EPJ B
Condensed Matter
and Complex Systems

EPJ .org
your physics journal

Eur. Phys. J. B (2013) 86: 193 DOI: 10.1140/epjb/e2013-31090-8

Spectral properties of Google matrix of Wikipedia and other networks

Leonardo Ermann, Klaus M. Frahm and Dima L. Shepelyansky

edp sciences
Società Italiana di Fisica
Springer

Eur. Phys. J. B (2013) 86: 193
DOI: 10.1140/epjb/e2013-31090-8

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Spectral properties of Google matrix of Wikipedia and other networks

Leonardo Ermann[1,2], Klaus M. Frahm[2], and Dima L. Shepelyansky[2,a]

[1] Departamento de Física Teórica, GIyA, Comisión Nacional de Energía Atómica, 1429 Buenos Aires, Argentina
[2] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

Received 5 December 2012
Published online 29 April 2013 –

Abstract. We study the properties of eigenvalues and eigenvectors of the Google matrix of the Wikipedia articles hyperlink network and other real networks. With the help of the Arnoldi method, we analyze the distribution of eigenvalues in the complex plane and show that eigenstates with significant eigenvalue modulus are located on well defined network communities. We also show that the correlator between PageRank and CheiRank vectors distinguishes different organizations of information flow on BBC and Le Monde web sites.

1 Introduction

With the appearance of the world wide web (WWW) [1] the modern society created huge directed networks where the information retrieval and ranking of network nodes becomes a formidable challenge. The mathematical grounds of ranking of nodes are based one the concept of Markov chains [2] and related class of Perron-Frobenius operators naturally appearing in dynamical systems (see, e.g., [3]). A concrete implementation of these mathematical concepts to the ranking of WWW nodes was started by Brin and Page in 1998 [4]. It is significantly based on the PageRank algorithm (PRA) which became a fundamental element of the Google search engine broadly used by internet users [5].

Already in 1998, Brin and Page pointed out that *"despite the importance of large-scale search engines on the web, very little academic research has been done on them"* [4]. Since that time the academic studies have been concentrated mainly on the properties of the PageRank vector determined by the PRA (see, e.g., [5–8]). Of course, the PageRank vector is at the basis of ranking of network nodes but the whole description of a directed network is given by the Google matrix G. Thus, it is important to understand the properties of the whole spectrum of eigenvalues of Google matrix and to analyze the meaning and significance of its eigenstates. Certain spectral properties of G matrix have been analyzed in references [9–15]. Here, we concentrate our spectral analysis on the Wikipedia articles network studied in reference [16]. The advantage of this network is due to a clear meaning of nodes, determined by the titles of Wikipedia articles thus simplifying the understanding of information flow in this network.

[a] e-mail: dima@irsamc.ups-tlse.fr

In addition to that, we analyze the statistical properties of eigenvalues and eigenstates of G for WWW networks of Cambridge University, Python, BBC and Le Monde crawled in March 2011.

The Google matrix elements of a directed network are defined as [4,5,17]:

$$G_{ij} = \alpha S_{ij} + (1-\alpha)/N, \qquad (1)$$

where the matrix S_{ij} is obtained from an adjacency matrix A_{ij} by normalizing all nonzero columns to one ($\sum_i S_{ij} = 1$) and replacing columns with only zero elements by $1/N$ (*dangling nodes*) with N being the matrix size. For the WWW an element A_{ij} of the adjacency matrix is equal to unity if a node j points to the node i and zero otherwise. The damping parameter α in the WWW context describes the probability $(1-\alpha)$ to jump to any node for a random surfer. For WWW, the Google search engine uses $\alpha \approx 0.85$ [5]. The matrix G belongs to the class of Perron-Frobenius operators [5], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \leq \alpha$. The right eigenvector at $\lambda = 1$, which is called the PageRank, has real nonnegative elements $P(i)$ and gives a probability $P(i)$ to find a random surfer at site i. Due to the gap $1-\alpha \approx 0.15$ between the largest eigenvalue and the other eigenvalues the PRA permits an efficient and simple determination of the PageRank by the power iteration method. Note that at $\alpha = 1$ the largest eigenvalue $\lambda = 1$ is typically highly degenerate due to many invariant subspaces which define many independent Perron-Frobenius operators which provide (at least) one eigenvalue $\lambda = 1$. This point and also a numerical method to determine the PageRank for the case $1-\alpha \ll 1$ are described in detail in reference [13].

Once the PageRank (at $\alpha = 0.85$) is found, all nodes can be sorted by decreasing probabilities $P(i)$. The node

Table 1. Parameters of all networks considered in the paper.

	N	N_ℓ	n_A
Wikipedia	3282257	71012307	3000
Cam. 2011	893176	15106706	4000
Python	541545	9031262	5000
BBC	319637	7278258	4000
Le Monde	134196	10621445	5000

rank is then given by index $K(i)$ which reflects the relevance of the node i. The top PageRank nodes are located at small values of $K(i) = 1, 2, \ldots$

In addition to a given directed network A_{ij}, it is useful to analyze an inverse network with inverted direction of links with elements of adjacency matrix $A_{ij} \to A_{ji}$. The Google matrix G^* of the inverse network is then constructed via corresponding matrix S^* according to the relations (1) using the same value of α as for the G matrix. The right eigenvector of G^* at eigenvalue $\lambda = 1$ is called CheiRank giving a complementary rank index $K^*(i)$ of network nodes [15,16,18–20]. It is known that the PageRank probability is proportional to the number of ingoing links characterizing how popular or known a given node is while the CheiRank probability is proportional to the number of outgoing links highlighting the node communicativity (see, e.g., [5–8,16,19]). The statistical properties of the node distribution on the PageRank-CheiRank plane are described in reference [19] for various directed networks.

The paper is composed as following: the spectrum of the Google matrix of various networks is analyzed in Section 2, statistical properties of eigenstates are discussed in Section 3, the communities related to Wikipedia eigenstates are examined in Section 4, the distribution of nodes in the PageRank-CheiRank plane is studied in Section 5, the link distribution over PageRank index is considered in Section 6, discussion of results is given in Section 7. An Appendix gives all parameters of the five directed networks considered here and describes in detail certain eigenvalues and eigenvectors.

2 Google matrix spectrum

We study the spectrum of eigenvalues of the Google matrix of five directed networks. For each network the number of nodes N and the number of links N_ℓ are given in Table 1 (see Appendix). The spectrum is obtained numerically using the powerful Arnoldi method described in [21–23]. The idea of the method is to construct a set of orthonormal vectors by applying the matrix (G, S, G^*, S^* or any other matrix of which we want to determine the largest eigenvalues) on some suitable normalized initial vector and orthonormalizing the result to the initial vector. Then the matrix is applied to the second vector and the result is orthonormalized to the first two vectors and so on. The used scalar products and normalization factors during the Gram-Schmidt process provide the matrix representation of the initial big matrix on the set of orthonormal vectors (which span a *Krylov space*) in a form of a Hessenberg matrix whose eigenvalues converge typically quite well versus the largest eigenvalues of the initial matrix even if the chosen number of orthonormal vectors, the Arnold dimension n_A, is quite modest (3000–5000 in this work) as compared to the initial matrix size.

Table 2. G and G^* eigespectrum parameters for all networks.

	N_s	N_d	$d_{\max}$	$N_{\rm circ.}$	N_1
Wikipedia	515	255	11	381	255
Wikipedia*	21198	5355	717	8968	5365
Cam. 2011	808	329	74	343	332
Cam. 2011*	186062	2039	5144	2044	2041
Python	198	23	72	26	23
Python*	1589	25	951	35	31
BBC	50	19	28	19	19
BBC*	39	28	6	28	28
Le Monde	83	64	18	64	64
Le Monde*	789	354	15	373	361

In this work, we are interested in the spectrum of the matrix $S = G(\alpha = 1)$ (or S^*) since the spectrum of $G(\alpha)$ (or $G^*(\alpha)$) is simply obtained by rescaling the complex eigenvalues with the factor α (apart from "one" largest eigenvalue $\lambda = 1$ which does not change).

The direct dionalization of the Google matrix G faces a number of numerical challenges. Thus, the highly degenerate unit eigenvalue $\lambda = 1$ of S creates convergence problems for the Arnoldi method. To resolve this numerical problem, we follow the approach developed in references [13,15] and follow the description given there. We first find the invariant isolated subsets. These subsets are invariant with respect to applications of S. We merge all subspaces with common members, and obtain a sequence of disjoint subspaces V_j of dimension d_j invariant by applications of S. The remaining part of nodes forms the wholly connected *core space*. Such a classification scheme can be efficiently implemented in a computer program and it provides a subdivision of network nodes in N_c core space nodes and N_s subspace nodes belonging to at least one of the invariant subspaces V_j inducing the block triangular structure of matrix S:

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix}, \quad (2)$$

where S_{ss} is itself composed of many small diagonal blocks for each invariant subspace and whose eigenvalues can be efficiently obtained by direct ("exact") numerical diagonalization.

The total subspace size N_S, the number of independent subspaces N_d, the maximal subspace dimension $d_{\max}$ and the number N_1 of S eigenvalues with $\lambda = 1$ are given in Table 2. The spectrum and eigenstates of the core space S_{cc} are determined by the Arnoldi method with Arnoldi dimension n_A giving the eigenvalues λ_i of S_{cc} with largest modulus and the corresponding eigenvectors ψ_j ($G\psi_i = \lambda_i\psi_i$). The values of n_A we used for the different networks are given in Table 1. According to Table 2, we have the average number of links per node $\zeta_\ell \approx 21.63$ (Wikipedia),

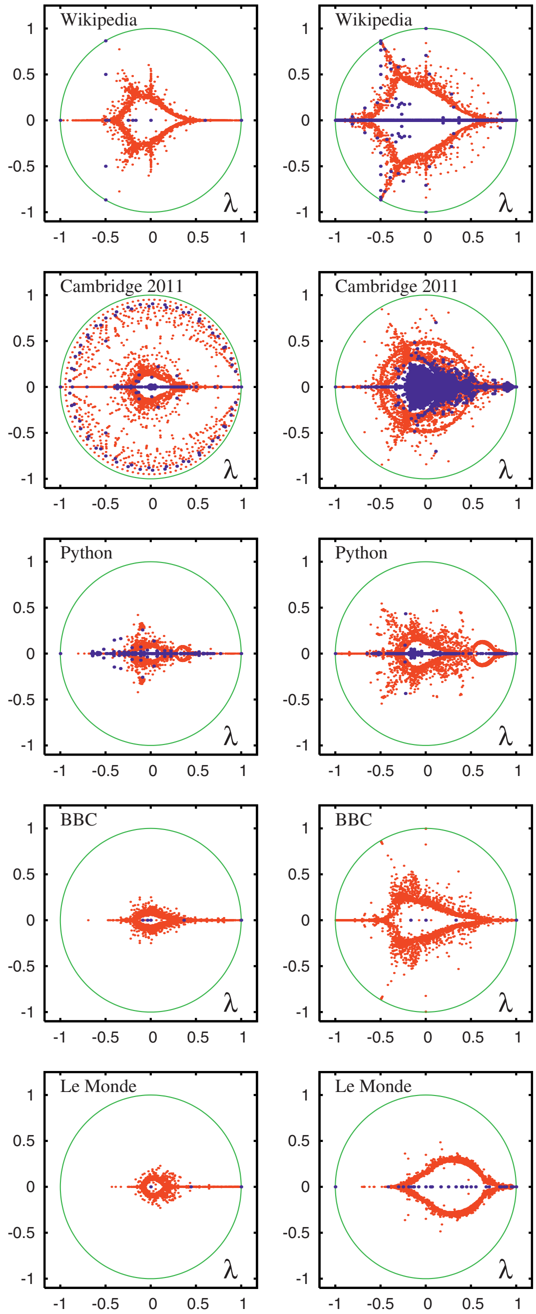

Fig. 1. Spectrum of eigenvalues λ the Google matrices G (left column) and G^* (right column) for Wikipedia, Cambridge 2011, Python, BBC and Le Monde ($\alpha = 1$). Red dots are core space eigenvalues, blue dots are subspace eigenvalues and the full green curve shows the unit circle. The core space eigenvalues were calculated by the projected Arnoldi method with Arnoldi dimensions n_A as given in Table 1.

16.91 (Cambridge 2011), 16.67 (Python), 22.77 (BBC), 79.14 (Le Monde).

The distributions of subspaces eigenvalues and largest n_A eigenvalues of the core space are shown in Figure 1 in the complex plane λ for all five networks. The blue points show the eigenvalues of isolated subspaces. We note that their number is relatively small compared to those of British University networks [24] (up to year 2006) analyzed in reference [13]. We attribute this to a larger number of ζ_ℓ links per node that reduces an effective size of isolated parts of network. Between 2006 and 2011, especially for Cambridge, it seems that the increased use of PHP and similar web software tends to considerably increase the value of ζ_ℓ. Indeed, we have $\zeta_\ell \approx 10$ for university networks up to 2006 [13] which used less this kind of PHP software. In Figure 1 the red points show n_A eigenvalues of the core space with largest $|\lambda|$. Due to finite n_A value there is an empty white space around $\lambda = 0$. There is no significant gap for core eigenvalues since λ_1 is rather close to 1 (see Tab. 3).

In global, we can say that the structure of the Wikipedia spectrum of S and S^* is somewhat similar to

Table 3. Eigenvalues of eigenvectors shown in Figures 1 and 2 by corresponding colors. Index m of λ_m numbers eigenvalues in the decreasing order of $|\lambda|$ in the core space.

	Color	Eigenvalue
Wikipedia	red	$\lambda_1 = 0.999987$
	green	$\lambda_2 = 0.977237$
	blue	$\lambda_{52} = -0.35003 + i\,0.77374$
	pink	$\lambda_{864} = -0.34293 + i\,0.43145$
Wikipedia*	red	$\lambda_1 = 0.999982$
	green	$\lambda_2 = 0.999902$
	blue	$\lambda_{662} = 0.0000000 + i\,0.84090$
	pink	$\lambda_{38} = -0.49626 + i\,0.85653$
Cam. 2011	red	$\lambda_1 = 0.999749$
	green	$\lambda_2 = 0.999270$
	blue	$\lambda_{350} = 0.41779 + i\,0.77856$
	pink	$\lambda_{144} = -0.52909 + i\,0.78693$
Cam. 2011*	red	$\lambda_1 = 0.999998$
	green	$\lambda_2 = 0.999994$
	blue	$\lambda_{765} = 0.24846 + i\,0.80915$
	pink	$\lambda_{249} = -0.48736 + i\,0.84568$
Python	red	$\lambda_1 = 0.999975$
	green	$\lambda_2 = 0.998864$
	blue	$\lambda_{3315} = 0.14484 + i\,0.19215$
	pink	$\lambda_{1337} = -0.14427 + i\,0.42051$
Python*	red	$\lambda_1 = 0.999995$
	green	$\lambda_2 = 0.999991$
	blue	$\lambda_{2559} = 0.37694 + i\,0.45231$
	pink	$\lambda_{3076} = 0.12214 + i\,0.47416$
BBC	red	$\lambda_1 = 0.99883$
	green	$\lambda_2 = 0.99251$
	blue	$\lambda_{1276} = -0.12414 + i\,0.24795$
	pink	$\lambda_{1148} = -0.22459 + i\,0.20024$
BBC*	red	$\lambda_1 = 0.999999$
	green	$\lambda_2 = 0.999994$
	blue	$\lambda_{16} = -0.00067 + i\,0.99930$
	pink	$\lambda_{90} = -0.49635 + i\,0.85848$
Le Monde	red	$\lambda_1 = 0.998837$
	green	$\lambda_2 = 0.983123$
	blue	$\lambda_{926} = 0.10295 + i\,0.22890$
	pink	$\lambda_{1118} = 0.08023 + i\,0.20595$
Le Monde*	red	$\lambda_1 = 0.999999$
	green	$\lambda_2 = 0.999959$
	blue	$\lambda_{2093} = 0.15987 + i\,0.48502$
	pink	$\lambda_{2474} = 0.17637 + i\,0.40917$

those of Cambridge 2006 (see Fig. 2 in Ref. [13]). For Cambridge 2011, the spectrum of S is drastically changed compared to the year 2006 but for S^* certain features remain common both for 2006 and 2011 (e.g., a circle $|\lambda| \approx 0.5$, triplet-star). For Python, BBC and Le Monde the imaginary parts $\mathrm{Im}(\lambda)$ of eigenvalues of S are relatively small compared to the networks of Wikipedia and Cambridge. We suppose that there are less symmetric links in the later cases. It is interesting that for S^* of Python, BBC and Le Monde the imaginary parts $\mathrm{Im}(\lambda)$ are significantly larger than for S.

The origin of nontrivial structures of the spectrum of G and G^* for directed networks discussed here and in references [11–13,15] still require detailed analysis. We note that well visible triplet and cross structures (see, e.g., Wikipedia spectrum in Fig. 1 and Fig. 2 of [13]) naturally appear in the spectra of random unistochastic matrices of size $N = 3$ and 4, which have been analyzed analytically and numerically in reference [25]. In view of this similarity, we suppose that networks with such structures have some triplet or quartet subgroup of nodes weakly coupled to the rest of the network. However, a detailed understanding of the spectrum requires a deeper analysis. In the next section, we turn to a study of eigenstate properties.

3 Statistical properties of eigenstates

The dependence of PageRank P and CheiRank P^* vectors on their indexes K and K^* at $\alpha = 0.85; 1 - 10^{-8}$ are shown in Figure 2. At $\alpha = 0.85$, we have an approximate algebraic decay of probability according to the Zipf law $P \sim 1/K^{\beta}, P^* \sim 1/K^{*\beta}$ (see, e.g., [14] and references therein). We find the following values β for PageRank (CheiRank): $0.96 \pm 0.002\,(0.73 \pm 0.003)$ Wikipedia; $0.81 \pm 0.007\,(0.90 \pm 0.004)$ Cambridge 2011; $1.12 \pm 0.01\,(1.17 \pm 0.006)$ Python; $1.20 \pm 0.006\,(0.96 \pm 0.004)$ BBC; $1.08 \pm 0.009\,(0.55 \pm 0.002)$ Le Monde. Formally, the statistical errors in β are relatively small but in some cases there are variations of slope in the decay of PageRank (CheiRank) probability that gives a dependence of β on a fitting range (e.g., that is why β here is a bit different from its values for Wikipedia given in Ref. [16]). We note that the value $\beta \approx 1$ for the PageRank remains relatively stable to all networks corresponding to the usual exponent $\mu \approx 2.1$ of algebraic decay of the ingoing link distribution leading to $\beta = 1/(\mu - 1) \approx 0.9$ (see, e.g., [6,7,14–16]).

For CheiRank the variations of β from one network to another are more significant being in agreement with the fact that for outgoing links the exponent $\mu \approx 2.7$ varies in a more significant manner.

For $\alpha = 1 - 10^{-8}$, we find that the main probability of PageRank and CheiRank eigenvectors is located on isolated subspaces with N_s nodes; after that value there is a significant drop of probability for $K, K^* > N_s$. This effect was already found and explained in detail in reference [13] and our new data confirm that it is indeed rather generic.

The modulus of four eigenfuctions $|\psi_i(j)|$ from the core space are shown in Figure 2 by color curves as a function of their own index K_i which order $|\psi_i(j)|$ in a monotonic decreasing order. For Python, BBC and Le Monde the decay of $|\psi_i(j)|$ with K_i is similar to the decay

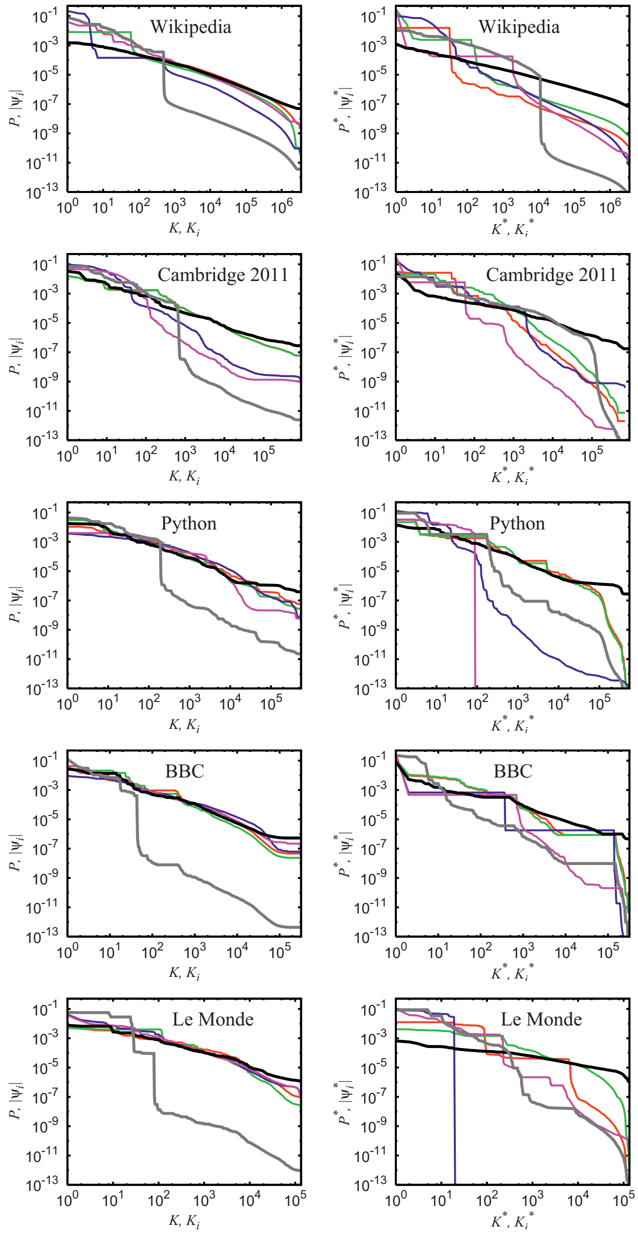

Fig. 2. PageRank P (left column) and CheiRank P^* (right column) vectors are shown as a function of the corresponding rank indexes K or K^* for the Google matrices of Wikipedia, Cambridge 2011, Python, BBC and Le Monde at the damping parameter $\alpha = 0.85$ (thick black curve) and $\alpha = 1 - 10^{-8}$ (thick gray curve). The thin color curves show for each panel the modulus of four core space eigenvectors $|\psi_i|$ of S (left column) and $|\psi_i^*|$ of S^* (right column) versus their ranking indexes K_i or K_i^*. Red and green curves are the eigenvectors corresponding to the two largest core space eigenvalues (in modulus) which are real and close to 1; blue and pink curves are the eigenvectors corresponding to two complex eigenvalues with large imaginary part. The chosen eigenvalues and other relevant quantities for each case are listed in Tables 1–3.

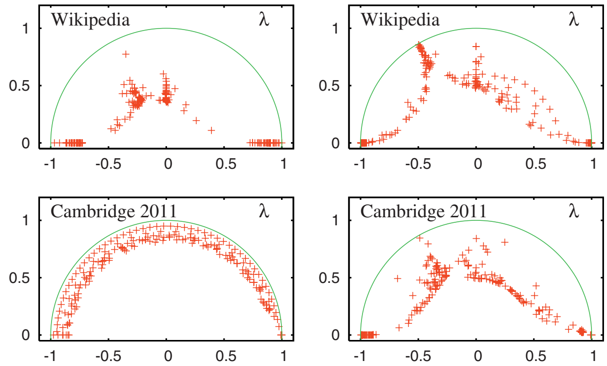

Fig. 3. A selection of 200 complex core space eigenvalues closest to the unit circle for the matrices S (left column) and S^* (right column) of Wikipedia and Cambridge 2011 networks. The characteristics of corresponding eigenvectors are shown in Figures 4 and 5.

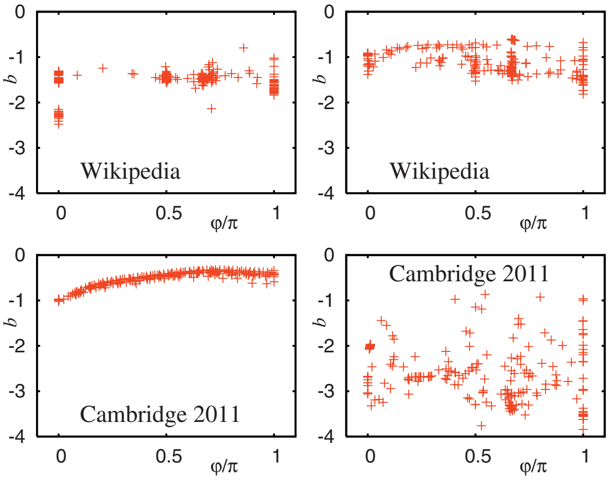

Fig. 4. Left column: algebraic exponent b obtained from a power law fit $|\psi_i(K_i)| \sim K_i^b$ for $K_i \geq 10^4$ shown as a function of the phase $\varphi = \arg(\lambda_i)$ of the complex eigenvalue λ_i associated to the eigenvector ψ_i of S. The shown data points correspond to the eigenvalue selection of Figure 3 for networks of Wikipedia and Cambridge 2011. Right column: the same as in the left column for the eigenvectors of S^*.

of PageRank probability with K. For Wikipedia and Cambridge 2011 we see that eigenvectors $|\psi_i(j)|$ are more localized. The eigenstates of S^* have a significantly more irregular decay compared to the eigenstates of S.

To analyze the properties of core eigenstates of Wikipedia and Cambridge 2011 in a better way, we select 200 core space eigenvalues of S and S^* being closest to the unitary circle $|\lambda| = 1$. These eigenvalues are shown in Figure 3. For these eigenvalues, we compute the corresponding eigenvectors $\psi_i(j)$ and by fitting a power law dependence $|\psi_i(K_i)| \sim K_i^b$ at $K_i \geq 10^4$ we determine the dependence of the exponent b on the phase of the eigenvalue $\varphi = \arg(\lambda_i)$. For Wikipedia, we have values of $|b|$ distributed mainly in the range (1–2) for S and in the range (0.5–1.5) for S^*. For Cambridge 2011, we have a more compact range (0.5–1) for S while for S^* there is a very broad variation of $|b|$ values in the range (1–4).

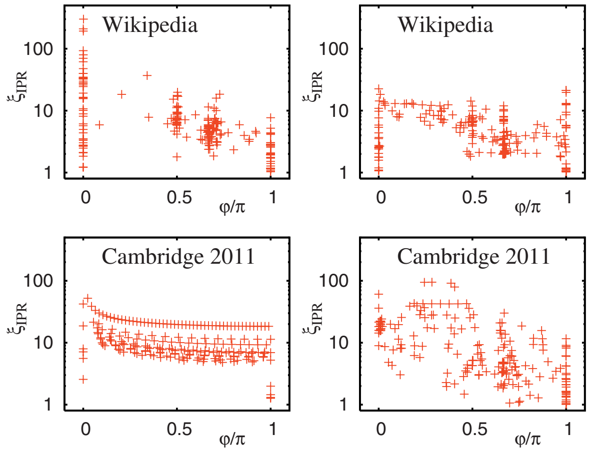

Fig. 5. Left column: inverse participation ratio $\xi_{\mathrm{IPR}} = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$ shown as a function of the phase $\varphi = \arg(\lambda_i)$ of the complex eigenvalue λ_i associated to the eigenvector ψ_i of S. The data points correspond to the eigenvalue selection of Figure 3 for networks of Wikipedia and Cambridge 2011. Right column: the same as in the left column for the eigenvectors of S^*.

The above approximate power law description of the eigenstate decay characterizes their behavior at large K values. The behavior at low K values can be characterized by the inverse participation ratio (IPR) $\xi_{\mathrm{IPR}} = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$, which gives an approximate number of nodes on which the main probability of an eigenstate $\psi_i(j)$ is located. We note that such a characteristic is broadly used in disordered mesoscopic systems allowing to detect the Anderson transition from localized phase with finite ξ to delocalized phase with ξ value comparable with the system size [26]. The IPR data are presented in Figure 5 for eigenvalues selection of Figure 3. We find that ξ_{IPR} values are by a factor 10^4 to 10^5 smaller than the network size N. This means that these eigenstates are well localized on a restricted number of nodes. We try to analyze what are these nodes in next section for the example of Wikipedia where the meaning of a node is clearly defined by the title of the corresponding Wikipedia article.

4 Communities of Wikipedia eigenstates

To understand the meaning of other eigenstates in the core space we order selected eigenstates by their decreasing value $|\psi_i(j)|$ and apply a frequency analysis on the first 1000 articles with $K_i \leq 1000$. The mostly frequent word of a given eigenvector is used to label the eigenvector name. These labels with corresponding eigenvalues are shown in Figure 6 in λ-plane. We identify four main categories for the selected eigenvectors shown by different colors in Figure 6: countries (red), biology and medicine (orange), mathematics (blue) and others (green). The category of others contains rather diverse articles

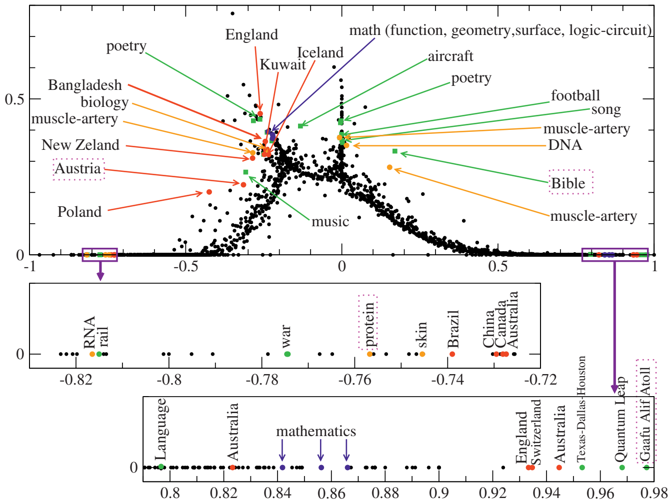

Fig. 6. Complex eigenvalue spectrum of the matrices S for Wikipedia. Highlighted eigenvalues represent different communities of Wikipedia and are labeled by the most repeated and important words following word counting of first 1000 nodes. Color are used in the following way: red for countries, orange for biology, blue for mathematics and green for others. Top panel shows complex plane for positive imaginary part of eigenvalues, while middle and bottom panels focus in the negative and positive real parts. Top 20 nodes with largest values of eigenstates $|\psi_i|$ and their eigenvalues λ_i are given in Tables 4–7 (4 names marked by dotted boxes in figure panels).

about poetry, Bible, football, music, American TV series (e.g., Quantum Leap), small geographical places (e.g., Gaafru Alif Atoll). Clearly these eigenstates select certain specific communities which are relatively weakly coupled with the main bulk part of Wikipedia that generates relatively large modulus of $|\lambda_i|$. The top 20 articles of eigenstate PageRank index K_i are listed in Tables 4–7.

The eigenvector of Table 4 has a positive real λ and is linked to the main article *Gaafu Alif Atoll* which in its turn is linked mainly to atolls in this region. Clearly this case represents well localized community of articles mainly linked between themselves that gives slow relaxation rate of this eigenmode with $\lambda = 0.9772$ being rather close to unity.

In Table 5, we have an eigenvector with real negative eigenvalue $\lambda = -0.8165$ with the top node *Photoactivatable fluorescent protein*. This node is linked to *Kaede (protein)* and *Eos (protein)* with the later being isolated from coral. Its picture is listed in *Portal:Berkshire/Selected picture* which has pictures of *St Paul's Cathedral* and *Legoland Windsor* that generates appearance of these, on a first glance unrelated articles, to be present in this eigenvector. Thus, this eigenvector also highlights a specific community which is somewhat stronger coupled to the global Wikipedia core, due to a link to selected pictures, with a smaller modulus of λ compared to the case of Table 4.

The eigenvector of Table 6 has a complex eigenvalue with $|\lambda| = 0.3733$ and the top article *Portal:Bible*. The top three articles of this eigenvector have very close values of $|\psi_i(j)|$ that seems to be the reason why we have

Table 4. Node rank for decreasing modulus of eigenstate $|\psi_i|$ corresponding to the eigenvalue $\lambda_2 = 0.97724$ (see Fig. 6).

	$\lambda_2 = 0.9772$ ("Gaafu Alif Atol")	$\lvert\psi_i\rvert$
1	Gaafu Alif Atoll	0.00816
2	Kureddhoo (Gaafu Alif Atoll)	0.00812
3	Hithaadhoo (Gaafu Alif Atoll)	0.00808
4	Dhigurah (Gaafu Alif Atoll)	0.00806
5	Maarandhoo (Gaafu Alif Atoll)	0.00806
6	Hulhimendhoo (Gaafu Alif Atoll)	0.00805
7	Araigaiththaa	0.00798
8	Baavandhoo	0.00798
9	Baberaahuttaa	0.00798
10	Bakeiththaa	0.00798
11	Beyruhuttaa	0.00798
12	Beyrumaddoo	0.00798
13	Boaddoo	0.00798
14	Budhiyahuttaa	0.00798
15	Dhevvalaabadhoo	0.00798
16	Dhevvamaagalaa	0.00798
17	Dhigudhoo	0.00798
18	Dhonhuseenahuttaa	0.00798
19	Falhumaafushi	0.00798
20	Falhuverrehaa	0.00798

Table 5. Node rank for decreasing modulus of eigenstate $|\psi_i|$ corresponding to the eigenvalue $\lambda_{80} = -0.8165$ (see Fig. 6).

	$\lambda_{80} = -0.8165$ ("protein")	$\lvert\psi_i\rvert$
1	Photoactivatable fluorescent protein	0.22767
2	Kaede (protein)	0.13942
3	Eos (protein)	0.13942
4	Fusion protein	0.05946
5	Green fluorescent protein	0.05723
6	Portal:Berkshire/Selected picture	0.01019
7	Persistent tunica vasculosa lentis	0.00552
8	Portal:Berkshire/Selected picture/Layout	0.00416
9	Portal:Berkshire/Selected picture/1	0.00416
10	Portal:Berkshire/Nominate/ Selected picture	0.00416
11	Persistent hyperplastic primary vitreous	0.00338
12	Tunica vasculosa lentis	0.00338
13	Tpr-met fusion protein	0.00319
14	St Paul's Cathedral	0.00256
15	Legoland Windsor	0.00255
16	Complementary DNA	0.00252
17	Gené	0.00221
18	Gene	0.00215
19	Gag-onc fusion protein	0.00181
20	Protein	0.00177

$\varphi = \arg(\lambda_i) = \pi \times 0.3496$ being very close to $\pi/3$. The Bible is strongly linked to various aspects of human society that leads to a relatively small modulus value of this well defined community.

In Table 7, we have an eigenvector which starts from the article *Lower Austria* with the eigenvalue modulus $|\lambda| = 0.3869$. This article is linked to such articles as *Austria* and *Upper Austria* with historical links to *Styria*. It also links to its city capital *Krems an der Donau*. The articles *World War II* and *Jew* appear due to a sentence

Table 6. Node rank for decreasing modulus of eigenstate $|\psi_i|$ corresponding to the eigenvalue $\lambda_{1481} = 0.1699 + i0.3325$ (see Fig. 6).

	$\lambda_{1481} = 0.1699 + i0.3325$ ("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

Table 7. Node rank for decreasing modulus of eigenstate $|\psi_i|$ corresponding to the eigenvalue $\lambda_{1395} = -0.3149+i0.2248$ (see Fig. 6).

	$\lambda_{1395} = -0.3149 + i0.2248$ ("Austria")	$\lvert\psi_i\rvert$
1	Lower Austria	0.04284
2	Austria	0.03112
3	Upper Austria	0.00817
4	Styria	0.00781
5	Burgenland	0.00307
6	World War II	0.00304
7	Krems an der Donau	0.00282
8	Jew	0.00272
9	Slovakia	0.00268
10	Bruck an der Leitha (district)	0.00265
11	History of Austria	0.00263
12	Wiener Neustadt	0.00260
13	Mostviertel	0.00251
14	States of Austria	0.00250
15	Waidhofen an der Ybbs	0.00249
16	MELK	0.00246
17	Melk	0.00246
18	Bundesland (Austria)	0.00239
19	Wachau	0.00233
20	Waldviertel	0.00226

"Before World War II, Lower Austria had the largest number of Jews in Austria". Due to links with very popular nodes the eigenvector of this community has a relative small modulus of λ.

Let us make here a few additional remarks about other eigenvectors. For example, we analyzed the meaning of eigenvector with $\lambda = -0.3500 + i0.7737 = |\lambda| \exp(i\theta)$ (located slightly above the word *England* in Fig. 6). Its top five amplitude modulus are *Screen Producers Association of Australia*, *Screen Producers Association of Australia (SPAA)*, *SPAA Conference*, *SPAA Fringe*, *Sydney*. This clearly shows that this vector selects a certain community of Australian Screen Producers. It is interesting to note that we have here $\theta = 114°$ being close to the angle $2\pi/3$ corresponding to 1/3 resonance rotations mainly between first three top nodes.

In fact, there are other eigenvalues which have θ being close to resonance values with $\theta/2\pi = 1/3, 1/4\ldots$ Thus, the eigenvector *England* has $\lambda = -0.2613 + i0.4527$ with $\theta = 120°$ corresponding to the resonance rotation between three nodes. Indeed, the top amplitudes of this eigenvector have titles *Charles William Hempel*, *Charles Frederick Hempel*, *Carl Frederick Hempel* with strong links between these titles leading to 1/3 rotation (this vector is marked as *England* since this word is the most frequent among top 1000 titles).

There are other eigenvalues close to 1/3 resonance rotation. Thus, we have $\lambda = -0.2621 + i0.4346$ with $\theta = 121°$ marked as *poetry* in Figure 6. This eigenvector has top amplitude modulus: *Poetry* (0.0622), *Portal:Poetry/poem archive* (0.03339), *Portal:Poetry/poem archive/2006 archive* (0.03289), *Portal:Poetry* (0.03180), *Walter Raleigh* (0.0064). We think that the top nodes 2, 3, 4 have practically the same amplitudes thus corresponding to the resonance 1/3 rotation between these three nodes.

There is also another eigenvector marked *poetry* in Figure 6 with $\lambda = -0.0026 + i0.4297$ and $\theta \approx 90°$. In fact this article speaks about *1000s in poetry* with approximately equal 6 amplitudes about poetry in various years that corresponds to a resonance 1/6 rotation generating $\theta \approx 90°$. There are also other vectors with resonance values $1/2, 1/4, 1/6$ that produce eigenvalues with a dominant imaginary part. We also note that there are other resonance eigenvalues among those given in Table 3 (e.g., λ_{38} with $\theta = 120.1°$). We think that such resonance θ values have close similarity with those of random matrix models of small size $N = 3, 4, 5, 6$ analyzed in reference [25] corresponding to the main part of information exchange between a small number of nodes.

The above analysis shows that the eigenvectors of the Google matrix of Wikipedia clearly identify certain communities which are relatively weakly connected with the Wikipedia core when the modulus of corresponding eigenvalue is close to unity. For moderate values of $|\lambda|$, we still have well defined communities which however have stronger links with some popular articles (e.g., countries) that lead to a more rapid decay of such eigenmodes.

The above results show that the analysis of eigenvectors highlights interesting features of communities and network structure. However, a priori it is not evident what is a correspondence between the numerically obtained eigenvectors and the specific community features in which someone has a specific interest. It is possible that for a well defined community it can be useful to construct a personalized Google matrix (see, e.g., [5]) and to perform analysis of its eigenstates.

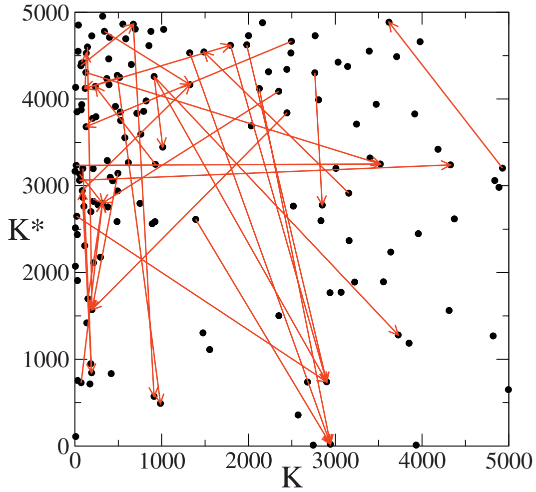

Fig. 7. Top 5000 values in PageRank-CheiRank plane (K, K^*) of Wikipedia. All nodes and all links in this region are shown by black circles and red arrows, respectively.

5 CheiRank versus PageRank plane

As it is discussed in references [15,16,18,19], it is useful to look on the distribution of network nodes on PageRank-CheiRank plane (K, K^*). For Wikipedia a large scale distribution is analyzed in references [16,19] and the networks of British Universities, Linux Kernel and Twitter are considered in references [15,19].

In Figure 7, we show for Wikipedia the distribution of nodes in (K, K^*) plane for a relatively small range of top 5000 values of K, K^*. All directed links in this region are also shown. In fact the number of such links and number of nodes in this region are relatively small. Indeed, a large scale density of nodes (see Fig. 3 in Ref. [16]) shows that the density of nodes is not very high at the top corner of PageRank-CheiRank plane. This happens due to the fact that top nodes of PageRank, whose components are proportional to the number of ingoing links, are usually not those of CheiRank, whose components are proportional to the number to outgoing links.

The correlation between PageRank and CheiRank vectors can be characterized by their correlator [18,19]:

$$\kappa = N \sum_{i=1}^{N} P(K(i))P^*(K^*(i)) - 1. \quad (3)$$

For our networks we find its values to be $\kappa = 4.08$ (Wikipedia), 41.5 (Cambridge 2011), 12.9 (Python), 140.2 (BBC), 0.85 (Le Monde). Except for the case of Le Monde, these values are relatively high showing that there is a significant correlation between PageRank and CheiRank probabilities on corresponding networks. We remind that for Linux Kernel networks the values of κ are close to zero corresponding to absence of correlations there [18,19].

The strong difference between κ values for BBC and Le Monde shows that the structure of these two web sites is very different. To analyze this difference in a better way we show the density of nodes for these two networks on small and large scales in Figure 8. For small scale, shown by top panels, it is clear that the density of nodes is significantly larger for BBC network. However, this difference becomes even more drastic on the large logarithmic scale of the whole network shown in bottom panels. Indeed, on a logarithmic scale we see that BBC network has a square like distribution region with a certain probability maximum around the diagonal $K \approx K^*$ while Le Monde network has a triangular type distribution which is typical for networks without correlations between PageRank and CheiRank vectors, like it is the case for the Linux Kernel networks (see Fig. 4 in Ref. [19]). Indeed, a random procedure of node generation on (K, K^*) plane gives such a triangular distribution without correlations between PageRank and CheiRank nodes (see procedure description and right panel of Fig. 4 in Ref. [16]). This analysis shows that BBC and Le Monde agencies handle information flows on their web sites in a drastically different manner. Thus for the BBC web site the most popular articles are at the same time also the most communicative ones while in contrast to that for the Le Monde web site the most popular and most communicative articles are very different.

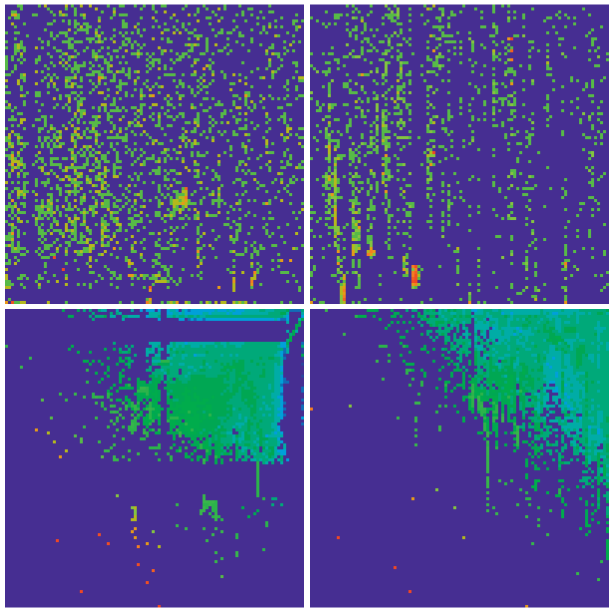

Fig. 8. Density of nodes $W(K, K^*)$ on PageRank-CheiRank plane (K, K^*) for the networks of BBC (left panels) and Le Monde (right panels). Top panels show density in the range $1 \le K, K^* \le 10^4$ with averaging over cells of size 100×100; bottom panels show density averaged over 100×100 logarithmically equidistant grids for $0 \le \ln K, \ln K^* \le \ln N$, the density is averaged over all nodes inside each cell of the grid, the normalization condition is $\sum_{K,K^*} W(K, K^*) = 1$. Color varies from blue at zero value to red at maximal density value. At each panel the x-axis corresponds to K (or $\ln K$ for the bottom panels) and the y-axis to K^* (or $\ln K^*$ for the bottom panels).

6 Links distribution over PageRank nodes

To understand the properties of directional flow on a network it is also useful to analyze the distribution of links

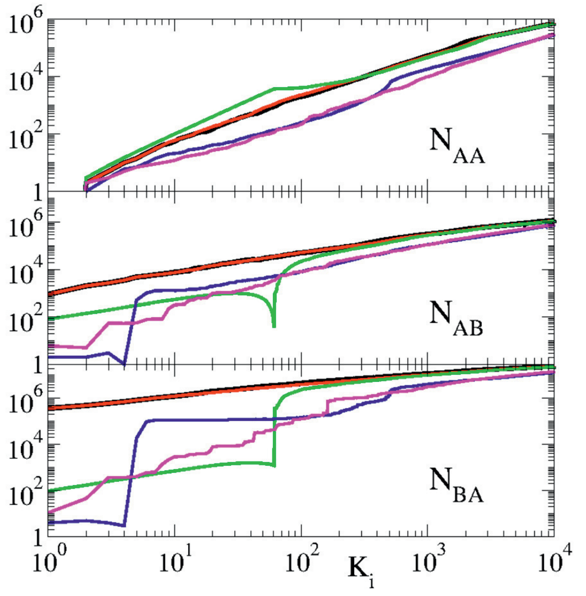

Fig. 9. Number of links between or inside sets A and B defined by the index K_i ordered by decreasing absolute value of Wikipedia eigenstates. The number of links starting and pointing to nodes inside the set A (N_{AA}) is shown in top panel as a function of K_i. The cases of links from set A to set B (N_{AB}) and from B to A (N_{BA}) are shown in middle and bottom panel, respectively. Note that the total number of links is conserved and the quantity N_{BB} can be obtained as $N_{BB} = N_\ell - N_{AA} - N_{AB} - N_{BA}$. The case of PageRank vector with damping parameter $\alpha = 0.85$ is shown by a black curve versus K index. The color curves show the cases of four core space eigenvectors $|\psi_i|$ of S versus their ranking indexes K_i. Red and green curves are the eigenvectors corresponding to the two largest core space eigenvalues (in modulus) being $\lambda_1 = 0.99998702$ and $\lambda_2 = 0.97723699$, respectively; blue and pink curves are the eigenvectors corresponding to two complex eigenvalues with large imaginary part being $\lambda_{52} = -0.35003316 + i0.77373677$ and $\lambda_{864} = -0.34293502 + i0.43144930$, respectively.

over PageRank nodes. We illustrate this approach for the Wikipedia network. Suppose that all nodes are ordered in a decreasing order of modulus of a given eigenvector. For the PageRank vector all nodes are numbered by the PageRank index K, while for a given eigenstate $\psi_i(j)$ all nodes are numbered by a local corresponding index K_i. We now divide all nodes on two parts A and B with $1, \ldots, K_i$ nodes for A and $K_i + 1, \ldots, N$ nodes for B. Then we determine the number of links N_{AA} starting and ending in part A, the number of links N_{AB} pointing from part A to part B and the number of links N_{BA} pointing from part B to part A. The number of links inside part B is then $N_{BB} = N_\ell - N_{AA} - N_{AB} - N_{BA}$. For the PageRank vector, the dependence of N_{AA} on K was analyzed for different networks in reference [15]. Here we generalize this concept to consider links between two parts A, B for various eigenvectors of the Google matrix.

According to the data of Figure 9, we find that for all eigenvectors $N_{AA} \propto K_i^{1.5}$ grows approximately in an algebraic way with the exponent being close to 1.5 being similar to the PageRank case considered in reference [15]. However, the dependence of N_{AB} and N_{BA} on K_i is rather different for different eigenstates. For the PageRank and the λ_1 eigenvector, we find practically the same behavior linked to the fact that at $\alpha = 0.85$, the PageRank vector is rather close to the first core space eigenvector (see discussion in Ref. [13]). Here, the interesting point is that at small values of K_i we have N_{BA} being larger than N_{AB} almost by a factor 100. This is due to the fact that low rank nodes at large K_i point preferentially to high rank nodes at low K_i. For other three eigenvectors with λ_2, λ_{52}, λ_{864}, we find well pronounced step-like behavior of N_{AB}, N_{BA} on K_i. We argue that the step size in K_i is given by the size of a community which has preferential links mainly inside the community. Indeed, for the eigenvector of λ_2 (see Tab. 3) we see that the community size is approximately $N_{\text{cs}} \approx 1/|\psi_1| \approx 100$ that corresponds to the step size in $K_i \approx 70$ for this case.

These results show that the analysis of the link distribution over the PageRank index provides interesting and useful information about characteristics and properties of directed networks.

7 Discussion

In this work, we performed a spectral analysis of eigenvalues and eigenstates of the Google matrix of Wikipedia and other networks. Our study shows that the spectrum of the core space component has eigenvalues in a close vicinity of $\lambda = 1$ and that there are isolated subspaces which give a degeneracy of the eigenvalue $\lambda = 1$. The eigenvalues and eigenstates with relatively large values of $|\lambda|$ can be efficiently determined by the powerful Arnoldi method. These eigenstates are mainly located on well defined network communities. We also find that the spectrum changes drastically from one network to another even if the distribution of links and decay of PageRank is rather similar for the networks considered. This means that the properties of directed networks strongly depend on the internal network structure. We show that the correlation between PageRank and CheiRank vectors highlights specific properties of information flow on directed network. For example, this correlation demonstrates a drastic difference between web sites of BBC and Le Monde. The distribution of links between PageRank nodes also provides an interesting information about the network structure. On the basis of our studies, we argue that the developed spectral analysis of Google matrix brings a deeper understanding of information flow on real directed networks.

We thank A.D. Chepelianskii for making to us available network data collected by him for networks of Cambridge University, Python, BBC, Le Monde in March 2011. Our research presented here is supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No. 288956). This work was granted access to the HPC resources of CALMIP (Toulouse) under the allocation 2012-P0110.

Appendix

The tables are given in the text of the paper. The notations used in the tables are: N is network size, N_ℓ is the number of links, n_A is the Arnoldi dimension used for the Arnoldi method for the core space eigenvalues, N_d is the number of invariant subspaces, $d_{\max}$ gives a maximal subspace dimension, $N_{\text{circ.}}$ notes number of eigenvalues on the unit circle with $|\lambda_i| = 1$, N_1 notes number of unit eigenvalues with $\lambda_i = 1$. We remark that $N_s \geq N_{\text{circ.}} \geq N_1 \geq N_d$ and $N_s \geq d_{\max}$ and the average subspace dimension is given by: $\langle d \rangle = N_s/N_d$. We note that the values of N, N_ℓ for network of Cambridge 2011 are slightly different from those given in [19] due to a slightly different procedure of cleaning of row data collection (e.g., count of pdf and other type nodes). Eigenvalues for eigenvectors are shown in Figure 1 with the colors red, green, blue or pink corresponding to colors of Table 3. The index m of λ_m in Tables 3–7 counts the order number of core eigenvalues in a decreasing order of $|\lambda_m|$.

References

1. Wikipedia, World Wide Web, http://en.wikipedia.org/wiki/World_Wide_Web
2. A.A. Markov, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete **15**, 135 (1906) (in Russian)
3. M. Brin, G. Stuck, *Introduction to dynamical systems* (Cambridge University Press, Cambridge, 2002)
4. S. Brin, L. Page, Computer Networks and ISDN Systems **30**, 107 (1998)
5. A.M. Langville, C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, Princeton, 2006)
6. D. Donato, L. Laura, S. Leonardi, S. Millozzi, Eur. Phys. J. B **38**, 239 (2004)
7. G. Pandurangan, P. Raghavan, E. Upfal, Internet Math. **3**, 1 (2005)
8. N. Litvak, W.R.W. Scheinhardt, Y. Volkovich, Lect. Notes Comput. Sci. **4936**, 72 (2008)
9. S. Serra-Capizzano, SIAM J. Matrix Anal. Appl. **27**, 305 (2005)
10. O. Giraud, B. Georgeot, D.L. Shepelyansky, Phys. Rev. E **80**, 026107 (2009)
11. B. Georgeot, O. Giraud, D.L. Shepelyansky, Phys. Rev. E **81**, 056109 (2010)
12. L. Ermann, A.D. Chepelianskii, D.L. Shepelyansky, Eur. Phys. J. B **79**, 115 (2011)
13. K.M. Frahm, B. Georgeot, D.L. Shepelyansky, J. Phys, A **44**, 465101 (2011)
14. L. Ermann, D.L. Shepelyansky, Acta Phys. Polonica A **120**, A158 (2011), www.quantware.ups-tlse.fr/QWLIB/tradecheirank/
15. K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **85**, 355 (2012), www.quantware.ups-tlse.fr/QWLIB/twittermatrix/
16. A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, Eur. Phys. J. B **77**, 523 (2010), www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/
17. Wikipedia, Google matrix, http://en.wikipedia.org/wiki/Google_matrix
18. A.D. Chepelianskii, *Towards physical laws for software architecture*, arXiv:1003.5455[cs.SE] (2010), www.quantware.ups-tlse.fr/QWLIB/linuxnetwork/
19. L. Ermann, A.D. Chepelianskii, D.L. Shepelyansky, J. Phys. A **45**, 275101 (2012), www.quantware.ups-tlse.fr/QWLIB/dvvadi/
20. Wikipedia, CheiRank, http://en.wikipedia.org/wiki/CheiRank
21. G.W. Stewart, *Matrix Algorithms Volume II: Eigensystems* (SIAM, Philadelphia, 2001)
22. G.H. Golub, C. Greif, BIT Num. Math. **46**, 759 (2006)
23. K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **76**, 57 (2010)
24. Academic Web Link Database Project http://cybermetrics.wlv.ac.uk/database/
25. K. Zyczkowski, M. Kus, W. Slomczynski, H.-J. Sommers, J. Phys. A **36**, 3425 (2003)
26. F. Evers, A.D. Mirlin, Rev. Mod. Phys. **80**, 1355 (2008)