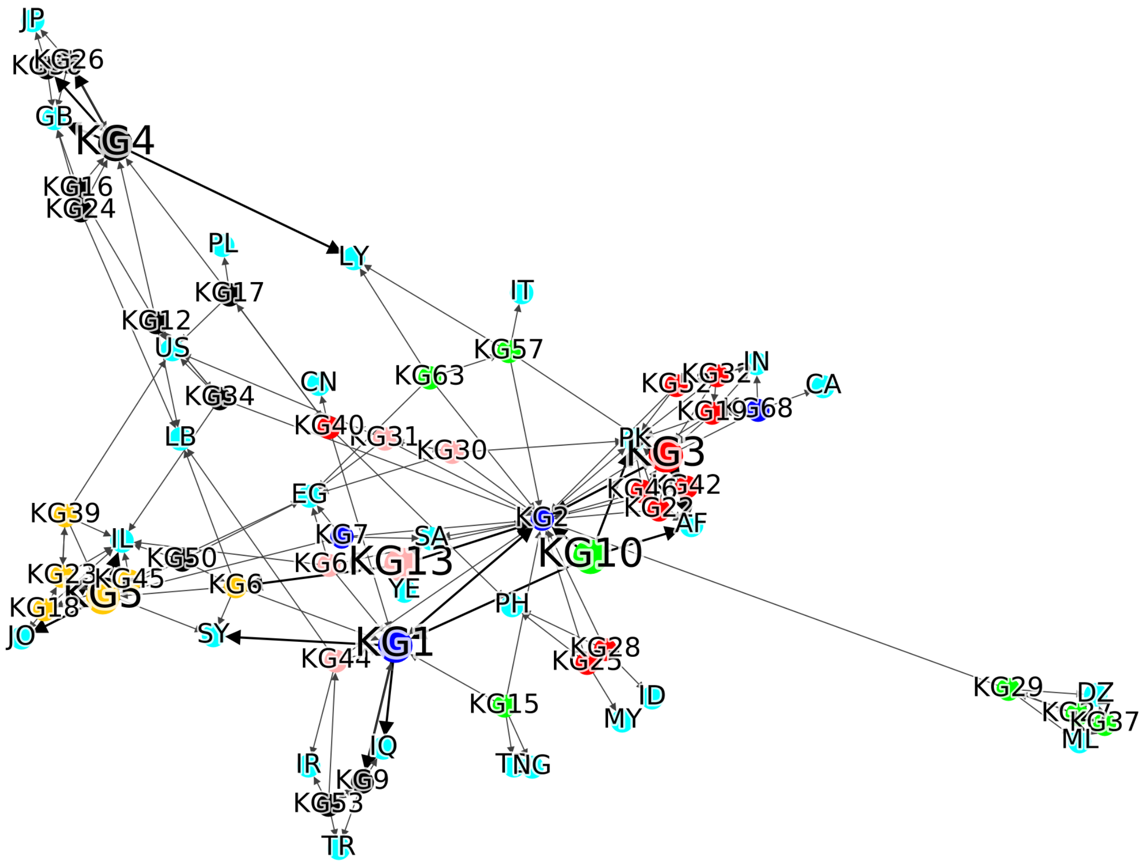
$G_{\rm qr}$ Arwiki



World terror networks

(*Zant, KF, Jaffrès-Runser, Shepelyansky, to appear in EPJB, arXiv:1710.03504*)

"Friend" network: $G_{\mathrm{qr}} + G_{rr}$, Countries and terror groups, Enwiki 2017



Random Perron-Frobenius matrices

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

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:

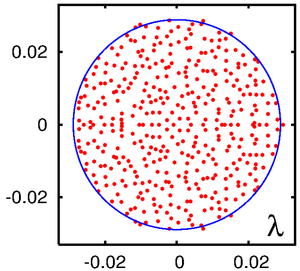
full $\quad \Rightarrow \quad R = 1/\sqrt{3N}$

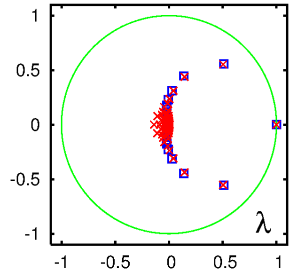
sparse with Q non-zero elements per column $\quad \Rightarrow \quad R \sim 1/\sqrt{Q}$

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

Numerical verification:

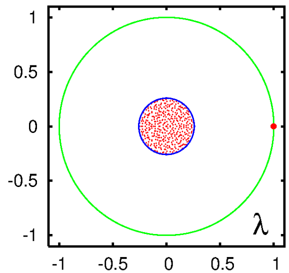
uniform full:
$N = 400$

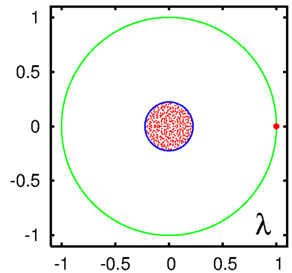




triangular random and average

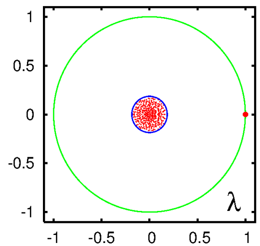
uniform sparse:
$N = 400,$
$Q = 20$

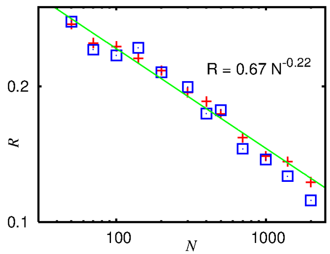




constant sparse:
$N = 400,$
$Q = 20$

power law:
$b = 2.5$





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

Conclusions

- New mathematical method to construct a reduced Google matrix $G_{\mathbf{R}}$ for sub-networks of Wikipedia etc.

- The need for an efficient numerical algorithm provides a decomposition of $G_{\mathbf{R}}$ in three contributions which are also important for the interpretation.

- Appearance of hidden/indirect links providing a friend/follower network structure which may be quite different for $G_{\mathbf{R}}$ or $G_{\mathbf{qr}}$.

- Only the link information of Wikipedia articles is used and not the details of their content ($\Rightarrow$ future approaches with links supporting positive, neutral or negative attitudes between nodes).

- Different language editions of Wikipedia allow to take into account multi-cultural aspects when constructing the reduced network structure.