PROJECT PERIODIC REPORT

Grant Agreement number: 288956

Project acronym: NADINE

Project title: New tools and Algorithms for Directed Network analysis

Funding Scheme: Small or medium-scale focused research project (STREP)

Periodic report: 1st X 2nd

Period covered: from 1.5.2012 to 31.10.2013

Name, title and organisation of the scientific representative of the project's coordinator[1]:

Dr. Dima Shepelyansky

Directeur de recherche au CNRS

Lab de Phys. Theorique, Universite Paul Sabatier, 31062 Toulouse, France

Tel: +331 5 61556068, Fax: +33 5 61556065, Secr.: +33 5 61557572

E-mail: dima@irsamc.ups-tlse.fr; URL: www.quantware.ups-tlse.fr/dima

Project website address: www.quantware.ups-tlse.fr/FETNADINE/

[1] Usually the contact person of the coordinator as specified in Art. 8.1. of the grant agreement

NADINE DELIVERABLE D2.1.

It is based on milestones M2, M3, M6(in progress) , M11(in progress) with deliverable publications:

[4] P1.4 K.M.Frahm, A.D.Chepelianskii and D.L.Shepelyansky, **"PageRank of integers"**, J. Phys. A: Math. Theor. v.45, p.405101 (2012) (arXiv:1205.6343[cs.IR], 2012)

[5] P1.5 K.M.Frahm and D.L. Shepelyansky **"Google matrix of Twitter"**, Eur. Phys. J. B v.85, p.355 (2012) (arXiv:1207.3414[cs.SI], 2012)

[6] P1.6 L.Ermann, K.M.Frahm and D.L. Shepelyansky **"Spectral properties of Google matrix of Wikipedia and other networks"**, Eur. Phys. J. B v.86, p.193 (2013) (arXiv:1212.1068 [cs.IR], 2012)

[7] P1.7 V.Kandiah and D.L.Shepelyansky, **"Google matrix analysis of DNA sequences"**, PLOS One v.8(5), p. e61519 (2013) (arXiv:1301.1626[q-bio.GN], 2013)

[8] P1.8 Y.-H.Eom, K.M.Frahm, A.Benczur and D.L. Shepelyansky, **"Time evolution of Wikipedia network ranking"**, submitted Eur. Phys. J. B (2013) (arXiv:1304.6601 [physics.soc-ph], 2013

[11] P1.11 K.M.Frahm and D.L. Shepelyansky, **"Poincare recurrences and Ulam method for the Chirikov standard map"**, Eur. Phys. J. B v.86, p.322(2013) (arXiv:1302.2761 [nlin.CD] , 2013

[12] P1.12 K.M.Frahm, Y.-H.Eom and D.L. Shepelyansky, **"Google matrix of the citation network of Physical Review"**, submitted to Phys. Rev. E Oct 21, 2013 (arXiv:1310.5624 [physics.soc-ph], 2013)

[

PageRank of integers

J. Phys. A: Math. Theor. **45** (2012) 405101 (20pp) doi:10.1088/1751-8113/45/40/405101

PageRank of integers

K M Frahm[1], **A D Chepelianskii**[2] **and D L Shepelyansky**[1]

[1] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France
[2] Department of Physics, Cavendish Laboratory, University of Cambridge, CB3 0HE, UK

E-mail: dima@irsamc.ups-tlse.fr

Received 29 May 2012, in final form 17 August 2012
Published 18 September 2012
Online at stacks.iop.org/JPhysA/45/405101

Abstract

We up a directed network tracing links from a given integer to its divisors and analyze the properties of the Google matrix of this network. The PageRank vector of this matrix is computed numerically and it is shown that its probability is approximately inversely proportional to the PageRank index thus being similar to the Zipf law and the dependence established for the World Wide Web. The spectrum of the Google matrix of integers is characterized by a large gap and a relatively small number of nonzero eigenvalues. A simple semi-analytical expression for the PageRank of integers is derived that allows us to find this vector for matrices of billion size. This network provides a new PageRank order of integers.

PACS numbers: 02.10.De, 02.50.−r, 89.75.Fb

(Some figures may appear in colour only in the online journal)

1. Introduction

Number theory [1] is the fundamental branch of mathematics where the theory of prime numbers, besides its beauty, finds important cryptographic applications [2]. It is established that the methods of random matrix theory and quantum chaos find their useful applications for the understanding of properties of prime numbers and the Riemann zeros [3–5].

In this work, we propose another matrix approach to number theory based on the Markov chains [6][3] and the Google matrix [7]. The latter finds important applications for the information retrieval and Google search engine of the World Wide Web (WWW) [8]. The right eigenvector of the Google matrix with the largest eigenvalue is known as the PageRank vector. The elements of this vector are non-negative and have the meaning of probability of finding a random surfer on the network nodes. The PageRank algorithm ranks all websites in decreasing order of

[3] English translation 'Extension of the limit theorems of probability theory to a sum of variables connected in a chain' reprinted in appendix B of the second part of [6].

components of the PageRank vector (see e.g. detailed description in [8]). Here, we propose a natural way to construct the Google matrix of positive integers using their division properties. We study the statistical properties of the PageRank vector of this matrix and discuss the properties of a new order of integers given by this ranking. The properties of the eigenvalues and eigenvectors are also discussed.

The paper is constructed as follows: in section 2, we give the definition of the Google matrix of integers; in section 3, the properties of its PageRank vector are analyzed; in section 4, the analysis of spectral properties is given; in sections 4 and 5, the analytical expressions for the PageRank vector are presented and in section 6, the discussion of the results is presented.

2. Google matrix of integers

The elements of the Google matrix $G(\alpha)$ of a directed network with N nodes are given by

$$G_{mn}(\alpha) = \alpha S_{mn} + (1-\alpha)/N. \tag{1}$$

Here the matrix S is obtained by normalizing to unity all columns of the adjacency matrix A_{mn}, and replacing the elements of columns with only zero elements, corresponding to dangling nodes, by $1/N$. An element A_{mn} of the adjacency matrix is equal to unity if a node n points to the node m and zero otherwise. The damping parameter α in the WWW context describes the probability $(1-\alpha)$ of jumping to any node for a random surfer. The value $\alpha = 0.85$ gives a good classification of pages for WWW [8]. The matrix G belongs to the class of Perron–Frobenius operators [8], its largest eigenvalue is $\lambda = 1$ and the other eigenvalues obey $|\lambda| \leqslant \alpha$. In typical WWW networks, the eigenvalue $\lambda = 1$ is strongly degenerate at $\alpha = 1$ (see e.g. [9]) and the introduction of $\alpha < 1$ becomes compulsory to define a unique right eigenvector at $\lambda = 1$ and to ensure the convergence of the PageRank vector by the power iteration method [8]. The right eigenvector at $\lambda = 1$ gives the probability $P(n)$ of finding a random surfer at site n and is called the PageRank. Once the PageRank is found, all nodes can be sorted by decreasing probabilities $P(n)$ and increasing index $K(n)$. The node rank is then given by the index $K(n)$ which reflects the relevance of the node corresponding to a positive integer n. For the WWW, the PageRank dependence on K is well described by a power law $P(K) \propto 1/K^{\beta_{\text{in}}}$ with $\beta_{\text{in}} \approx 0.9$ [8, 9]. This is consistent with the relation $\beta_{\text{in}} = 1/(\mu_{\text{in}} - 1)$ corresponding to the average proportionality of the PageRank probability $P(n)$ to its in-degree distribution $w_{\text{in}}(k) \propto 1/k^{\mu_{\text{in}}}$ where $k(n)$ is a number of ingoing links for a node n [8]. For the WWW, it is established that for the ingoing links $\mu_{\text{in}} \approx 2.1$ (with $\beta_{\text{in}} \approx 0.9$), while for the out-degree distribution w_{out} of outgoing links, a power law has the exponent $\mu_{\text{out}} \approx 2.7$ [10, 11]. Here we analyze the properties of PageRank and use the notation $\beta = \beta_{\text{in}}$. Finally, we note that usually for WWW, the analysis is done for the exponent μ (see e.g. [10, 11]) related to $\mathrm{d}K \sim \mathrm{d}P/P^{-\mu} \sim w_{\text{in}}(k)$, but here we prefer to analyze the exponent β which is related to μ by a simple relation $\beta = 1/(\mu - 1)$.

To construct the Google matrix of integers, we define for $m, n \in \{1, \ldots, N\}$ the adjacency matrix by $A_{mn} = k$ where k is a 'multiplicity' defined as the largest integer such that m^k is a divisor of n and if $1 < m < n$, and $k = 0$ if $m = 1$ or $m = n$ or if m is not a divisor of n. Thus, we have $k = 0$ if m is not a divisor of n and $k \geqslant 1$ if m is a divisor of n different from 1 and n. The total size N of the matrix is fixed by the maximal considered integer.

This defines a network where an integer number n is linked to its divisors m different from 1 and n itself and where the transition probability is proportional to the multiplicity k, the number of times we can divide n by m. The number 1 and the prime numbers are therefore not linked to any other number and correspond to dangling nodes in the language of WWW networks. For example, the number $n = 24$ has links pointing to

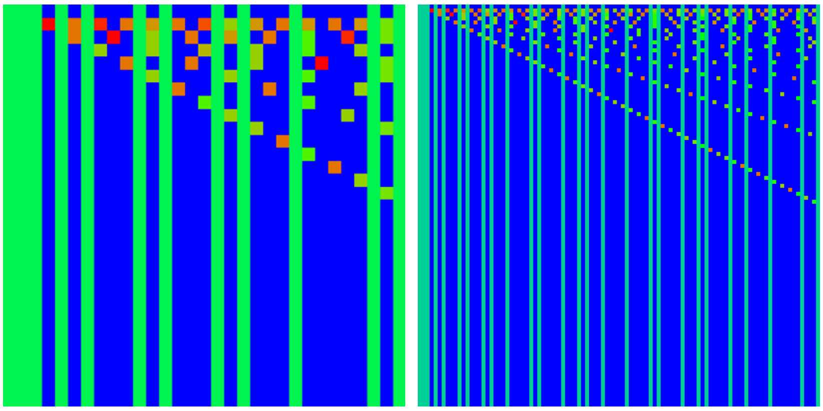

Figure 1. The Google matrix of integers: the amplitudes of the matrix elements G_{mn} at $\alpha = 1$ are shown by color: blue for minimal zero elements and red for maximal unity elements, with $1 \leqslant n \leqslant N$ corresponding to the x-axis (with $n = 1$ corresponding to the left column) and $1 \leqslant m \leqslant N$ to the y-axis (with $m = 1$ corresponding to the upper row). The matrix sizes are $N = 31$ in the left panel and $N = 101$ in the right panel.

$m(k) = 2(3), 3(1), 4(1), 6(1), 8(1), 12(1)$ (multiplicity is given in parentheses) so that the nonzero matrix elements in this column are $3/8$, $1/8$, $1/8$, $1/8$, $1/8$, $1/8$, respectively. We find the total number of links $N_\ell = \sum_{mn} A_{mn}$, taking into account the multiplicity, to be $N_\ell = 6005$ at $N = 1000$, $N_\ell = 1066\,221$ at $N = 10^5$, $N_\ell = 152\,720\,474$ at $N = 10^7$ and $N_\ell = 19\,877\,650\,264$ at $N = 10^9$. The fit of the dependence $N_\ell = N\,(a_\ell + b_\ell \ln N)$ gives $a_\ell = -0.901 \pm 0.018$, $b_\ell = 1.003 \pm 0.001$.

From the adjacency matrix A, we first construct a matrix S_0 by normalizing the sum in each column, containing at least one non-zero element, to unity and the matrix S is obtained from S_0 by replacing the elements of columns with only zero elements, corresponding to dangling nodes 1 and prime numbers, by $1/N$. The Google matrix G is finally obtained from S by equation (1) for an arbitrary damping factor. The PageRank is the right eigenvector of the matrix G with the maximal eigenvalue $\lambda = 1$: $GP = \lambda P = P$.

The examples of the Google matrix G at $\alpha = 1$ for $N = 31$, 101 are shown in figure 1. We see that most elements are concentrated above the main matrix diagonal since the divisors m are smaller than the number n itself. The only exceptions are given by the columns at 1 and the prime numbers p which have no divisors (apart from 1 and p) and hence they correspond to the dangling nodes with no direct links pointing to them. The amplitude of the elements in these columns is uniformly $1/N$. The structure of the matrix clearly shows the presence of diagonals $m = n/2$, $n/3$, $\ldots$ corresponding to the small divisors $m' = 2, 3, \ldots$, which appear rather often in the division of integers. This structure is preserved up to the largest size $N = 10^9$ considered in this work.

As we will see in section 4, the eigenvalue $\lambda_0 = 1$ of the matrix S is non-degenerate (contrary to typical realistic WWW networks [9]) and in addition, its spectrum has a large gap with λ_0 and the other eigenvalues $|\lambda_i| < 0.6$. In such a case, the PageRank vector $P(K)$ has a very small variation when the damping factor α is changed in the range $0.85 \leqslant \alpha \leqslant 1$ and the convergence of the power method to calculate the PageRank is well assured, actually quite fast, even for the damping parameter $\alpha = 1$. Therefore, we limit in this work our studies to the case $\alpha = 1$ at which G coincides with the matrix S and from now on we denote S as ‘the Google matrix’.

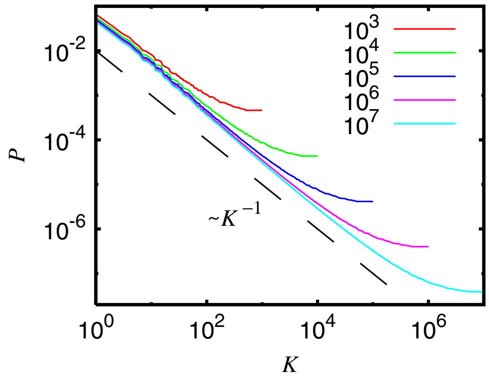

Figure 2. Dependence of PageRank probability $P(K)$ on the PageRank index K for the matrix sizes $N = 10^3$, 10^4, 10^5, 10^6, 10^7; the dashed straight line shows the Zipf law dependence $P \sim 1/K$.

Finally, we note that certain networks constructed from integers have been considered in [12, 13] but these networks were nondirectional and the Google matrix analysis was not performed there.

3. PageRank order of integers

We first determine the PageRank vector of the Google matrix numerically by the power iteration method [8] or by the Arnoldi method [14] using an Arnoldi dimension of size n_A, which allows us to find several eigenvalues and eigenvectors with largest $|\lambda|$ for a full matrix size of a few millions (see more details in [9, 15]).

The dependence of PageRank probability $P(K)$ on the PageRank index K is shown in figure 2. We see that with the growth of the system size N, the dependence $P(K)$ converges to a fixed distribution $P(K)$ on initial $K \leqslant N/10$ values with the tail of distribution $P(K)$ at $K > N/10$, which is sensitive to the cut-off at the finite matrix size N. In the convergent part, a formal fit (for $10 < K < 10^5$) gives the dependence $P \sim A/K^{\beta}$ with $\ln A = 0.0431 \pm 0.000\,49$, $\beta = 1.040 \pm 0.0015$ being close to the Zipf law with $\beta = 1$ [16]. The small value of $\beta - 1$ indicates that there can be a logarithmic correction. Indeed, the fit $1/(PK) = a_1 + b_1 \ln K$ (for $10 < K < 10^3$) gives the values $a_1 = 16.050 \pm 0.187$, $b_1 = 2.468 \pm 0.036$. Thus, it is possible that in the limit of $N \to \infty$, we have the asymptotic behavior $P \sim 1/(K \ln K)$. Such a scaling looks to be more probable due to usual logarithmic corrections in the density of primes [2]. However, for the available finite matrix sizes, the regime of linear behavior of $1/(PK)$ versus $\ln K$ is quite limited and it is not obvious how to distinguish between the above two fitting dependences.

The dependence of PageRank probability P on the integer index n is shown in figure 3. It is characterized by a global decay $P \propto 1/n$ with the presence of various branches which are especially well visible for the rescaled quantity nP. This structure is preserved with the increase of matrix size for the values of $n < N/100$. The direct check shows that the highest plateau corresponds to the prime numbers p.

Another way to analyze the structures visible in figure 3 is to consider the dependence of n on the PageRank index K obtained from the PageRank probability $P(K_n)$. In fact K gives a new order of integers imposed by the PageRank. The dependence $n(K)$ is shown in figure 4 on a large scale. In the first approximation, we find the layered structure with a sequence of parallel lines $n \propto K$. This global structure is preserved with the increase of the matrix size from $N = 10^5$ to 10^7.

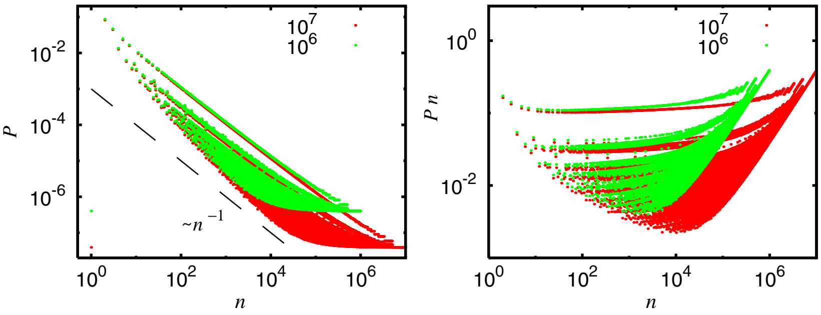

Figure 3. Dependence of PageRank probability P on the integer number n for matrix sizes $N = 10^6, 10^7$ (left panel: green and red points, respectively), and rescaled probability nP on n (right panel); data are shown in log–log scale.

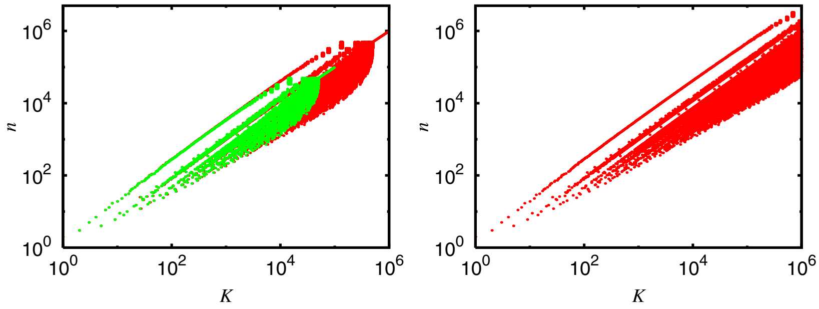

Figure 4. Dependence of the integer number n on the PageRank index K for sizes $N = 10^5, 10^6$ (left panel: green and red points, respectively) and 10^7 (right panel); data are shown in log–log scale.

A more detailed view of this structure is shown in figure 5. There are well-defined separated branches with approximately linear dependence $n \approx \kappa K$ with $\kappa \approx 4.5$ for the highest branch, which corresponds to the highest plateau in figure 3 (right panel). This branch contains only primes. The lower branch contains semi-primes (products of two primes) and so on down to smaller and smaller values of κ. The whole structure looks to have a self-similar structure as it shows a zoom to a smaller scale. The increase of the size N gives some modifications of the structure keeping its global pattern (see figure 5, bottom panels). There is a certain clustering on the (n, K) plane of rectangles containing close values of K and integer numbers n. The rectangles in the upper prime-branch contain exclusively prime numbers for $n = p$. Note that the neighboring non-prime values appear in other rectangles on the right side for larger values of K. For example, in the bottom-left panel of figure 5, we have a rectangle at $K \sim 2.6 \times 10^4$ and $n \sim 10^5$ with primes but there is at $K \sim 7 \times 10^4$ another rectangle of semi-primes, also with the values $n \sim 10^5$.

The direct analysis shows that the rectangles in figure 5 correspond to flat plateaux with degenerate values of $P(K_n)$ (see the global dependence shown in figure 2) appearing for finite matrix size N. This degeneracy results from only rational numbers appearing in the elements of the Google matrix and from its very sparse structure. Inside such flat regions, the ordering in K is somewhat arbitrary and depends on the precise sorting algorithm used. The K index shown in figure 5 was obtained by the Shellsort method that may indeed produce quite a

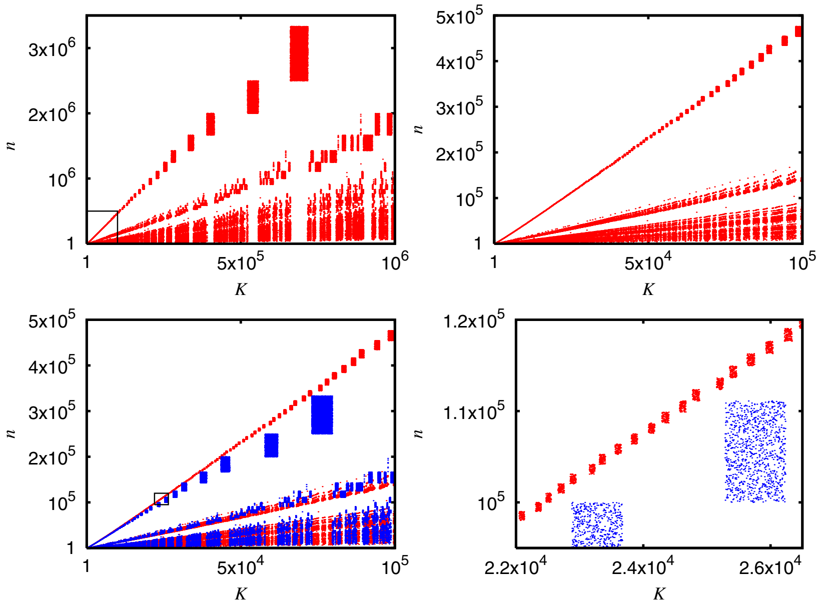

Figure 5. Top panels: the dependence of the integer number n on the PageRank index K for size $N = 10^7$ shown by red points (left panel); the right panel shows zoom of data in a rectangle from the left panel. Bottom panels: in addition to the data of the top right panel, data for $N = 10^6$ are shown (left panel); the right panel shows zoom of data in a rectangular region from the left panel. Data are shown in usual scale.

random ordering for degenerate values, thus generating the rectangles seen in figure 5. We have verified that when using a modified sorting algorithm with a secondary criterion, to sort with increasing n inside a degenerate region, the rectangles are replaced by lines from the left bottom corner to the right top corner. With increasing values of N, these rectangles are reduced in size. We numerically find that the first degenerate plateau appears at $K = K_d$ and that this number increases with the matrix size N, e.g. $K_d = 27$ at $N = 1000$, 177 at 10^5, 1287 at 10^7 and 10 386 at 10^9. This dependence is well described by the fit $K_d = a_d K^{b_d}$ with $a_d = 1.284 \pm 0.078$, $b_d = 0.432 \pm 0.004$. We return to discussion of the convergence at large N a bit later.

Since we find an approximate linear growth of n with K inside each branch, it is useful to consider the dependence of the ratio n/K on K, which is shown in figure 6. The upper branch of primes is well described by the dependence $n/K = b_2 \ln K + a_2$ with $b_2 = 0.322$, $a_2 = 1.358$. This shows that in the previous relation, κ is not a constant but grows logarithmically with K. We have an approximate relation $b_2 = 0.322 \approx 1/b_1 = 1/2.468$. The lower branches also have an approximately logarithmic growth of the ratio n/K with K.

Finally, let us discuss the stability of the PageRank order of integers with respect to the variation of the matrix size N. The dependence $P(K)$ is definitely converging to a fixed function for $K \ll N$ as is well seen in figure 2. However, for a fixed integer n, its PageRank index K_n has a visible variation with the increase of matrix size N. These variations are visible in figure 5 (bottom panels). At the same time, the global structure of the K_n or $n(K)$ dependence shows signs of convergence with the growth of N. A more detailed analysis of variation of $\Delta K = |K_n(N_1) - K_n(N_2)|$ for two matrix sizes $N_2 = 10N_1$ is shown in figure 7. We see that there is a significant decrease in variations ΔK with increase in N_1, even if a small change of K_n values is visible even at relatively low $n \sim 100$. On the basis of these data, we make a

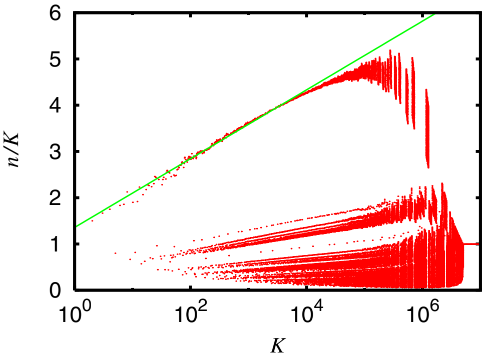

Figure 6. Dependence of the ratio n/K on the PageRank index K for size $N = 10^7$; data are shown in semi-log scale. The straight line shows the fit dependence $n/K = a_2 + b_2 \ln K$ for the upper branch in the range $10 \leqslant K \leqslant 10^4$ with $a_2 = 1.3583 \pm 0.0099$, $b_2 = 0.3227 \pm 0.0014$.

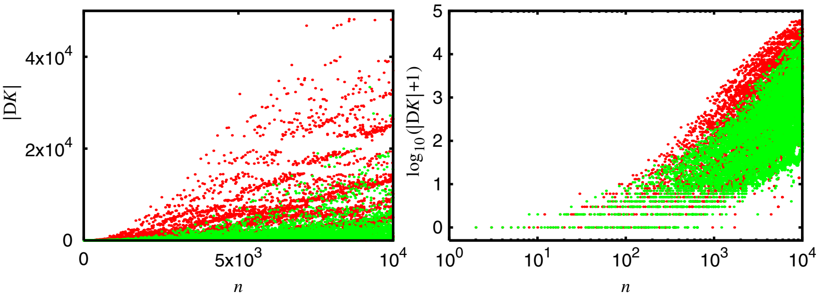

Figure 7. Dependence of $|\Delta K| = |K_n(N_2) - K_n(N_1)|$ on the integer n for matrix sizes $N_1 = 10^6, N_2 = 10^7$ (green points) and $N_1 = 10^5, N_2 = 10^6$ (red points). The left and right panels show the same data either in normal or in log–log scales.

conjecture that in the limit of $N \to \infty$, we will have a convergence to a fixed PageRank order of integers K_n. However, we expect that this convergence is very slow, probably logarithmic in N, thus being the reason that, even at $N = 10^7$, we find some variations in K_n. We note that the density of states of Riemann zeros also shows very slow convergence so that enormously large values of $n \sim N \sim 10^{20}$ are required to obtain stable results [3, 4].

4. Spectral properties of the Google matrix of integers

4.1. Arnoldi method

To study numerically the spectrum of the Google matrix $S = G$ of integers at $\alpha = 1$, we first employ the Arnoldi method [14, 15]. This method uses a normalized initial vector ξ_0 and generates a *Krylov space* by the vectors $S^j\, \xi_0$ for $j = 0, \ldots, n_A - 1$, where n_A is called the Arnoldi dimension. Using Gram–Schmidt orthogonalization, one determines an orthogonal basis of the Krylov space and the matrix representation of S in this basis. This provides a matrix $\bar{S}$ of modest dimension n_A of Hessenberg form which can be diagonalized by standard QR-methods and whose eigenvalues, called *Ritz eigenvalues*, are in general very accurate approximations of the largest eigenvalues of the original (very large) matrix S.

In this work, we have used the Arnoldi dimension $n_A = 1000$ and two different initial vectors: first a random initial vector and second a uniform initial vector with identical

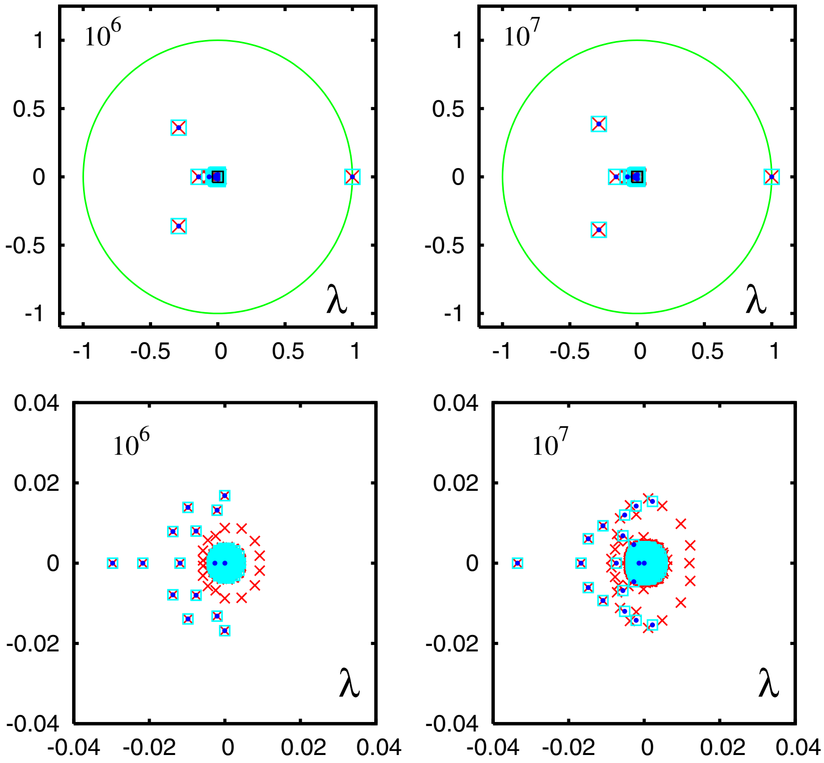

Figure 8. Spectrum of the Google matrix of integers for the matrix size $N = 10^6$ (left panels) and 10^7 (right panels); the red crosses (light blue squares) represent numerical data from the Arnoldi method with Arnoldi dimension $n_A = 1000$ and a random initial vector (with the unit initial vector), and the dark blue points represent the exact eigenvalues obtained as the zeros of the reduced polynomial of equation (6). The top panels show the whole spectrum and the bottom panels show a zoom of the region represented by black squares in the top panels. The eigenvalues have significantly higher accuracy for the Arnoldi method with unit initial vector. The unit circle $|\lambda| = 1$ is shown in green.

components $1/\sqrt{N}$ (thus normalized by the Euclidean norm $\|(\cdots)\|_2$). The spectrum of the matrix S is shown in figure 8 for two sizes $N = 10^6$, 10^7. We see that there are only three eigenvalues within the ring $0.05 < |\lambda| < 0.5$ while the majority of eigenvalues is concentrated inside a range of $|\lambda| < 0.05$. The first few largest eigenvalues are accurately obtained from both initial vectors used for the Arnoldi method and also coincide (up to numerical precision) with the eigenvalues determined by a semi-analytical approach (see below). However, for the range $|\lambda| < 0.05$, the situation becomes more subtle, as discussed below.

We note that figure 8 shows a large gap between $\lambda_0 = 1$ and the next eigenvalue, thus justifying our above choice of the damping factor $\alpha = 1$.

4.2. Analytical discussion of spectrum

The Google matrix S at $\alpha = 1$ has a very particular structure that allows us to establish some important properties for the spectrum and its eigenvalues. We can write

$$S = S_0 + v\,d^T, \tag{2}$$

where v and d are two vectors of size N with components $v_n = 1/N$ and $d_n = 1$ for the prime numbers $n = p$ or $n = 1$ and $d_n = 0$ for the other non-prime numbers (different from 1). For later use, we also introduce the vector e with components $e_n = 1$ and therefore $v = e/N$. In addition, d^T denotes the transposed line vector of d. The matrix S_0 is the contribution that arises from the adjacency matrix A by normalizing the non-vanishing columns of the latter and the tensor product $v\,d^T$ represents the values $1/N$ that are put in the zero columns of S_0 when constructing the full matrix S. The normalization condition of the non-vanishing columns of S_0 can be formally written as $e^T\,S_0 = e^T - d^T$ which is just the line vector with components 0 for the vanishing columns of S_0 (for prime numbers n or $n = 1$) and 1 for the non-vanishing columns of S_0 (for the other non-prime numbers different from 1). This expression provides the useful identity

$$d^T = \mathrm{e}^T(\mathbb{1} - S_0). \tag{3}$$

Furthermore, we observe that the matrix S_0 has a trigonal form with vanishing entries on the diagonals because $(S_0)_{mn} \neq 0$ only if m is a divisor of n different from 1 and n, and therefore for any non-vanishing matrix element $(S_0)_{mn}$, we have $m \leqslant n/2 < n$. This matrix structure can also be seen in figure 1. As a consequence, S_0 is nilpotent with $S_0^l = 0$ for some integer l. In the following, let us assume that l is the minimal number such that $S_0^l = 0$. Obviously in our model, $l = [\log_2(N)]$ is actually a very modest number as compared to the full matrix size N.

We now discuss how the form of equation (2) affects the eigenvalues of the full matrix S. Let ψ be a right eigenvector of S and λ its eigenvalue:

$$\lambda\psi = S\psi = S_0\psi + C\,v, \quad C = d^T\,\psi = \sum_{n\text{ prime or }n=1}^{N} \psi_n. \tag{4}$$

If $C = 0$, we find that ψ is an eigenvector of S_0. Then $\lambda = 0$ since the matrix S_0 is nilpotent and cannot have non-vanishing eigenvalues. The matrix S_0 is actually non-diagonalizable and can only be transformed to a Jordan form with quite large Jordan blocks and 0 as the diagonal element of each of the Jordan blocks.

Suppose now that $C \neq 0$ implying that $\lambda \neq 0$ since the equation $S_0\psi = -C\,v$ does not have a solution for ψ because S_0 has many zero rows and $v_n = 1/N \neq 0$ for each $n = 1, \ldots, N$. Since $\lambda \neq 0$, the trigonal matrix $\lambda\mathbb{1} - S_0$ is invertible and from equation (4), we obtain

$$\psi = C\,(\lambda\mathbb{1} - S_0)^{-1}\,v = \frac{C}{\lambda}\sum_{j=0}^{l-1}\left(\frac{S_0}{\lambda}\right)^j v. \tag{5}$$

Note that the sum is finite since $S_0^l = 0$. The eigenvalue λ is determined by the condition that this expression of ψ has to satisfy the condition $C = d^T\,\psi$. Multiplying this condition by λ^l/C, we find that λ is a zero of the following *reduced polynomial* of degree l:

$$\mathcal{P}_r(\lambda) = \lambda^l - \sum_{j=0}^{l-1}\lambda^{l-1-j}\,c_j = 0, \quad c_j = d^T\,S_0^j\,v. \tag{6}$$

This calculation shows that there are at most l eigenvalues $\lambda \neq 0$ of S given as the zeros of this reduced polynomial.

We note that using $S_0^l = 0$ and identity (3), one finds that the coefficients c_j obey the following sum rule:

$$\sum_{j=0}^{l-1} c_j = d^T\left(\sum_{j=0}^{l-1} S_0^j\right)\,v = \mathrm{e}^T(\mathbb{1} - S_0)(\mathbb{1} - S_0)^{-1}\,v = 1 \tag{7}$$

since $\mathrm{e}^T\, v = \sum_n v_n = 1$. This sum rule ensures that $\lambda = 1$ is a zero of the reduced polynomial and the PageRank as the eigenvector of $\lambda = 1$ is obtained from (5):

$$P = C \sum_{j=0}^{l-1} S_0^j\, v, \quad C^{-1} = \sum_{j=0}^{l-1} \mathrm{e}^T\, S_0^j\, v, \tag{8}$$

where the identity for C^{-1} is due to the normalization of P.

Since the degree $l = [\log_2(N)]$ of the reduced polynomial is very modest, $9 \leqslant l \leqslant 29$ for $10^3 \leqslant N \leqslant 10^9$, we have determined numerically the coefficients c_j, which only require a finite number of successive multiplications of S_0 to the initial vector v, and determined the zeros of the reduced polynomial by the very efficient Newton–Maehly method in the complex plane. The resulting l eigenvalues (and the trivial highly degenerate eigenvalue $\lambda = 0$ of S) obtained from this semi-analytical method are also shown in figure 8.

The numerical determination of the zeros shows that they are all simple zeros of the reduced polynomial but at this point, we are not yet sure that they are also non-degenerate as far as the full matrix S is concerned. In theory we might still have the principal vectors ϕ associated with some eigenvalue $\lambda \neq 0$ such that $S\phi = \lambda\phi + \psi$ with ψ being the eigenvector at λ. However, we can exclude this scenario by determining the full characteristic polynomial of S:

$$\begin{aligned}\mathcal{P}_S(\lambda) &= \det(\lambda\mathbb{1} - S_0 - v\, d^T)\\ &= \lambda^N \det(\mathbb{1} - S_0/\lambda)\, \det[\mathbb{1} - (\mathbb{1} - S_0/\lambda)^{-1}\, v\, d^T/\lambda]\\ &= \lambda^N [1 - d^T(\mathbb{1} - S_0/\lambda)^{-1}\, v/\lambda] = \lambda^{N-l}\, \mathcal{P}_r(\lambda)\end{aligned} \tag{9}$$

since $\det(\mathbb{1} - S_0/\lambda) = 1$, $\det(\mathbb{1} - u\, w^T) = (1 - w^T\, u)$ for the arbitrary vectors u and w, and the matrix inverse has been expanded in a finite sum in a similar way as in equation (5). According to equation (9), we observe that the simple zeros of $\mathcal{P}_r(\lambda)$ are also simple zeros of $\mathcal{P}_S(\lambda)$ and have therefore an algebraic multiplicity equal to 1. This proves that there are no principal vectors and no non-trivial Jordan-block structure for $\lambda \neq 0$. On the other hand, the eigenvalue $\lambda = 0$ has the algebraic multiplicity $N - l$ with many large Jordan blocks.

The l-dimensional subspace associated with the eigenvalues $\lambda \neq 0$ is according to equation (5) generated by the l vectors $v^{(j)} = S_0^j\, v$ with $j = 0, \ldots, l-1$, which form a basis of this subspace. Using equations (2) and (6), we may easily determine the matrix representation of S with respect to this basis by

$$S\, v^{(j)} = c_j\, v^{(0)} + v^{(j+1)} = \sum_{k=0}^{l} \bar{S}_{k+1,j+1}\, v^{(k)}, \qquad j = 0, \ldots, l-1, \tag{10}$$

where for simplicity of notation for the case $j = l - 1$, we write $v^{(l)} = 0$. The $l \times l$-matrix $\bar{S}$ has the explicit form

$$\bar{S} = \begin{pmatrix} c_0 & c_1 & \cdots & c_{l-2} & c_{l-1}\\ 1 & 0 & \cdots & 0 & 0\\ 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{11}$$

One easily verifies that the characteristic polynomial $\mathcal{P}_{\bar{S}}(\lambda)$ of this matrix coincides with the reduced polynomial (6) and its l eigenvalues are therefore exactly the l non-vanishing eigenvalues of the full matrix S. Using the sum rule (7), one notes that the l-dimensional vector $(1, \ldots, 1)^T$ is a right eigenvector of $\bar{S}$ with eigenvalue $\lambda = 1$, thus confirming the PageRank expression $P \propto \sum_{j=0}^{l-1} v^{(j)}$ (see also equation (8)).

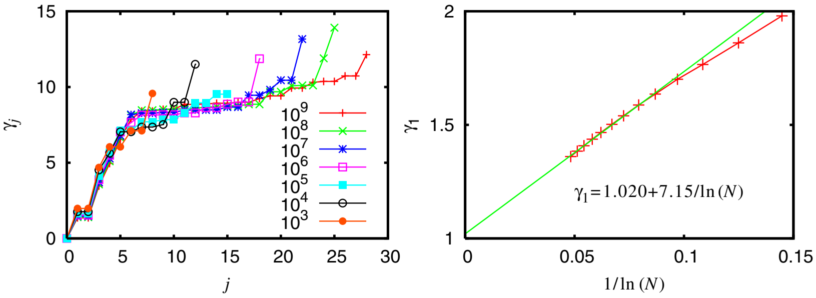

Figure 9. Left panel: the dependence of $\gamma_j = 2\ln|\lambda_j|$ on the index j for the l non-vanishing eigenvalues of S and various matrix sizes N. Right panel: the dependence of γ_1 on $(\ln N)^{-1}$ (red line with crosses). The green line corresponds to the fit $\gamma_1(N) = \gamma_1(\infty) + \Delta\gamma/\ln N$ for the range $10^5 \leqslant N \leqslant 10^9$ (i.e. $(\ln N)^{-1} < 0.09$) with $\gamma_1(\infty) = 1.020 \pm 0.006$ and $\Delta\gamma = 7.14 \pm 0.09$.

A direct numerical diagonalization of matrix (11) is tricky and fails to produce the smaller eigenvalues (below 10^{-2}) due to numerical rounding errors since the coefficients c_j decay very rapidly, e.g. $c_{22} \sim 10^{-38}$ for $N = 10^7$ with $l = 23$. However, we may numerically diagonalize the 'equilibrated' matrix, $\rho^{-1}\bar{S}\rho$, which has the same eigenvalues as $\bar{S}$ and where ρ is a diagonal matrix with the diagonal matrix elements $\rho_{jj} = 1/c_{j-1}$. The eigenvalues obtained from the equilibrated matrix coincide very precisely (up to numerical precision 10^{-14}) with the zeros obtained from the reduced polynomial by the Newton–Maehly method. In figure 8, we also show these l zeros for $N = 10^6$ and $N = 10^7$. Apparently, both variants of the Arnoldi method fail to confirm the analytical result that there are only l non-vanishing eigenvalues, a point we attribute to the numerical instability of the highly degenerate and defective eigenvalue $\lambda = 0$ and which we will discuss below.

To study the evolution of the eigenvalue spectrum with N, it is actually convenient to introduce the variable $\gamma_j = -2\ln|\lambda_j|$. The dependence of γ_j on the index j is shown in the left panel of figure 9. It appears that the γ-spectra for different values of N fall roughly on the same curve except for the last one or two values of each spectrum. This universal curve can be roughly approximated by a piecewise linear function with two slopes $\approx 4/3$ for $0 \leqslant j \leqslant 6$ and $\approx 1/7$ for $6 \leqslant j \leqslant 28$.

We note that the convergence of the first nonzero γ_1 is compatible with the law $\gamma_1(N) \approx \gamma_1(\infty) + \Delta\gamma/\ln N$ with $\gamma_1(\infty) = 1.020 \pm 0.006$ and $\Delta\gamma = 7.14 \pm 0.09$ obtained from a fit in the range $10^5 \leqslant N \leqslant 10^9$. This fit is actually very accurate as can be seen from the small error of $\gamma_1(\infty)$ and the right panel of figure 9. Once more, such a dependence indicates a very slow logarithmic convergence with the system size N.

In figure 10, we show the amplitude $|\psi_1|$ of the second eigenvector ψ_1 at $\lambda_1 = -0.284\,22 + \mathrm{i}\,0.387\,26$ for $N = 10^7$ versus the K index. Despite some fluctuations, this eigenvector seems to be close to the PageRank as far as the overall distribution of very large and small values is concerned. This behavior does not come as a surprise in view of the expansion (see equation (5))

$$\psi_1 \propto \sum_{j=0}^{l-1} \lambda_1^{-j-1}\, v^{(j)}. \tag{12}$$

In principle, the fact that $|\lambda_1|$ is well below 1 indicates that the contributions of $v^{(j)}$ for the larger values of j increase. However, as we will discuss in the next section, the overall size of $v^{(j)}$ decays with increasing j much faster than the increase by the factor λ_1^{-j-1} and therefore

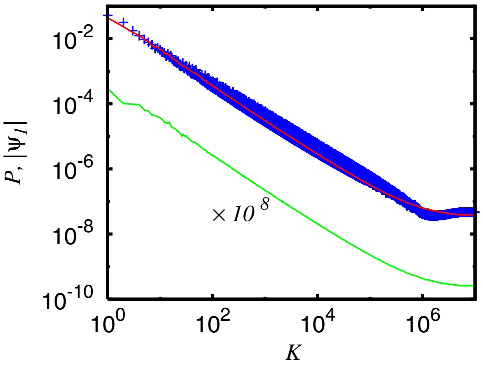

Figure 10. Dependence of the PageRank vector P (red curve) and the eigenvector $|\psi_1|$ (blue crosses) on the PageRank index K for $N = 10^7$. Here the eigenvalue is $\lambda_1 = -0.284\,22 + \mathrm{i}\,0.387\,26$ ($|\lambda_1| = 0.480\,37$, $\gamma_1 = 1.4663$, and the corresponding ψ_1 is normalized by the condition $\sum_n |\psi_1(n)| = 1$); the green curve shows the difference $|\Delta P|$ between the numerically computed PageRank P (red curve) and semi-analytical computation of PageRank; for clarity, $|\Delta P|$ is multiplied by a factor of 10^8.

mainly the first few terms of this sum contribute to ψ_1 in a similar way as for the PageRank (see section 5).

Finally in figure 10, also the numerical difference of the PageRank determined by the standard power method and the semi-analytical expression (8) is shown. The relative difference is $\sim 10^{-10}$ for the full range of K, thus numerically confirming the accuracy of equation (8).

4.3. Numerical problems due to Jordan blocks

The question arises why the Arnoldi method for both initial vectors, random and uniform (and also direct numerical diagonalization for small matrix sizes $N \leqslant 10^4$), fails to confirm the analytical result that there are only $l = [\log_2(N)]$ non-zero eigenvalues $\lambda \neq 0$ of S. The reason is that the big subspace of dimension $N - l$ associated with the eigenvalue $\lambda = 0$ with a lot of large Jordan blocks is numerically very problematic. This effect for such a *defective eigenvalue* is well known in the theory of numerical diagonalization methods [14]. To understand this a bit clearer, consider a 'perturbed' Jordan block of size D:

$$\begin{pmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \varepsilon & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{13}$$

which has a characteristic polynomial $\lambda^D - (-1)^D\varepsilon$ and therefore complex eigenvalues that scale as $|\lambda| \sim \varepsilon^{1/D}$ as a function of the perturbation ε, while for $\varepsilon = 0$ we have $\lambda = 0$ with multiplicity D. Therefore, a value of $\varepsilon \sim 10^{-15}$ due to numerical rounding errors may still produce strong numerical errors in the eigenvalues if D is sufficiently large. In our case, figure 8 shows that the eigenvalues obtained by the Arnoldi method are accurate for $|\lambda| \geqslant 10^{-2}$.

As can be seen in figure 8, there is also a difference in quality between the two initial vectors chosen for the Arnoldi method. Using a random initial vector, the Arnoldi method produces some wrong isolated eigenvalues in the intermediate regime $0.01 \leqslant |\lambda| \leqslant 0.02$ and in the case $N = 10^7$, some of the semi-analytical eigenvalues in the same regime are not accurately found. However, for uniform initial vector, the Arnoldi method produces rather accurate eigenvalues even for $|\lambda| \approx 0.005$. The reason is that the uniform initial vector corresponds (up to normalization) to the vector $v = e/N$. In view of equation (10), the Arnoldi method generates, at least in theory, exactly the l-dimensional subspace spanned by the vectors $v^{(j)}$ and should exactly break off at $n_A = l$ with a vanishing coupling matrix element from the subspace to the remaining space. However, due to numerical rounding errors and the fact that the vectors $v^{(j)}$ are badly conditioned, i.e. mathematically they are linearly independent but numerically nearly linearly dependent, the coupling matrix element is of the order of 10^{-3} (for $N = 10^7$). As a consequence, the Arnoldi method continues to generate new vectors producing a cloud of 'artificial' eigenvalues inside a circle or radius ~ 0.005. These eigenvalues are generated by the above-explained mechanism of perturbed Jordan blocks.

The Arnoldi method with a random initial vector produces a similar but slightly larger cloud of such artificial eigenvalues. However, here, even without any numerical rounding errors, the method should not break off due to a bad choice of the initial vector. Actually, in this case, the method even produces some 'bad' eigenvalues outside the Jordan-block-generated cloud.

We mention that it is possible to improve the numerical behavior of the Arnoldi method with uniform initial vector by the following 'tricks': first we chose a different matrix representation of S where the first basis vector (associated with the number '1') is replaced by the uniform vector e and second where the scalar product used for the Gram–Schmidt orthogonalization is modified with stronger weights $\sim n^2$ for the larger components. This modified Arnoldi method produces a very small coupling matrix element $\sim 10^{-10}$ (for $N = 10^7$) at $n_A = l$ and numerically very accurate eigenvalues (up to 10^{-10}) for *all* l non-vanishing eigenvalues. If we force the Arnoldi iterations to continue ($n_A \gg l$), we obtain again a Jordan-block-generated cloud of eigenvalues but whose size is considerably reduced as compared to both original variants of the method.

5. Self-consistent determination of PageRank and analytic approximation

The eigenvalue equation of the PageRank, $P = C\,v + S_0\,P$ with $C = d^T\,P$ (see equation (2)), can be interpreted as a self-consistent equation for P defining a very effective iterative method to determine P in a few iterations. Let us define the following iteration procedure:

$$P^{(0)} = 0, \quad P^{(j+1)} = C\,v + S_0\,P^{(j)}, \quad j = 0,\ 1,\ 2,\ \ldots\ . \tag{14}$$

In principle, the constant $C = d^T\,P$ is only obtained once the exact PageRank is known. Therefore, in a practical application of this iteration, one first chooses some arbitrary non-vanishing value for C and normalizes the PageRank once the procedure has converged. However, for reasons of notation, we chose to keep the value $C = d^T\,P$ in equation (14) from the very beginning.

We note that iteration (14) can formally be solved by the sum

$$P^{(j)} = C\sum_{i=0}^{j-1} S_0^i\,v = C\sum_{i=0}^{j-1} v^{(i)}. \tag{15}$$

Since $S_0^l = 0$ for $l = [\log_2(N)]$, the iteration not only converges but it actually provides the exact PageRank $P = P^{(l)}$ after a finite number of iterations when $j = l$ and in which case, equation (15) coincides with our previous result (8).

We mention that the power method, where one successively multiplies the matrix $S = v\,d^T + S_0$ by an initial (normalized) vector, is somewhat similar to (14) but with a very crucial difference. In the power method, the constant C is updated at each iteration according to $C^{(j)} = d^T\,P^{(j)}$ and here the initial vector must be different from 0. We recall that

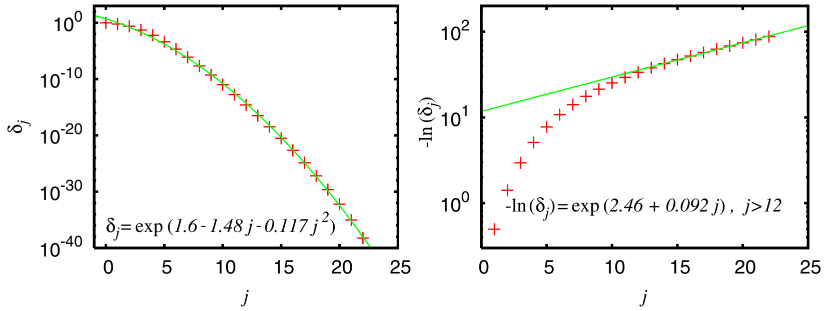

Figure 11. Decay of the quantity $\delta_j = \|P^{(j)} - P\|_1$ representing the error of the approximate PageRank $P^{(j)}$ after j iterations of equation (14) (for $N = 10^7$). The left panel shows δ_j versus j and the green line is obtained from the fit: $\ln(\delta_j) = a_3 - b_3\, j - c_3\, j^2$ with $a_3 = 1.6 \pm 0.4$, $b_3 = 1.48 \pm 0.08$ and $b_3 = 0.117 \pm 0.004$. The right panel shows $-\ln(\delta_j)$ versus j and the green line is obtained from the fit: $\ln[-\ln(\delta_j)] = a_4 + b_4\, j$ for $j > 12$ with $a_4 = 2.46 \pm 0.03$ and $b_4 = 0.092 \pm 0.002$. Note that both panels use a logarithmic representation for the vertical axis.

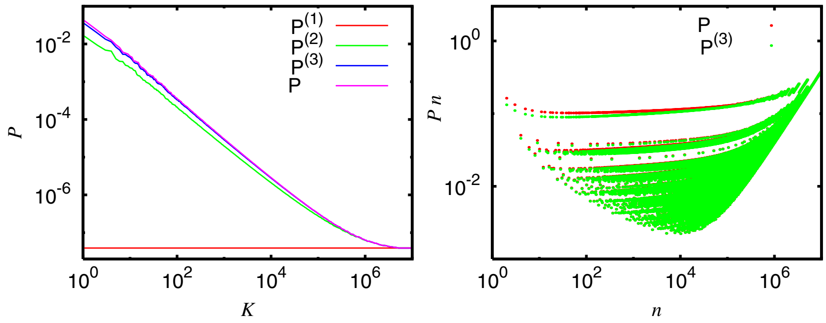

Figure 12. Left panel: comparison of the first three PageRank approximations $P^{(j)}$ for $j = 1, 2, 3$ obtained from equation (14) and the exact PageRank P versus the PageRank index K. Right panel: comparison of the dependence of the rescaled probabilities nP and $nP^{(3)}$ on n. Both panels correspond to the case $N = 10^7$.

the power method converges exponentially with an error $\sim|\lambda_1|^j$ where λ_1 being the second eigenvalue of S with $|\lambda_1| \approx 0.5$ for $N = 10^9$ and an extrapolated value $|\lambda_1| \approx 0.6$ in the limit $N \to \infty$. As can be seen in figure 11, iteration (14) actually converges much faster than $|\lambda_1|^j$, which is simply due to fixing the constant C from the beginning and not updating it with the iterations.

The norm $\delta_j = \|P^{(j)} - P\|_1$ of the error vector after j iterations decays much faster than exponentially with j as shown in figure 11. For $N = 10^7$, one can quite well approximate the error norm by the fit $\delta_j \approx \exp(1.6–1.48\, j - 0.117\, j^2)$ representing a quadratic function in the exponential. Furthermore, for j close to l, we have the approximate ratio $\delta_j/\delta_{j-1} \approx 10^{-2}$ and not 0.5–0.6 as the power method would imply. For $j > 12$, one can actually identify a regime of superconvergence where the logarithm of the error behaves exponentially, $-\ln(\delta_j) \approx \exp(2.46 + 0.092\, j)$, but the parameter range for j is too small to decide if there is really superconvergence. However, both fits clearly indicate that the convergence is considerably faster than exponential.

As a consequence of the very rapid convergence dependent on the required precision, it is sufficient to apply iteration (14) only a few times $j \ll l$ to obtain a reasonable approximation. For example, figure 12 shows for $N = 10^7$ that on a logarithmic scale, $P^{(3)}$ and P are already very close.

This allows us to obtain a very simple analytical approximation of the PageRank: $P \approx P^{(3)} = v^{(0)} + v^{(1)} + v^{(2)}$. For this, let us rewrite the recursion $v^{(j+1)} = S_0\, v^{(j)}$ in a different way:

$$v_n^{(j+1)} = \sum_{m=2}^{[N/n]} \frac{M(mn,m)}{Q(mn)}\, v_{mn}^{(j)} \quad \text{if} \quad n \geqslant 2 \quad \text{and} \quad v_1^{(j+1)} = 0, \tag{16}$$

where for given two integers n and $m > 1$, the multiplicity $M(n,m)$ is the largest integer such that $m^{M(n,m)}$ is a divisor of n and $Q(n) = \sum_{m=2}^{n-1} M(n,m)$ is the number of divisors of n (different from 1 and n itself) counting divisors several times according to their multiplicity. The appearance of the multiplicity $M(mn,n)$ in (16) is not very convenient for numerical evaluations. Either one recalculates the multiplicity at each use or one sacrifices a big amount of memory to store them. It is actually possible to rewrite equation (16) in a way that the multiplicities no longer appear explicitly. For this, we note that the case $M(mn,n) \geqslant 2$ implies only those values of m such that n is a divisor of m implying $m = \tilde{m}n$ and $mn = \tilde{m}n^2$. This produces a second sum where one uses the multiples of n^2 and in a similar way, a further sum with multiples of n^3 for the cases $M(mn,n) \geqslant 3$ and so on. For $n \geqslant 2$, we may therefore rewrite equation (16) in the following equivalent expression:

$$v_n^{(j+1)} = \sum_{m=2}^{[N/n]} \frac{1}{Q(mn)}\, v_{mn}^{(j)} + \sum_{\nu \geqslant 2}^{n^\nu \leqslant N} \sum_{m=1}^{[N/n^\nu]} \frac{1}{Q(mn^\nu)}\, v_{mn^\nu}^{(j)}, \tag{17}$$

where each term in the sum of ν takes into account the contributions with $M(mn,m) = \nu$. Note that the extra sums start at $m = 1$ since $n \geqslant 2$ and therefore $mn^\nu > n$ even for $m = 1$. The above PageRank iteration (14) can be written in a similar way (see below) but for practical purposes, numerical or analytical, it is actually more convenient to use the recurrence for the vectors $v^{(j)}$ and to add them to obtain the PageRank according to equation (15).

Both equations (16) and (17) are also very efficient for a numerical evaluation, especially in terms of memory usage, since the matrix S_0 is represented by 'only' N integer values $Q(n)$, $n = 1, \ldots, N$, which is much less than the number ($\sim N \ln N$) of non-zero double-precision matrix elements of S_0 (even completely taking into account the sparse structure of S_0). When using equation (16), one can recalculate at each time the multiplicities $M(n,m)$, which is not very expensive. However, it turns out that the additional sums in equation (17) are slightly more effective than this recalculation. Furthermore, for the iteration of $v^{(j)}$, the number of non-vanishing elements is reduced by a factor of 2 at each iteration. As a consequence, we may replace in equations (16) and (17) N by $[N\,2^{-j}]$ and thus considerably reduce the computation time. We note that the direct iteration of $P^{(j)}$ instead $v^{(j)}$ does not have this advantage. Actually, in terms of numerical computation time, the approximation to stop after a few iterations is not very important since in any case the higher order corrections require less computation time. Using iteration (17), we have been able to determine numerically the vectors $v^{(j)}$ and therefore the PageRank, the coefficients c_j and the resulting $l = [\log_2 N]$ non-zero eigenvalues of S for system sizes up to $N = 10^9$.

In addition, equation (16) allows also for some analytical approximate evaluation of the first vectors. The initial vector is $v_n^{(0)} = 1/N$. Let us try to evaluate the next two vectors $v_n^{(1)}$ and $v_n^{(2)}$ for the most important case where n is a prime number p. Furthermore, in sum (16), the most important contributions arise for m also being a prime number q such that $Q(qp) = 2$ and $M(qp,p) = 1$ (except for the case $q = p$, which we neglect) resulting in

$$v_p^{(1)} \approx \sum_{q=2,\ \text{prime}}^{[N/p]} \frac{1}{2N} = \frac{1}{2N}\,\pi\left(\left[\frac{N}{p}\right]\right) \approx \frac{1}{2p(\ln N - \ln p)}, \tag{18}$$

where $\pi(n) \approx n/\ln(n)$ (for $n \gg 1$) is the number of prime numbers below n. However, these values of $v_n^{(1)}$ at the prime numbers $n = p$ do not contribute to (16) for the next iteration $j = 1$ when trying to determine $v^{(2)}$. To obtain the leading contributions in $v^{(2)}$, we need $v_n^{(1)}$ for $n = p_1 p_2$ being a product of two prime numbers. In this case, we have $Q(q\, p_1\, p_2) = 2^3 - 2 = 6$ if $q,\ p_1$and p_2 are three different prime numbers. Assuming $p_1 \neq p_2$ and neglecting the complications from the few cases $q = p_1$ or $q = p_2$, we find that

$$v_{p_1 p_2}^{(1)} \approx \frac{1}{6N}\,\pi\left(\left[\frac{N}{p_1\, p_2}\right]\right) \approx \frac{1}{6p_1\, p_2\, (\ln N - \ln p_1 - \ln p_2)}. \tag{19}$$

For the case $n = p^2$, i.e. $p_1 = p_2 = p$, we have $Q(qp^2) = 5$ (since p has multiplicity 2) resulting in

$$v_{p^2}^{(1)} \approx \frac{1}{5N}\,\pi\left(\left[\frac{N}{p^2}\right]\right) \approx \frac{1}{5p^2\, (\ln N - 2\,\ln p)}. \tag{20}$$

From (16) for $j = 1$ and (19), we obtain

$$v_p^{(2)} \approx \frac{1}{12N} \sum_{q=2,\ \text{prime}}^{[N/(2p)]} \pi\left(\left[\frac{N}{pq}\right]\right). \tag{21}$$

Here we have reduced the sum from $q \leqslant [N/p]$ to $q \leqslant [N/(2p)]$ since $\pi([N/(pq)])$ is non-zero only for $N/(pq) \geqslant 2$ and therefore $q \leqslant N/(2p)$. Now, we replace the sum $\sum_q(\cdots)$ over the prime numbers by an integral $\int \mathrm{d}q\, \pi'(q)\, (\cdots)$ where $\pi'(q) \approx 1/\ln(q)$ is the average density of prime numbers at q resulting in

$$\begin{aligned} v_p^{(2)} &\approx \frac{1}{12N}\int_2^{N/(2p)} \mathrm{d}q\, \pi\left(\left[\frac{N}{pq}\right]\right)\, \pi'(q) \\ &\approx \frac{1}{12p}\int_2^{N/(2p)} \frac{\mathrm{d}q}{q}\, \frac{1}{(\ln(N/p) - \ln q)\ln q} \\ &= \frac{1}{12p}\int_{\ln 2}^{\ln(N/(2p))} \mathrm{d}u\, \frac{1}{(\ln(N/p) - u)\, u} \\ &= \frac{1}{6p\, \ln(N/p)}\left(\ln\ln\left(\frac{N}{2p}\right) - \ln\ln 2\right). \end{aligned} \tag{22}$$

From (18) and (22), we obtain the PageRank approximation at integer values,

$$P_p \approx P_p^{(3)} \approx C\left(\frac{1}{N} + v_p^{(1)} + v_p^{(2)}\right) \approx \frac{C}{2p\ln N}\left(1 - \ln\ln 2 + \frac{\ln\ln N}{3}\right), \tag{23}$$

where we have assumed that $N \gg p$ and replaced $\ln(N/p) = \ln N - \ln p \approx \ln N$ and C is the same constant as used in (14).

The important point with this expression is that it is of the form $P_p \approx C_N/p$ where C_N is a constant depending on N. In order to compare with our above results, especially in figure 2, we have to replace p by the K index. Assuming that the K index is dominated by the prime numbers, we have $K = \pi(p) \approx p/\ln p$ implying $p \approx K\ln p \approx K\,\ln K$, thus providing the behavior $P(K) \approx C_N/(K\,\ln K)$ already conjectured above based on the numerical results. Concerning the numerical value of the constant C_N, we find that, at $N = 10^7$, it is roughly one order of magnitude too small compared to the numerical results.

We recall that the considerations leading to expression (23) are based on a lot of assumptions and quite crude approximations, especially the replacement of $\pi(n) \approx n/\ln(n)$, even if $n = \mathcal{O}(1)$, and we have neglected a lot of contributions from numbers with more factors in their prime factor decomposition, which are most likely responsible for the reduced numerical prefactor. Furthermore, the assumption that the PageRank is dominated by prime

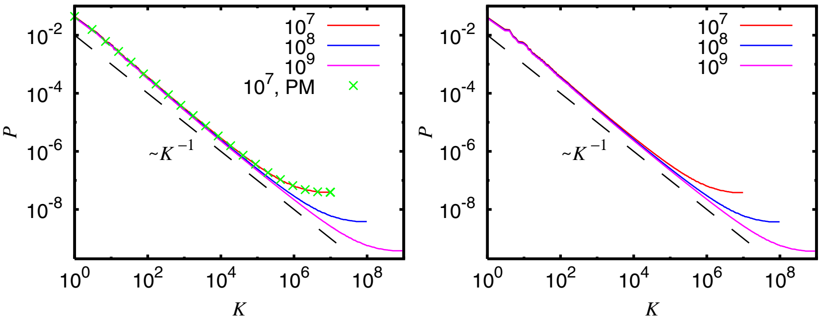

Figure 13. Left panel: the full lines correspond to the dependence of PageRank probability $P(K)$ on the PageRank index K for the matrix sizes $N = 10^7, 10^8, 10^9$ with the PageRank evaluated from expression (8) using the efficient numerical method based on equation (17). The green crosses correspond to the PageRank obtained by the power method (PM) for $N = 10^7$; the dashed straight line shows the Zipf law dependence $P \sim 1/K$. Right panel: the same as in the left panel (without data from the power method) for a simplified model for the Google matrix of integers where all multiplicities $M(n, m)$ are replaced by 1, i.e. n is linked to its divisors m only once even if n can be divided several times by m. The PageRank was numerically evaluated by the same efficient method using equations (8) and (16) with $M(n, m) = 1$.

numbers is not completely exact since certain non-prime numbers with a small number of factors intermix with larger prime numbers in the PageRank, thus modifying the dependence of the prime numbers on the K index from $p \approx K \ln(K)$ to $p \approx K(1.36 + 0.323 \ln K)$ according to the fit in figure 6 for $N = 10^7$. However, despite the approximations, we recover the leading parametric dependence of $P \sim 1/(K \ln K)$.

The PageRank dependence $P(K)$ obtained from expression (8) using the efficient numerical method based on equation (17) is shown in figure 13 (left panel) for $N = 10^7, 10^8, 10^9$. For $N = 10^7$, these data agree with the computation result by the Arnoldi power method with the numerical accuracy of the order of 10^{-10} (see also figure 10). This confirms the efficiency of our semi-analytical computation of the PageRank.

We note that it may be useful to consider a simplified model for the Google matrix of integers when multiplicity of all divisors is taken to be unity. The numerical fit of data shows that, in this case, the number of links scales as $N_\ell = N(a_\ell + b_\ell \ln N)$ with $a_\ell = -1.838 \pm 0.002$, $b_\ell = 0.999 \pm 0.0002$. For this model, we have the same expression (16) but with the replacements $M(nm, m) \to 1$ and $Q(n) \to Q^*(n)$ where $Q^*(n)$ is the number of divisors of the integer n excluding 1 and n itself without multiplicities, e.g. $Q^*(2) = 0$, $Q^*(3) = 0$, $Q^*(4) = 1, \ldots$. Note that this quantity is given by the expression $Q^*(n) = (\prod_j(\mu_j + 1)) - 2$ where μ_j are the exponents in the prime factor decomposition of $n = \prod_j p_j^{\mu_j}$.

The dependence of the PageRank on K for the simplified model is shown in the right panel of figure 13. It shows practically the same behavior as in the main model shown in the left panel. In this case, the analytical expression for the PageRank P, obtained from the first three terms, has a very simple form

$$P_n \approx P_n^{(3)} = \sigma_N \left(1 + \sum_{m_1=1}^{[N/n]} \frac{1}{Q^*(m_1 n)} + \sum_{m_1=2}^{[N/n]} \sum_{m_2=2}^{[N/(nm_1)]} \frac{1}{Q^*(m_1 n)} \frac{1}{Q^*(m_2 m_1 n)} \right), \tag{24}$$

where N is the matrix size and σ_N is the global normalization constant determined by the condition $\sum_{n=1}^{n=N} P_n = 1$. This simple formula gives a good description of the PageRank behavior shown in the right panel of figure 13. Indeed, the direct count shows that the ratio

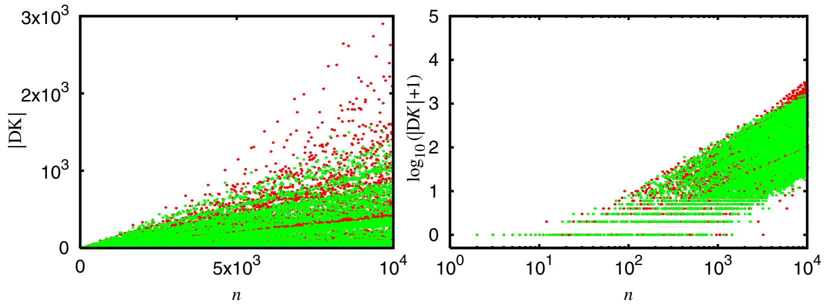

Figure 14. Dependence of $|\Delta K| = |K_n(N_2) - K_n(N_1)|$ on the integer n for matrix sizes $N_1 = 10^8, N_2 = 10^9$ (green points) and $N_1 = 10^7, N_2 = 10^8$ (red points). The left and right panels show the same data in normal and log–log scales. Note the strongly reduced vertical scale of the left panel as compared to the left panel of figure 7. The vertical scale of the right panel was not reduced allowing a direct comparison with the right panel of figure 7. The data were obtained by the same efficient numerical method as in the left panel of figure 13.

R_{ms} of the total number of links N_ℓ for both models (counted with or without multiplicities) approaches unity for large matrix sizes. For example, we have $R_{ms} = 1.184$ ($N = 1000$), 1.102 (10^5), 1.070 (10^7) and 1.052 (10^9). Thus, we think that in the limit of large N, both models converge to the same type of behavior. It is possible that the simplified model may be more suitable for further analytical analysis. However, in this work, we present data for the simplified model only in the right panel of figure 13.

Using the PageRank data obtained by the self-consistent approach for large $N = 10^7, 10^8, 10^9$, we can analyze the convergence of the PageRank order K_n at larger sizes compared to those used in figure 7. These new results for variation of $|\Delta K|$ are presented in figure 14. They show that the variation $|\Delta K|$ decreases with the increase of N from 10^7 up to 10^9 even if the process is slow. A direct comparison shows that the first deviation in the order K_n appears at $K = K_s = 13$ (comparing $N = 10^6$ versus 10^7), $K_s = 27$ (10^7 versus 10^8), $K_s = 30$ (10^8 versus 10^9). We find that the stable range interval K_s grows with N but this growth seems logarithmic like with $K_s \sim \ln N$. Such a growth seems to be natural in the view of logarithmic convergence of the second eigenvalue λ_1 discussed above and all logarithmic factors appearing in the density of primes. We also note that the value of K_s is significantly smaller than the value of K_d at which the first degenerate flat plateau appears in the PageRank $P(K)$ and hence these degeneracies do not affect the order of the first K_s integers.

On the basis of the obtained results, we conclude that for our maximal matrix size $N = 10^9$, we have convergence of the first 32 values of K_n. These numbers n, corresponding to the values of $K = 1, 2, \ldots, 32$, are $n = 2, 3, 5, 7, 4, 11, 13, 17, 6, 19, 9, 23, 29, 8, 31, 10, 37, 41, 43, 14, 47, 15, 53, 59, 61, 25, 67, 12, 71, 73, 22, 21$. There are about 30% of non-primes among these values. We mention that the positions of the first non-primes $4,\ 6,\ 9$ can already be obtained from the first-order approximations of $v^{(1)}$ discussed above. According to (18), the relative weight of a prime number in the first order is $1/(2p)$. For the two square numbers 4 and 9, the weight is according to (20) either $1/(5 \times 4) = 1/(2 \times 10)$ or $1/(5 \times 9) = 1/(2 \times 22.5)$, explaining that 4 is between the primes 7 and 11 and that 9 is between 19 and 23. For the product $6 = 2 \times 3$, we have according to (19) the weight $1/(6 \times 6) = 1/(2 \times 18)$ implying that 6 is between 17 and 19. However, this simple argument does not work for other numbers, for example, for 10 (or 14), it would imply an incorrect position between 29 and 31 (41 and 43). We mention that more numerical data are available at the web page [17].

For the simplified model, we find at $N = 10^9$ for the first values $K = 1, 2, \ldots, 32$ a slightly different order of integers $n = 2, 3, 5, 4, 7, 11, 13, 17, 9, 6, 19, 8, 23, 29, 31, 10, 37, 41, 43, 14, 47, 15, 53, 25, 59, 16, 61, 12, 67, 71, 22, 21$. Here the absence of multiplicities increases the weight for the square numbers of primes to $1/(4p^2)$, implying that these numbers are slightly advanced in the K order as compared to our main model. The modified weight for 9 is $1/(2 \times 18)$ coherent with the new position between 17 and 19 (with 6 having the same first-order weight as 9 and also being between 17 and 19). For 4, the weight is increased from $1/(2 \times 10)$ to $1/(2 \times 8)$. However, this increase is not sufficient to explain the new position of 4 between 5 and 7.

One might mention as a curiosity a special 'prime integer network model' where a non-prime number n is only linked to its prime factors (and not to all of its divisors). In this case, the matrix S_0 is strongly simplified such that $S_0^2 = 0$, i.e. $l = 2$ being independent of the system size, and hence there are only two non-vanishing eigenvalues of the Google matrix, which are $\lambda_0 = 1$ and $\lambda_1 = c_0 - 1 \approx -1 + 1/\ln N$ where $c_0 = (\pi(N) + 1)/N \approx 1/\ln N$ is the ratio of the number of primes and unity to N. This is simply seen from the definition of c_j in equation (6) and the trace $c_0 = \lambda_0 + \lambda_1$ of matrix (11), which is of size 2×2 for this case. According to (5), the PageRank P and the second eigenvector ψ_1 are given by $P \propto e + v^{(1)}$ and $\psi_1 \propto e - v^{(1)}/(1 - 1/\ln N)$ where e is the vector with all components equal to unity and $v^{(1)}$ is a vector such that $v_n^{(1)} = 0$ for the non-prime numbers n or $n = 1$ and $v_n^{(1)}$ for the prime numbers $n = p$ is given by an equation similar to equation (16) for $j = 0$ with $v_{nm}^{(0)}$ being replaced by unity and multiplicities and number of divisors adapted for the prime integer network model. Here both versions, with or without multiplicities, are possible. The eigenvalues do not depend on the version but the eigenvectors do. For both cases, it is pretty obvious that the K index gives exactly the sequence of prime numbers below N in increasing order followed by a large degenerated plateau for the non-prime integer numbers. Note that here the second eigenvalue converges to -1 with a correction $1/\ln(N)$ for large N, thus closing the gap in $|\lambda|$ of the Google matrix.

6. Discussion

In this work, we constructed the Google matrix of integers based on links between a given integer n and its divisors. The numerical analysis based on the Arnoldi method allowed us to show that the PageRank $P(K_n)$ of this directed network decays with the PageRank index K_n of an integer n approximately as $P(K_n) \sim 1/(K_n \ln K_n)$, being similar to those of the Zipf law and those found for the WWW. However, the spectrum of the Google matrix has a large gap appearing between the unit eigenvalue and other eigenvalues, while the spectrum of the Google matrix of WWW usually has no gap. We developed an efficient semi-analytical method to compute the PageRank of integers which allowed us to determine the dependence $P(K_n)$ up to the matrix size of 1 billion. We show that the dependence of PageRank on the integer number n is characterized by a series of branches corresponding to primes, semi-primes and numbers with higher products of primes. Our data show a logarithmic-like convergence of the PageRank order of integers to a fixed order in the limit of matrix size going to infinity.

Acknowledgments

This work was supported in part by the EC FET Open project 'New tools and algorithms for directed network analysis' (NADINE no 288956).

References

[1] Hardy G H and Wright E M 2008 *An Introduction to the Theory of Numbers* 6th edn (Oxford: Oxford University Press)

[2] Crandall R and Pomerance C 2005 *Prime Numbers: A Computational Perspective* (Berlin: Springer)

[3] Berry M V 1991 Some quantum-to-classical asymptotics *Les Houches Lecture Series LII (1989)* ed M-J Giannoni, A Voros and J Zinn-Justin (Amsterdam: North-Holland) p 251

[4] Berry M V and Keating J P 1999 *SIAM Rev.* **41** 236

[5] Srednicki M 2011 *Phys. Rev. Lett.* **107** 100201

[6] Markov A A 1906 *Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga (Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya* vol 15*)* p 135 (in Russian)
Howard R A *Dynamic Probabilistic Systems (Markov Models* vol 1*)* (New York: Dover)

[7] Brin S and Page L 1998 *Comput. Netw. ISDN Syst.* **30** 107

[8] Langville A M and Meyer C D 2006 *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton, NJ: Princeton University Press)

[9] Frahm K M, Georgeot B and Shepelyansky D L 2011 *J. Phys. A: Math. Theor.* **44** 465101

[10] Donato D, Laura L, Leonardi S and Millozzi S 2004 *Eur. Phys. J.* B **38** 239

[11] Pandurangan G, Raghavan P and Upfal E 2005 *Internet Math.* **3** 1

[12] Corso G 2004 *Phys. Rev.* E **69** 036106

[13] Achter J D 2004 *Phys. Rev.* E **70** 058103

[14] Stewart G W 2001 *Matrix Algorithms: Volume II. Eigensystems* (Philadelphia, PA: SIAM)

[15] Frahm K M and Shepelyansky D L 2010 *Eur. Phys. J.* B **76** 57

[16] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Boston, MA: Addison-Wesley)

[17] http://www.quantware.ups-tlse.fr/QWLIB/pagerankofintegers/

EPJ B
Condensed Matter
and Complex Systems

EPJ.org
your physics journal

Eur. Phys. J. B (2012) 85: 355

DOI: 10.1140/epjb/e2012-30599-6

Google matrix of Twitter

K.M. Frahm and D.L. Shepelyansky

edp sciences

Società Italiana
di Fisica

Springer

Eur. Phys. J. B (2012) 85: 355
DOI: 10.1140/epjb/e2012-30599-6

THE EUROPEAN
PHYSICAL JOURNAL B

Regular Article

Google matrix of Twitter

K.M. Frahm and D.L. Shepelyansky[a]

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

Received 14 July 2012 / Received in final form 14 July 2012
Published online 24 October 2012 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2012

Abstract. We construct the Google matrix of the entire Twitter network, dated by July 2009, and analyze its spectrum and eigenstate properties including the PageRank and CheiRank vectors and 2DRanking of all nodes. Our studies show much stronger inter-connectivity between top PageRank nodes for the Twitter network compared to the networks of Wikipedia and British Universities studied previously. Our analysis allows to locate the top Twitter users which control the information flow on the network. We argue that this small fraction of the whole number of users, which can be viewed as the social network elite, plays the dominant role in the process of opinion formation on the network.

1 Introduction

Twitter is an online directed social network that enables its users to exchange short communications of up to 140 characters [1]. In March 2012 this network had around 140 million active users [1]. Being founded in 2006, the size of this network demonstrates an enormously fast growth with 41 million users in July 2009 [2], only three years after its creation. The crawling and statistical analysis of the entire Twitter network, collected in July 2009, was done by the KAIST group [2] with additional statistical characteristics available at LAW DSI of Milano University[1]. This network has scale-free properties with an average power law distribution of ingoing and outgoing links[1] [2] being typical for the World Wide Web (WWW), Wikipedia and other social networks (see e.g [3–5]). In this work we use this Twitter dataset to construct the Google matrix [6,7] of this directed network and we analyze the spectral properties of its eigenvalues and eigenvectors. Even if the entire size of Twitter 2009 is very large the powerful Arnoldi method (see e.g. [8–11]) allows to obtain the spectrum and eigenstates for the largest eigenvalues.

A special analysis is performed for the PageRank vector, used in the Google search engine [6,7], and the CheiRank vector studied for the Linux Kernel network [12,13], Wikipedia articles network [5], world trade network [14] and other directed networks [15]. While the components of the PageRank vector are on average proportional to a number of ingoing links [16], the components of the CheiRank vector are on average proportional to a number of outgoing links [5,12] that leads to a two-dimensional ranking of all network nodes [15]. Thus our studies allow to analyze the spectral properties of the entire Twitter network of an enormously large size which is by one-two orders of magnitude larger compared to previous studies [5,11,13,15].

The paper is organized as follows: the construction of the Google matrix and its global structure are described in Section 2; the properties of spectrum and eigenvectors of the Google matrix of Twitter are presented in Section 3; properties of 2DRanking of Twitter network are analyzed in Section 4 and the discussion of the results is given in Section 5. Detailed data and results of our statistical analysis of the Twitter matrix are presented at the web page[2].

2 Google matrix construction

The Google matrix of the Twitter network is constructed following the standard rules described in [6,7]: we consider the elements A_{ij} of the adjacency matrix being equal to unity if a user (or node) j points to user i and zero otherwise. Then the Google matrix of the network with N users is given by

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

where the matrix S is obtained by normalizing to unity all columns of the adjacency matrix $A_{i,j}$ with at least one non-zero element, and replacing columns with only zero elements, corresponding to the dangling nodes, by $1/N$. The damping factor α in the WWW context describes the probability $(1-\alpha)$ to jump to any node for a random surfer. The value $\alpha = 0.85$ gives a good classification for WWW [7] and thus we also use this value here. The matrix G belongs to the class of Perron-Frobenius

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

[1] Twitter web data of [2] are downloaded from the web site maintained by S. Vigna, http://law.dsi.unimi.it/webdata/twitter-2010.

[2] http://www.quantware.ups-tlse.fr/QWLIB/twittermatrix/.

operators [7], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \leq \alpha$. The right eigenvector at $\lambda = 1$ gives the probability $P(i)$ to find a random surfer at site i and is called the PageRank. Once the PageRank is found, all nodes can be sorted by decreasing probabilities $P(i)$. The node 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 $K(i) = 1, 2, \ldots$

The PageRank dependence on K is well described by a power law $P(K) \propto 1/K^{\beta_{in}}$ with $\beta_{in} \approx 0.9$. This is consistent with the relation $\beta_{in} = 1/(\mu_{in} - 1)$ corresponding to the average proportionality of PageRank probability $P(i)$ to its in-degree distribution $w_{in}(k) \propto 1/k^{\mu_{in}}$ where $k(i)$ is a number of ingoing links for a node i [7,16]. For the WWW it is established that for the ingoing links $\mu_{in} \approx 2.1$ (with $\beta_{in} \approx 0.9$) while for the out-degree distribution w_{out} of outgoing links the power law has the exponent $\mu_{out} \approx 2.7$ [3,4]. Similar values of these exponents are found for the WWW British university networks [11], the procedure call network of Linux Kernel software introduced in [12] and for Wikipedia hyperlink citation network of English articles (see e.g. [5]).

In addition to the Google matrix G we also analyze the properties of matrix G^* constructed from the network with inverted directions of links, with the adjacency matrix $A_{i,j} \to A_{j,i}$. After the inversion of links the Google matrix G^* is constructed via the procedure (1) described above. The right eigenvector at unit eigenvalue of the matrix G^* is called the CheiRank [5,12]. In analogy with the PageRank the probability values of CheiRank are proportional to number of outgoing links, due to links inversion. All nodes of the network can be ordered in a decreasing order with the CheiRank index $K^*(i)$ with $P^* \propto 1/K^{*\beta_{out}}$ with $\beta_{out} = 1/(\mu_{out} - 1)$. Since each node i of the network is characterized both by PageRank $K(i)$ and CheiRank $K^*(i)$ indexes the ranking of nodes becomes two-dimensional. While PageRank highlights well-know popular nodes, CheiRank highlights communicative nodes. As discussed in [5,12,15], such 2DRanking allows to characterize an information flow on networks in a more efficient and rich manner. It is convenient to characterize the interdependence between PageRank and CheiRank vectors by the correlator

$$\kappa = N \sum_{i=1}^{N} P(K(i))P^*(K^*(i)) - 1. \tag{2}$$

As it is shown in [12,15], we have $\kappa \approx 0$ for Linux Kernel network, transcription gene networks and $\kappa \approx 2{-}4$ for University and Wikipedia networks.

In this work we apply the Google matrix analysis developed in [5,11–15] to the Twitter 2009 network available at[1] [2]. The total size of the Google matrix is $N = 41\,652\,230$ and the number of links is $N_\ell = 1\,468\,365\,182$. This matrix size is by one-two orders of magnitude larger than those studied in [11,13,15]. The number of links per node is $\xi_\ell = N_\ell/N \approx 35$ being by a factor $1.5{-}3.5$ larger than for Wikipedia network or Cambridge University 2006 network [15]. The matrix elements of G and G^* are shown in Figure 1 on a scale of top 200 (top panels) and 400 (middle panels) values of K (for G) and K^* (for G^*) and in a coarse grained image for the whole matrix size scale (bottom panels).

It is interesting to note that the coarse-grained image has well visible hyperbolic onion curves of high density which are similar to those found in [15] for Wikipedia and University networks. In [15] the appearance of such curves was attributed to existence of specific categories. We assume that for the Twitter network such curves are a result of enhanced links between various categories of users (e.g. actors, journalists, etc.) but a detailed origin is still to be established.

In the following sections we also compare the properties of the Twitter network with those of the Wikipedia articles network from [5]. Some spectral properties of the Wikipedia network with $N = 3\,282\,257$ nodes and $N_\ell = 71\,012\,307$ links are analyzed in [11,15]. We also compare certain parameters with the networks of Cambridge and Oxford Universities of 2006 with $N = 212\,710$ and $N = 200\,823$ nodes and with $N_\ell = 2\,015\,265$ and $N_\ell = 1\,831\,542$ links respectively. The properties of these networks are discussed in [11,15]. The gallery of the Google matrix G images for these networks, as well as for the Linux Kernel network, are presented in [15]. The comparison with the data shown in Figure 1 here shows that for the Twitter network we have much stronger interconnection matrix at moderate K values. We return to this point in Sections 4 and 5.

3 Spectrum and eigenstates of Twitter

To obtain the spectrum of the Google matrix of Twitter we use the Arnoldi method [8–10]. However, at first, following the approach developed in [11], we determine the invariant subspaces of the Twitter network. For that for each node we find iteratively the set of nodes that can be reached by a chain of non-zero matrix elements of S. Usually, there are several such invariant isolated subsets and the size of such subsets is smaller than the whole matrix size. These subsets are invariant with respect to applications of matrix 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, it provides a subdivision of network nodes in N_c core space nodes (typically $70{-}80\%$ of N for British University networks [11]) and N_s subspace nodes belonging to at least one of the invariant subspaces V_j inducing the block triangular structure,

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix}. \tag{3}$$

Here the subspace-subspace block S_{ss} is actually composed of many diagonal blocks for each of the invariant subspaces. Each of these blocks corresponds to a column sum normalized matrix of the same type as G and has

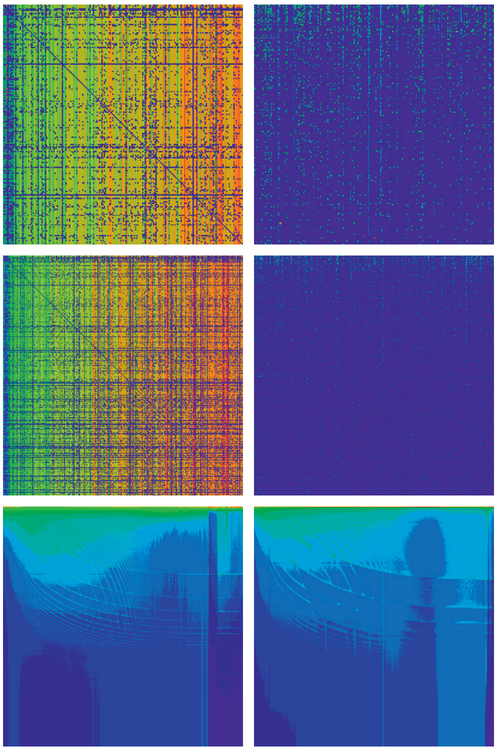

Fig. 1. (Color online) Google matrix of Twitter: matrix elements of G (left column) and G^* (right column) are shown in the basis of PageRank index K (and K') of matrix $G_{KK'}$ (left column panels) and in the basis of CheiRank index K^* (and $K^{*'}$) of matrix $G^*{}_{K^*K^{*'}}$ (right column panels). Here, x (and y) axis shows K (and $K^{'}$) (left column) (and respectively K^* and $K^{*'}$ on right column) with the range $1 \leq K, K' \leq 200$ (top panels); $1 \leq K, K' \leq 400$ (middle panels); $1 \leq K, K' \leq N$ (bottom panels). All nodes are ordered by PageRank index K of the matrix G and thus we have two matrix indexes K, K' for matrix elements in this basis (left column) and respectively $K^*, K^{*'}$ for matrix G^* (right column). Bottom panels show the coarse-grained density of matrix elements $G_{K,K'}$ and $G^*_{K^*K^{*'}}$; the coarse graining is done on 500×500 square cells for the entire Twitter network. We use a standard matrix representation with $K = K' = 1$ on top left panel corner (left column) and respectively $K^* = K^{*'} = 1$ (right column). Color shows the amplitude of matrix elements in top and middle panels or their density in the bottom panels changing from blue for minimum zero value to red at maximum value. Here the PageRank index K (and CheiRank index K^*) has been calculated for the damping factor $\alpha = 0.85$. However, the matrix elements G are shown for the damping factor $\alpha = 1$ since a value $\alpha < 1$ only adds a uniform background value and modifies the overall scale in the density plots.

therefore at least one unit eigenvalue thus explaining the high degeneracy. Its eigenvalues and eigenvectors are easily accessible by numerical diagonalization (for full matrices) thus allowing to count the number of unit eigenvalues.

We find for the G matrix of Twitter 2009 that there are $N_s = 40\,307$ subset sites with a maximal subspace dimension of 44 (most subspaces are of dimension 2 or 3). For the matrix G^* we find $N_s = 180\,414$ also with a lot of subspaces of dimension 2 or 3 and a maximal subspace dimension of 2959. The remaining eigenvalues of S can be obtained from the projected core block S_{cc} which is not column sum normalized (due to non-zero matrix elements in the block S_{sc}) and has therefore eigenvalues strictly inside the unit circle $|\lambda_j^{(\mathrm{core})}| < 1$. We have applied the Arnoldi method (AM) [8–10] with Arnoldi dimension $n_A = 640$ to determine the largest eigenvalues of S_{cc} which required a machine with 250 GB of physical RAM memory to store the non-zero matrix elements of S and the 640 vectors of the Krylov space.

In general the Arnoldi method provides numerically accurate values for the largest eigenvalues (in modulus) but their number depends crucially on the Arnoldi dimension. In our case there is a considerable density of real eigenvalues close to the points 1 and -1 where convergence is rather difficult. Comparing the results for different values of n_A, we find that for the matrix S (S^*) the first 200 (150) eigenvalues are correct within a relative error below 0.3% while the marjority of the remaining eigenvalues with $|\lambda_j| \geq 0.5$ ($|\lambda_j| \geq 0.6$) have a relative error of 10%. However, the well isolated complex eigenvalues, well visible in Figure 2, converge much better and are numerically accurate (with an error $\sim 10^{-14}$). The first three core space eigenvalues of S (S^*) are also numerically acurrate with an error of $\sim 10^{-14}$ ($\sim 10^{-8}$).

The composed spectrum of subspaces and core space eigenvalues obtained by the Arnoldi method is shown in Figure 2 for G and G^*. The obtained results show that the fraction of invariant subspaces with $\lambda = 1$ ($g_1 = N_s/N \approx 10^{-3}$) is by orders of magnitude smaller than the one found for British Universities ($g_1 \approx 0.2$ at $N \approx 2 \times 10^5$) [11]. We note that the cross and triple-star structures are visible for Twitter spectrum in Figure 2 but they are significantly less pronounced as compared to the case of Cambridge and Oxford network spectrum (see Fig. 2 in [11]). It is interesting that such a triplet and cross structures naturally appear in the spectra of random unistochastic matrices of size $N = 3$ and 4 which have been analyzed analytically and numerically in [17]. A similar star-structure spectrum appears also in sparse regular graphs with loops studied recently in [18] even if in the later case the spectrum goes outside of unit circle. This shows that even in large size networks the loop structure between 3 or 4 dominant types of nodes is well visible for University networks. For Twitter network it is less pronounced probably due to a larger number ξ_ℓ of links per node. At the same time a circle structure in the spectrum remains well visible both for Twitter and University networks. The integrated number of eigenvalues as a function of $|\lambda|$ is shown in the bottom panels of Figure 2. Further detailed analysis is required for a better understanding of the origin of such spectral structures.

It is interesting to note that a circular structure, formed by eigenvalues λ_i with $|\lambda_i|$ being close to unity (see red and blue point in top left and right panels of Fig. 3), is rather similar to those appearing in the Ulam networks of intermittency maps (see Fig. 4 in [19]). Following an analogy with the dynamics of these one-dimensional maps

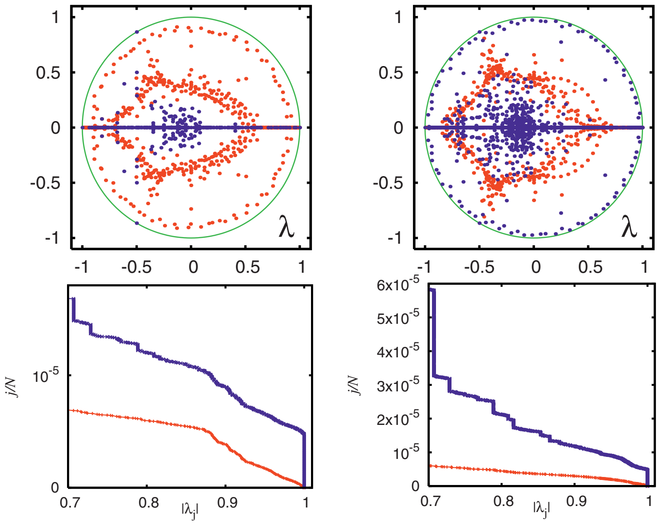

Fig. 2. (Color online) Spectrum of the Twitter matrix S (S^* with inverted direction of links) for the Twitter network shown on left panels (right panels). Top panel: subspace eigenvalues (blue dots) and core space eigenvalues (red dots) in λ-plane (green curve shows unit circle); there are 17 504 (66 316) invariant subspaces, with maximal dimension 44 (2959) and the sum of all subspace dimensions is $N_s = 40\,307$ (180 414). The core space eigenvalues are obtained from the Arnoldi method applied to the core space subblock S_{cc} of S with Arnoldi dimension 640 as explained in reference [11]. Bottom panels: fraction j/N of eigenvalues with $|\lambda| > |\lambda_j|$ for the core space eigenvalues (red bottom curve) and all eigenvalues (blue top curve) from raw data of top panels. The number of eigenvalues with $|\lambda_j| = 1$ is 34135 (129 185) of which 17505 (66 357) are at $\lambda_j = 1$; this number is (slightly) larger than the number of invariant subspaces which have each at least one unit eigenvalue. Note that in the bottom panels the number of eigenvalues with $|\lambda_j| = 1$ is artificially reduced to 200 in order to have a better scale on the vertical axis. The correct number of those eigenvalues corresponds to $j/N = 8.195 \times 10^{-4}$ (3.102×10^{-3}) which is strongly outside the vertical panel scale.

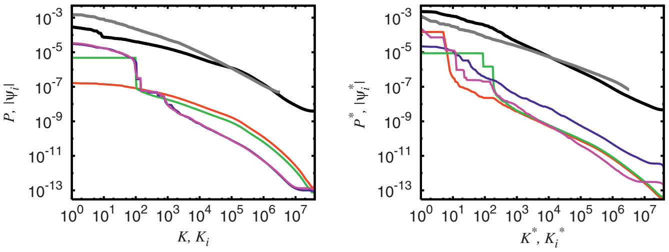

Fig. 3. (Color online) The left (right) panel shows the PageRank P (CheiRank P^*) versus the corresponding rank index K (K^*) for the Google matrix of Twitter at the damping parameter $\alpha = 0.85$ (thick black curve); for comparison the PageRank (CheiRank) of the Google matrix of Wikipedia network [5] is shown by the gray curve at same α. The colored thin curves (shifted down by factor 1000 for clarity) show the modulus of four core space eigenvectors $|\psi_i|$ ($|\psi_i^*|$) of S (S^*) versus their own ranking indexes K_i (K_i^*). Red and green lines are the eigenvectors corresponding to the two largest core space eigenvalues (in modulus) $\lambda_1 = 0.99997358$, $\lambda_2 = 0.99932634$ ($\lambda_1 = 0.99997002$, $\lambda_2 = 0.99994658$); blue and pink lines are the eigenvectors corresponding to the two complex eigenvalues $\lambda_{151} = 0.09032572 + i\,0.90000530$, $\lambda_{161} = -0.47504961 + i\,0.76576321$ ($\lambda_{457} = 0.38070896 + i\,0.39207668$, $\lambda_{105} = -0.45794117 + i\,0.80825210$). Eigenvalues and eigenvectors are obtained by the Arnoldi method with Arnoldi dimension 640 as for the data in Figure 2.

we may say that the eigenstates related to such a circular structure corresponds to quasi-isolated communities, being similar to orbits in a vicinity of intermittency region, where the information circulates mainly inside the community with only a very little flow outside of it.

The eigenstates of G and G^* with $|\lambda|$ being unity or close to unity are shown in Figure 3. For the PageRank P (CheiRank P^*) we compare its dependence on the corresponding index K (K^*) with the PageRank (CheiRank) of the Wikipedia network analyzed in [5,11,15] which size N (number of links N_ℓ) is by a factor of 10 (20) smaller. Surprisingly we find that the PageRank $P(K)$ of Twitter, approximated by the algebraic decay $P(K) = a/K^\beta$, has a slower drop as compared to Wikipedia case. Indeed, we have $\beta = 0.540 \pm 0.004$ ($a = 0.00054 \pm 0.00002$) for the PageRank of Twitter in the range $1 \le \log_{10} K \le 6$ (similar value as in [20] for the range $\log_{10} K \le 5.5$) while we have $\beta = 0.767 \pm 0.0005$ ($a = 0.0086 \pm 0.00035$) for the same range of PageRank of Wikipedia network. Also we have a sharper drop of CheiRank with $\beta = 0.857 \pm 0.003$ ($a = 0.0148 \pm 0.0004$) compared to those of PageRank of Twitter while for CheiRank of Wikipedia network we find an opposite tendency ($\beta = 0.620 \pm 0.001, a = 0.0015 \pm 0.00002$) in the same index range. Thus for Twitter network the PageRank is more delocalized compared to CheiRank (e.g. $P(1) < P^*(1)$) while usually one has the opposite relation (e.g. for Wikipedia $P(1) > P^*(1)$). We attribute this to the enormously high inter-connectivity between the top PageRank nodes $K \le 10^4$ which is well visible in Figure 1.

We should also point out a specific property of PageRank and CheiRank vectors which has been already noted in [21]: there are some degenerate plateaus in $P(K(i))$ or $P^*(K^*(i))$ with absolutely the same values of P or P^* for a few nodes. For example, for the Twitter network we have the appearance of the first degenerate plateau at $P = 7.639 \times 10^{-7}$ for $196489 \le K \le 196491$. As a result the PageRank index K can be ordered in various ways. We attribute this phenomenon to the fact that the matrix elements of G are composed from rational elements that leads to such type of degeneracy. However, the sizes of such degenerate plateaus are relatively short and they do not influence significantly the PageRank order. Indeed, on large scales the curves of $P(K)$, $P^*(K^*)$ are rather smooth being characterized by a finite slope (see Fig. 3). Similar type of degenerate plateaus exits for networks of Wikipedia, Cambridge and Oxford Universities.

Other eigenvectors of G and G^* of Twitter network are shown by color curves in Figure 3. We see that the shape of eigenstates with λ_1 and λ_2, shown as a function of their monotonic decrease index K_i, is well pronounced in $P(K)$. Indeed, these vectors have a rather small gap separating

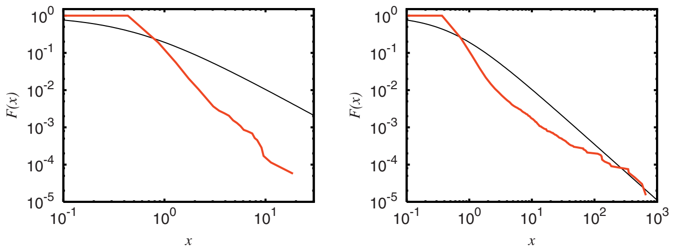

Fig. 4. (Color online) Fraction of invariant subspaces F with dimensions larger than d as a function of the rescaled variable $x = d/\langle d\rangle$, where $\langle d\rangle$ is the average subspace dimension. Left (right) panel corresponds to the matrix S (S^*) for the Twitter network (thick red curve) with $\langle d\rangle = 2.30$ (2.72). The tail can be fitted for $x \geq 0.5$ ($x \geq 10$) by the power law $F(x) = a/x^b$ with $a = 0.092 \pm 0.011$ and $b = 2.60 \pm 0.07$ ($a = 0.0125 \pm 0.0008$ and $b = 0.94 \pm 0.02$). The thin black line is $F(x) = (1 + 2x)^{-1.5}$ which corresponds to the universal behavior of $F(x)$ found in reference [11] for the WWW of British university networks.

them from unity ($|\Delta\lambda| \sim 2 \times 10^{-5}$) and thus they significantly contribute to the PageRank at $\alpha = 0.85$. At the same time we note that the gap values are significantly smaller than those for certain British Universities (see e.g. Fig. 4 in [11]). We argue that a larger number of links ξ_ℓ for Twitter is at the origin of moderate spectral gap between the core space spectrum and $\lambda = 1$. The eigenvectors of G^* have less slope variations and their decay is rather similar to the decay of CheiRank vector $P^*(K^*)$.

Finally, in Figure 4 we use the approach developed in [11] and analyze the dependence of the fraction of invariant subspaces $F(x)$ with dimensions larger than d on the rescaled variable $x = d/\langle d\rangle$ where $\langle d\rangle$ is the average subspace dimension. In [11] it was found that the British University networks are characterized by a universal functional distribution $F(x) = 1/(1 + 2x)^{3/2}$. For the Twitter network we find significant deviations from such a dependence as it is well seen in Figure 4. The tail can be fitted by the power law $F(x) \sim x^{-b}$ with the exponent $b = 2.60$ for G and $b = 0.94$ for G^*. It seems that with the increase of number of links per node ξ_ℓ we start to see deviations from the above universal distribution: it is visible for Wikipedia network (see Fig. 7 in [11]) and becomes even more pronounced for the Twitter network. We assume that a large value of ξ_ℓ for Twitter leads to a change of the percolation properties of the network generating other type of distribution F which properties should be studied in more detail in further.

4 CheiRank versus PageRank of Twitter

As discussed in [5,12,15] each network node i has its own PageRank index $K(i)$ and CheiRank index $K^*(i)$ and, hence, the ranking of network nodes becomes a two-dimensional (2DRanking). The distribution of Twitter nodes in the PageRank-CheiRank plane (K, K^*) is shown in Figure 5 (left column) in comparison to the case of the Wikipedia network from [5,15] (right column). There are much more nodes inside the square of size $K, K^* \leq 1000$ for Twitter as compared to the case of Wikipedia. For the squares of larger sizes the densities become comparable. The global logarithmic density distribution is shown in the bottom panels of Figure 5 for both networks. The two densities have certain similarities in their distributions: both have a maximal density along a certain ridge along a line $\ln K^* = \ln K +$ const. However, for the Twitter network we have a significantly larger number of nodes at small values $K, K^* < 1000$ while in the Wikipedia network this area is practically empty.

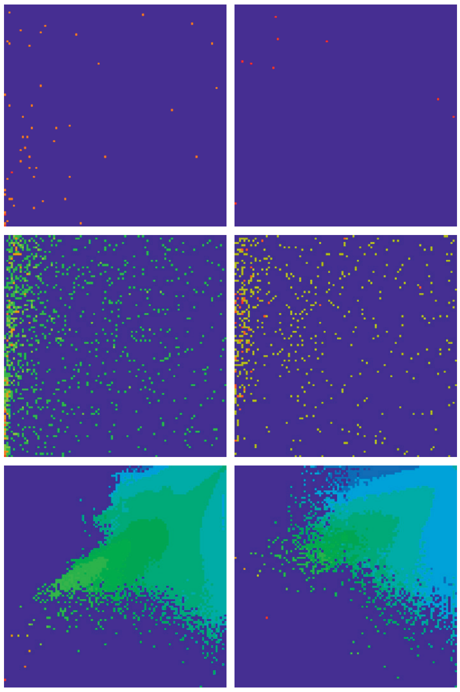

Fig. 5. (Color online) Density of nodes $W(K, K^*)$ on PageRank-CheiRank plane (K, K^*) for Twitter (left panels) and Wikipedia (right panels). Top panels show density in the range $1 \leq K, K^* \leq 1000$ with averaging over cells of size 10×10; middle panels show the range $1 \leq K, K^* \leq 10^4$ with averaging over cells of size 100×100; bottom panels show density averaged over 100×100 logarithmically equidistant grids for $0 \leq \ln K, \ln K^* \leq \ln N$, 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).

The striking difference between the Twitter and Wikipedia networks is in the number of points N_K, located inside a square area of size $K \times K$ in the PageRank-CheiRank plane. This is directly illustrated in Figure 6: at $K = 500$ there are 40 times more nodes for Twitter, at $K = 1000$ we have this ratio around 6. We note that a similar dependence N_K was studied in [15] for Wikipedia, British Universities and Linux Kernel networks (see Fig. 8 there), where in all cases the initial growth of N_K was significantly smaller as compared to the Twitter network considered here.

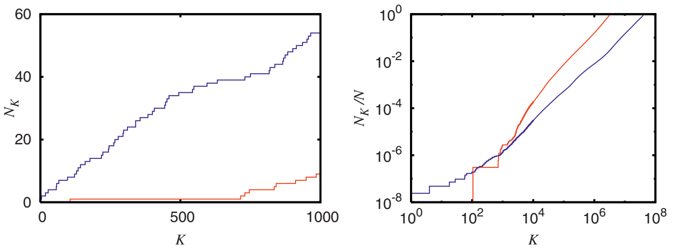

Fig. 6. (Color online) Dependence of number of nodes N_K, counted inside the square of size $K \times K$ on PageRank-CheiRank plane, on K for Twitter (blue curve) and Wikipedia (red curve); left panel shows data for $1 \leq K \leq 1000$ in linear scale, right panel shows data in log-log scale for the whole range of K.

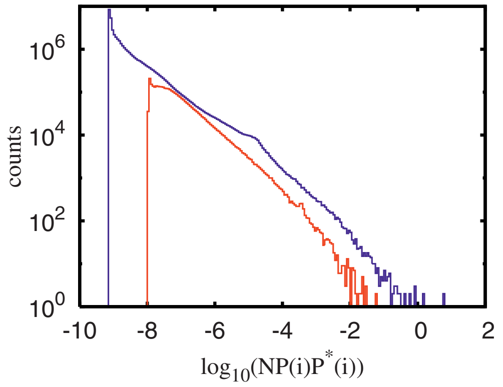

Fig. 7. (Color online) Histogram of frequency appearance of correlator components $\kappa_i = NP(K(i))P^*(K^*(i))$ for networks of Twitter (blue) and Wikipedia (red). For the histogram the whole interval $10^{-10} \leq \kappa_i \leq 10^2$ is divided in 240 cells of equal size in logarithmic scale.

Another important characteristics of 2DRanking is the correlator κ (2) between PageRank and CheiRank vectors. We find for Twitter the value $\kappa = 112.60$ which is by a factor 30–60 larger compared to this value for Wikipedia (4.08), Cambridge and Oxford University networks of 2006 considered in [5,11,15]. The origin of such a large value of κ for the Twitter network becomes more clear from the analysis of the distribution of individual node contributions $\kappa_i = NP(K(i))P^*(K^*(i))$ in the correlator sum (2) shown in Figure 7. We see that there are certain nodes with very large κ_i values and even if there are only few of them still they give a significant contribution to the total correlator value. We note that there is a similar feature for the Cambridge University network in 2011 as discussed in [15] even if there one finds a smaller value $\kappa = 30$. Thus we see that for certain nodes we have strongly correlated large values of $P(K(i))$ and $P^*(K^*(i))$ explaining the largest correlator value κ among all networks studied up to now. We will argue below that this is related to a very strong inter-connectivity between top K PageRank users of the Twitter network.

5 Discussion

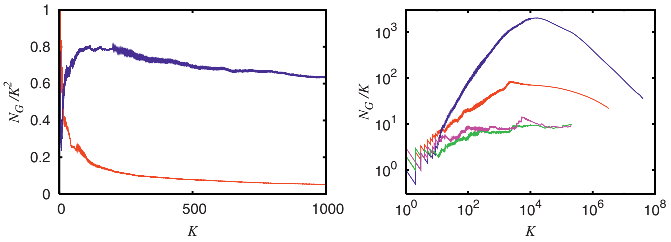

Fig. 8. (Color online) Left panel: dependence of the area density $g_K = N_G/K^2$ of nonzero elements of the adjacency matrix among top PageRank nodes on the PageRank index K for Twitter (blue curve) and Wikipedia (red curve) networks, data are shown in linear scale. Right panel: linear density N_G/K of same matrix elements shown for the whole range of K in log-log scale for Twitter (blue curve), Wikipedia (red curve), Oxford University 2006 (magenta curve) and Cambridge University 2006 (green curve) (curves from top to bottom at $K = 100$).

In this work we study the statistical properties of the Google matrix of Twitter network including its spectrum, eigenstates and 2DRanking of PageRank and CheiRank vectors. The comparison with Wikipedia shows that for Twitter we have much stronger correlations between PageRank and CheiRank vectors. Thus for the Twitter network there are nodes which are very well known by the community of users and at the same time they are very communicative being strongly connected with top PageRank nodes. We attribute the origin of this phenomenon to a very strong connectivity between top K nodes for Twitter as compared to the Wikipedia network. This property is illustrated in Figure 8 where we show the number of nonzero elements N_G of the Google matrix, taken at $\alpha = 1$ and counted in the top left corner with indexes being smaller or equal to K (elements in columns of dangling nodes are not taken into account). We see that for $K \leq 1000$ we have for Twitter the 2D density of nonzero elements to be on a level of 70% while for Wikipedia this density is by a factor 10 smaller. For these two networks the dependence of N_G on K at $K \leq 1000$ is well described by a power law $N_G = aN^b$ with $a = 0.72 \pm 0.01$, $b = 1.993 \pm 0.002$ for Twitter and $a = 2.10 \pm 0.01$, $b = 1.469 \pm 0.001$ for Wikipedia. Thus for Twitter the top $K \leq 1000$ elements fill about 70% of the matrix and about 20% for size $K \leq 10^4$. For Wikipedia the filling factor is smaller by a factor 10–20. An effective number of links per node for top K nodes is given by the ratio N_G/K which is equal to ξ_ℓ at $K = N$. The dependence of this ratio on K is shown in Figure 8 in right panel. We see a striking difference between Twitter network and networks of Wikipedia, Cambridge and Oxford Universities. For Twitter the maximum value of N_G/K is by two orders of magnitude larger as compared to the Universities networks, and by a factor 20 larger than for Wikipedia. Thus the Twitter network is characterized by a very strong connectivity between top PageRank nodes which can be considered as the Twitter elite [20].

It is interesting to note that for $K \leq 20$ the Wikipedia network has a larger value of the ratio N_G/K^2 compared to the Twitter network, but the situation is changed for larger values of $K > 20$. In fact the first top 20 nodes of Wikipedia network are mainly composed from world countries (see [5]) which are strongly interconnected due to historical reasons. However, at larger values of K Wikipedia starts to have articles on various subjects and the ratio N_G/K^2 drops significantly. On the other hand, for the Twitter network we see that a large group of very important persons (VIP) with $K < 10^4$ is strongly interconnected. This dominant VIP structure has certain similarities with the structure of transnational corporations and their ownership network dominated by a small tightly-knit core of financial institutions [22]. The existence of a solid phase of industrially developed, strongly linked countries is also established for the world trade network obtained from the United Nations COMTRADE data base [23]. It is possible that such super concentration of links between top Twitter users results from a global increase of number of links per node characteristic for such type of social networks. Indeed, the recent analysis of the Facebook network shows a significant decrease of degree of separation during the time evolution of this network [24]. Also the number of friendship links per node reaches as high value as $\xi_\ell \approx 100$ at the current Facebook snapshot (see Tab. 2 in [24]). This significant growth of ξ_ℓ during the time evolution of social networks leads to an enormous concentration of links among society elite at top PageRank users and may significantly influence the process of strategic decisions on such networks in the future. The growth of ξ_ℓ leads also to a significant decrease of the exponent β of algebraic decay of PageRank which is known to be $\beta \approx 0.9$ for the WWW (see e.g. [3,4,7]) while for the Twitter network we find $\beta \approx 0.5$ (see also [20]). This tendency may be a precursor of a delocalization transition of the PageRank vector emerging at a large values of ξ_ℓ. Such a delocalization would lead to a flat PageRank probability distribution and a strong drop of the efficiency of the information retrieval process. It is known that for the Ulam networks of dynamical maps such a delocalization indeed takes place under certain conditions [19,25].

Our results show that the strong inter-connectivity of VIP users with about top 1000 PageRank indexes dominates the information flow on the network. This result is in line with the recent studies of opinion formation of the Twitter network [20] showing that the top 1300 PageRank users of Twitter can impose their opinion for the whole network of 41 million size. Thus we think that the statistical analysis presented here plays a very important role for a better understanding of decision making and opinion formation on the modern social networks.

The present size of the Twitter network is by a factor 3.5 larger as compared to its size in 2009 analyzed in this work. Thus it would be very interesting to extend the present analysis to the current status of the Twitter network which now includes all layers of the world society. Such an analysis will allow to understand in an better way the process of information flow and decision making on social networks.

This work is supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No. 288956). We thank S.Vigna for providing us a friendly access to the Twitter dataset[1] [2]. We also acknowledge the France-Armenia collaboration grant CNRS/SCS No. 24943 (IE-017) on "Classical and quantum chaos".

References

1. Wikipedia (The Free Encyclopedia) Twitter, `http://en.wikipedia.org/wiki/Twitter` (2012)
2. H. Kwak, C. Lee, H. Park, S. Moon, *Proc. 19th Int. Conf. WWW2010* (ACM, New York, 2010), p. 591
3. D. Donato, L. Laura, S. Leonardi, S. Millozzi, Eur. Phys. J. B **38**, 239 (2004)
4. G. Pandurangan, P. Raghavan, E. Upfal, Internet Math. **3**, 1 (2005)
5. A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, Eur. Phys. J. B **77**, 523 (2010)
6. S. Brin, L. Page, Comput. Netw. ISDN Syst. **30**, 107 (1998)
7. A.M. Langville, C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, Princeton, 2006)
8. G.W. Stewart, *Matrix Algorithms Eigensystems* (SIAM, 2001), Vol. II
9. G.H. Golub, C. Greif, BIT Num. Math. **46**, 759 (2006)
10. K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **76**, 57 (2010)
11. K.M. Frahm, B. Georgeot, D.L. Shepelyansky, J. Phys. A: Math. Theor. **44**, 465101 (2011)
12. A.D. Chepelianskii, `arXiv:1003.5455[cs.SE]` (2010)
13. L. Ermann, A.D. Chepelianskii, D.L. Shepelyansky, Eur. Phys. J. B **79**, 115 (2011)
14. L. Ermann, D.L. Shepelyansky, Acta Phys. Polonica A **120**, A158 (2011)
15. L. Ermann, A.D. Chepelianskii, D.L. Shepelyansky, J. Phys. A: Math. Theor. **45**, 275101 (2012)
16. N. Litvak, W.R.W. Scheinhardt, Y. Volkovich, Lect. Notes Comput. Sci. **4936**, 72 (2008)
17. K. Zyczkowski, M. Kus, W. Slomczynski, H.-J. Sommers, J. Phys. A: Math. Gen. **36**, 3425 (2003)
18. F.L. Metz, I. Neri, D. Bolle, Phys. Rev. E **84**, 055101(R) (2011)
19. L. Ermann, D.L. Shepelyansky, Phys. Rev. E **81**, 036221 (2010)
20. V. Kandiah, D.L. Shepelyansky, Physica A **391**, 5779 (2012)
21. K.M. Frahm, A.D. Chepelianskii, D.L. Shepelyansky, J. Phys. A: Math. Theor. **45**, 405101 (2012)
22. S. Vitali, J.B. Glattfelder, S. Battiston, PLoS ONE **6**, e25995 (2011)
23. L. Ermann, D.L. Shepelyansky, Acta Phys. Polonica A **120**, A158 (2011)
24. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, `arXiv:1111.4570v3 [cs.SI]` (2012)
25. D.L. Shepelyansky, O.V. Zhirov, Phys. Rev. E **81**, 036213 (2010)

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 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 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. 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]:

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

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

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

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

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

	N_s	N_d	$d_{\max}$	$N_{\text{circ.}}$	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

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 $A_{ij} \rightarrow A_{ji}$. 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.

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:

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

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),

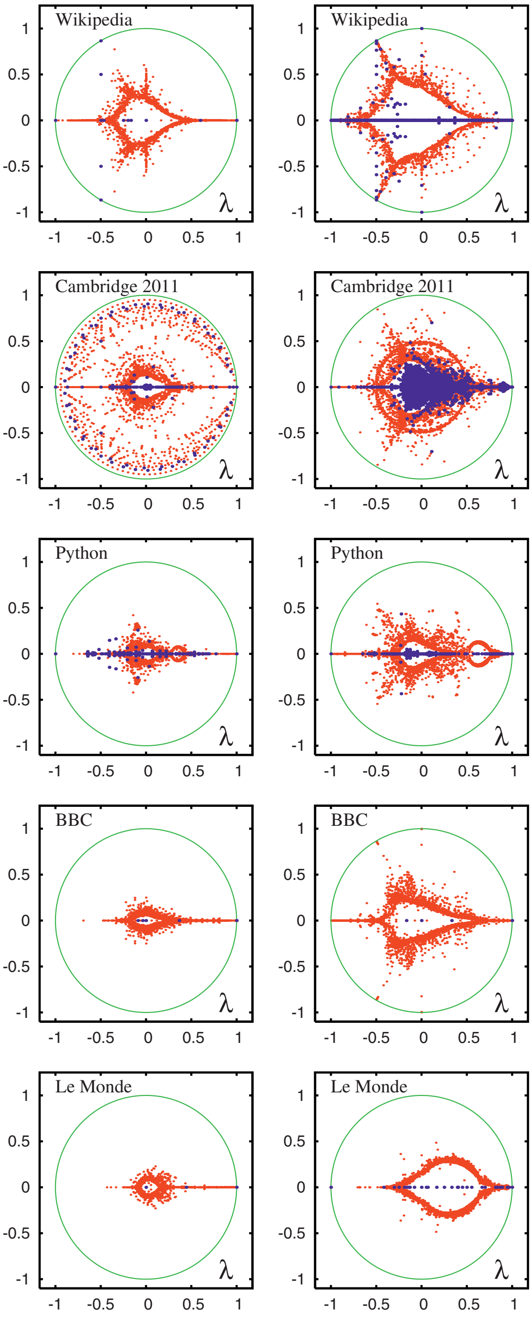

Fig. 1. Spectrum of eigenvalues λ the Google matrices G (left column) and G^* (right column) for Wikipedia, Cambridge 2011, Python, BBC and Le Monde ($\alpha = 1$). Red dots are core space eigenvalues, blue dots are subspace eigenvalues and the full green curve shows the unit circle. The core space eigenvalues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

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	$\lambda_{2474} = 0.17637 + i\,0.40917$

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 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

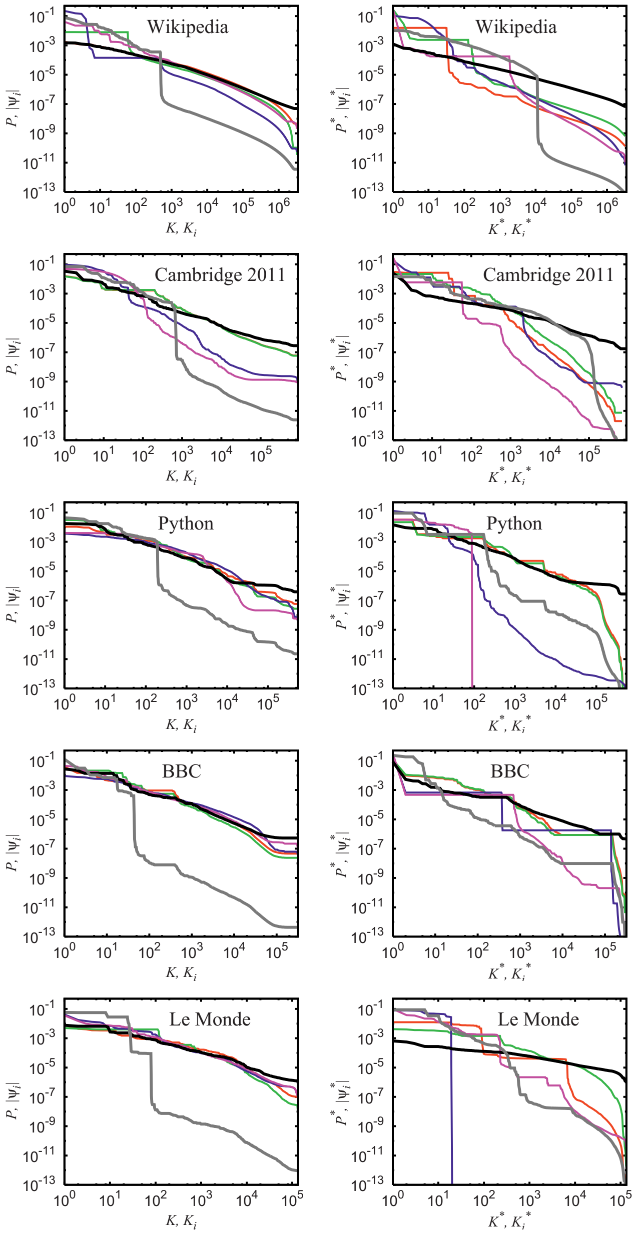

Fig. 2. PageRank P (left column) and CheiRank P^* (right column) vectors are shown as a function of the corresponding rank indexes K or K^* for the Google matrices of Wikipedia, Cambridge 2011, Python, BBC and Le Monde at the damping parameter $\alpha = 0.85$ (thick black curve) and $\alpha = 1 - 10^{-8}$ (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.

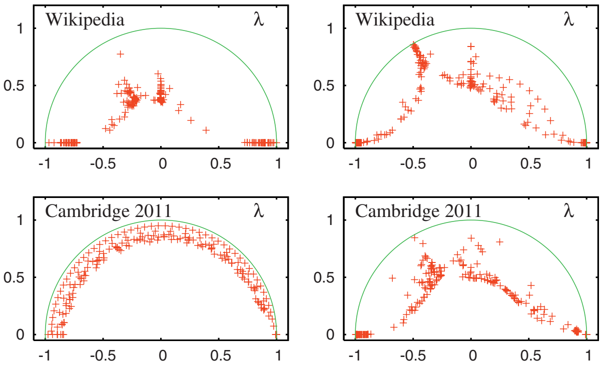

Fig. 3. A selection of 200 complex core space eigenvalues closest to the unit circle for the matrices S (left column) and S^* (right column) of Wikipedia and Cambridge 2011 networks. The characteristics of corresponding eigenvectors are shown in Figures 4 and 5.

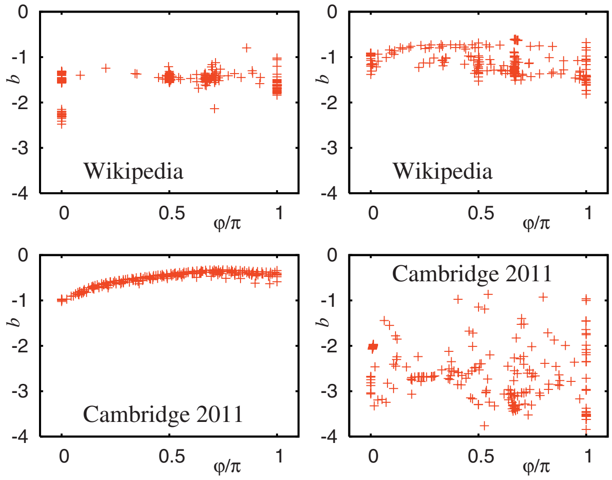

Fig. 4. Left column: algebraic exponent b obtained from a power law fit $|\psi_i(K_i)| \sim K_i^b$ for $K_i \geq 10^4$ shown as a function of the phase $\varphi = \arg(\lambda_i)$ of the complex eigenvalue λ_i associated to the eigenvector ψ_i of S. The shown data points correspond to the eigenvalue 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).

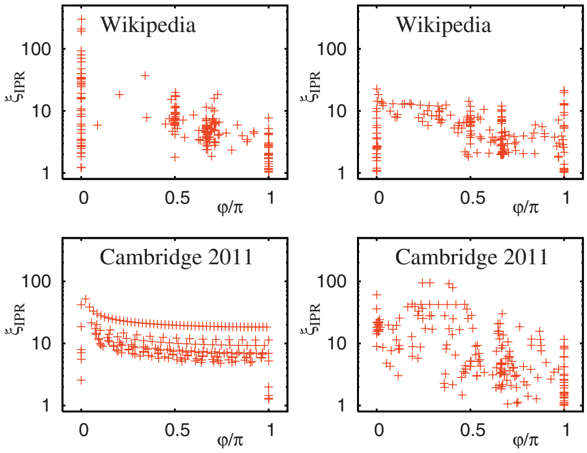

Fig. 5. Left column: inverse participation ratio $\xi_{\mathrm{IPR}} = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$ shown as a function of the phase $\varphi = \arg(\lambda_i)$ of the complex eigenvalue λ_i associated to the eigenvector ψ_i of S. The data points correspond to the eigenvalue selection 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

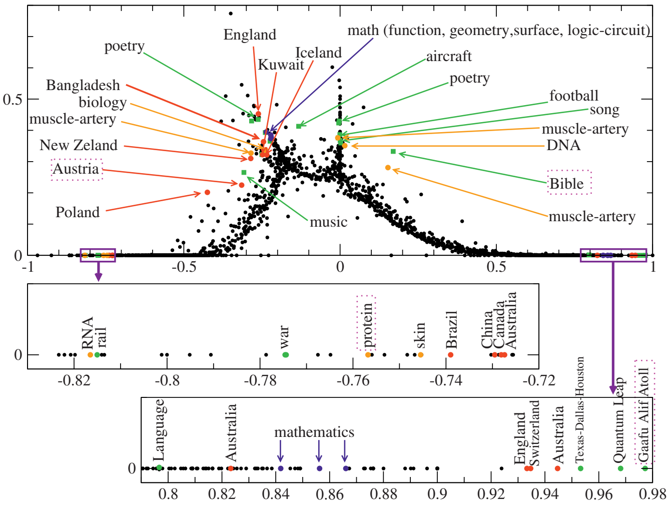

Fig. 6. Complex eigenvalue spectrum of the matrices S for Wikipedia. Highlighted eigenvalues represent different communities of Wikipedia and are labeled by the most repeated and important words following word counting of first 1000 nodes. Color are used in the following way: red for countries, orange 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).

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”)	$\|\psi_i\|$
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”)	$\|\psi_i\|$
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

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 $\varphi = \arg(\lambda_i) = \pi \times 0.3496$ 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")	$\|\psi_i\|$
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")	$\|\psi_i\|$
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.

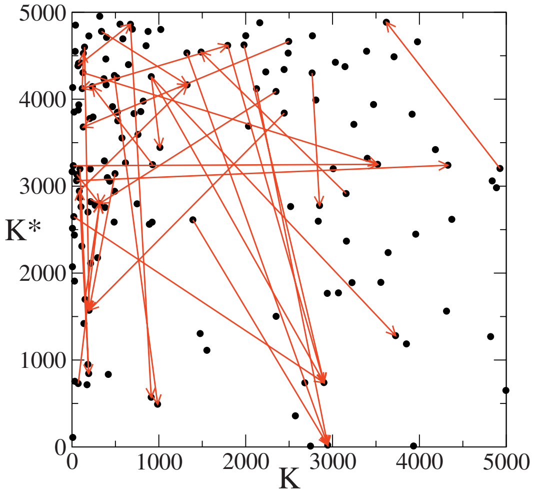

Fig. 7. Top 5000 values in PageRank-CheiRank plane (K, K^*) of Wikipedia. All nodes and all links in this region are shown by black circles and red arrows, respectively.

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]:

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

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.

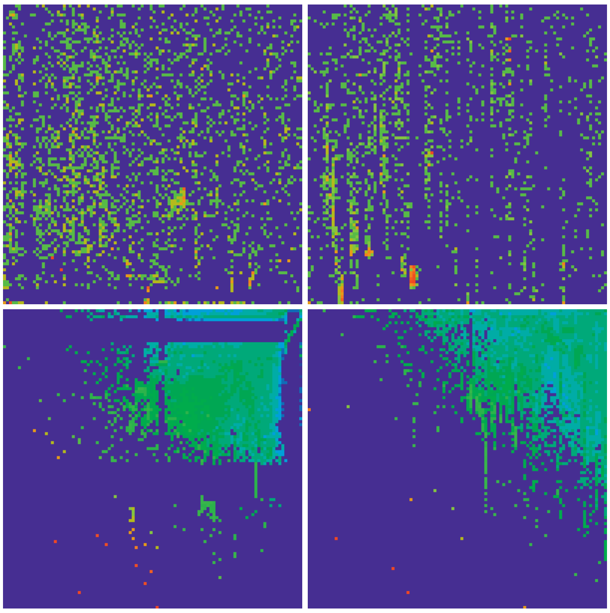

Fig. 8. Density of nodes $W(K, K^*)$ on PageRank-CheiRank plane (K, K^*) for the networks of BBC (left panels) and Le Monde (right panels). Top panels show density in the range $1 \leq K, K^* \leq 10^4$ with averaging over cells of size 100×100; bottom panels show density averaged over 100×100 logarithmically equidistant grids for $0 \leq \ln K, \ln K^* \leq \ln N$, 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

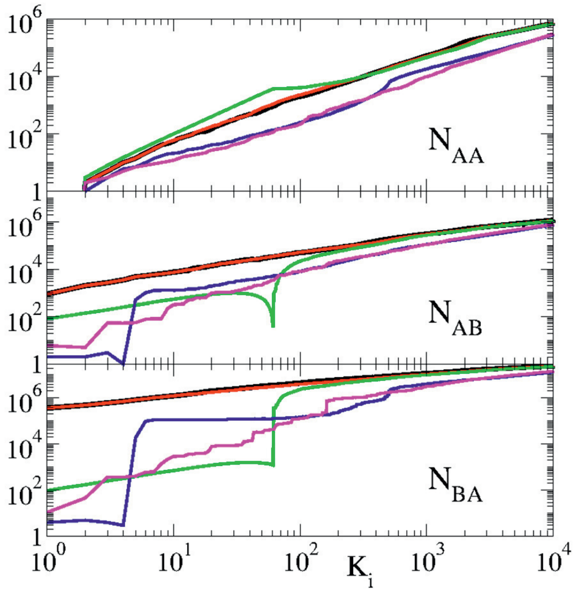

Fig. 9. Number of links between or inside sets A and B defined by the index K_i ordered by decreasing absolute value of Wikipedia eigenstates. The number of links starting and pointing to nodes inside the set A (N_{AA}) is shown in top panel as a function of K_i. The cases of links from set A to set 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)

Google Matrix Analysis of DNA Sequences

Vivek Kandiah, Dima L. Shepelyansky*

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, Toulouse, France

Abstract

For DNA sequences of various species we construct the Google matrix G of Markov transitions between nearby words composed of several letters. The statistical distribution of matrix elements of this matrix is shown to be described by a power law with the exponent being close to those of outgoing links in such scale-free networks as the World Wide Web (WWW). At the same time the sum of ingoing matrix elements is characterized by the exponent being significantly larger than those typical for WWW networks. This results in a slow algebraic decay of the PageRank probability determined by the distribution of ingoing elements. The spectrum of G is characterized by a large gap leading to a rapid relaxation process on the DNA sequence networks. We introduce the PageRank proximity correlator between different species which determines their statistical similarity from the view point of Markov chains. The properties of other eigenstates of the Google matrix are also discussed. Our results establish scale-free features of DNA sequence networks showing their similarities and distinctions with the WWW and linguistic networks.

Citation: Kandiah V, Shepelyansky DL (2013) Google Matrix Analysis of DNA Sequences. PLoS ONE 8(5): e61519. doi:10.1371/journal.pone.0061519

Editor: Ali Torkamani, The Scripps Research Institute, United States of America

Received January 8, 2013; **Accepted** March 4, 2013; **Published** May 9, 2013

Funding: This research is supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No. 288956); the France-Armenia collaboration grant CNRS/SCS No. 24943 (IE-017) on "Classical and quantum chaos"; VK is supported by CNRS – Region Midi-Pyrénées grant. No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dima@irsamc.ups-tlse.fr

Introduction

The theory of Markov chains [1] finds impressive modern applications to information retrieval and ranking of directed networks including the World Wide Web (WWW) where the number of nodes is now counted by tens of billions. The PageRank algorithm (PRA) [2] uses the concept of the Google matrix G and allows to rank all WWW nodes in an efficient way. This algorithm is a fundamental element of the Google search engine used by a majority of Internet users. A detailed description of this method and basic properties of the Google matrix can be found e.g. in [3,4].

The Google matrix belongs to the class of Perron-Frobenius operators naturally appearing in dynamical systems (see e.g. [5]). Using the Ulam method [6] a discrete approximant of Perron-Frobenius operator can be constructed for simple dynamical maps following only one trajectory in a chaotic component [7] or using many independent trajectories counting their probability transitions between phase space cells [8,9], [10]. The studies of Google matrix of such directed Ulam networks provides an interesting and detailed analysis of dynamical properties of maps with a complex chaotic dynamics [7,8], [9,10].

In this work we use the Google matrix approach to study the statistical properties of DNA sequences of the species: Homo sapiens (HS, human), Canis familiaris (CF, dog), Loxodonta africana (LA, elephant), Bos Taurus (bull, BT), Danio rerio (DR, zebrafish), taken from the publicly available database [11]. The analysis of Poincaré recurrences in these DNA sequences [12] shows their similarities with the statistical properties of recurrences for dynamical trajectories in the Chirikov standard map and other symplectic maps [7]. Indeed, a DNA sequence can be viewed as a long symbolic trajectory and hence, the Google matrix, constructed from it, highlights the statistical features of DNA from a new viewpoint.

An important step in the statistical analysis of DNA sequences was done in [13] applying methods of statistical linguistics and determining the frequency of various words composed of up to 7 letters. A first order Markovian models have been also proposed and briefly discussed in this work. Here we show that the Google matrix analysis provides a natural extension of this approach. Thus the PageRank eigenvector gives the frequency appearance of words of given length. The spectrum and eigenstates of G characterize the relaxation processes of different modes in the Markov process generated by a symbolic DNA sequence. We show that the comparison of word ranks of different species allows to identify proximity between species.

At present the investigations of statistical properties of DNA sequences are actively developed by various bioinformatic groups (see e.g. [14,15], [16], [17,18]). The development of various methods of statistical analysis of DNA sequences become now of great importance due to a rapid growth of collected genomic data. We hope that the Google matrix approach, which already demonstrated its efficiency for enormously large networks [2,3], will find useful applications for analysis of genomic data sets.

Results

Construction of Google matrix from DNA sequence

From [11] we collected DNA sequences of HS represented as a single string of length $L \approx 1.5 \cdot 10^{10}$ base pairs (bp) corresponding to 5 individuals. Similar data are obtained for BT ($2.9 \cdot 10^9$ bp), CF ($2.5 \cdot 10^9$ bp), LA ($3.1 \cdot 10^9$ bp), DR ($1.4 \cdot 10^9$ bp). For HS, CF, LA, DR the statistical properties of Poincaré recurrences in these sequences are analyzed in [12]. All strings are composed of 4 letters A,G,G,T and undetermined letter N_l. The strings can be found at the web page [19].

For a given sequence we fix the words W_k of m letters length corresponding to the number of states $N=4^m$. We consider that there is a transition from a state i to state j inside this basis N when we move along the string from left to right going from a word W_k to a next word W_{k+1}. This transition adds one unit in the transition matrix element $T_{ij} \to T_{ij}+1$. The words with letter N_l are omitted, the transitions are counted only between nearby words not separated by words with N_l. There are approximately $N_t \approx L/m$ such transitions for the whole length L since the fraction of undetermined letters N_l is small. Thus we have $N_t = \sum_{i,j=1}^{N} T_{ij}$. The Markov matrix of transitions S_{ij} is obtained by normalizing matrix elements in such a way that their sum in each column is equal to unity: $S_{ij} = T_{ij}/\sum_i T_{ij}$. If there are columns with all zero elements (dangling nodes) then zeros of such columns are replaced by $1/N$. Such a procedure corresponds to one used for the construction of Google matrix of the WWW [2,3]. Then the Google matrix of DNA sequence is written as

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

where α is the damping factor for which the Google search uses usually the value $\alpha \approx 0.85$ [3]. The matrix G belongs to the class of Perron-Frobenius operators. It has the largest eigenvalue $\lambda = \lambda_1 = 1$ with all other eigenvalues $|\lambda_i| \le \alpha$. For WWW usually there are isolated subspaces so that at $\alpha = 1$ there are many degenerate $\lambda = 1$ eigenvalues [4] so that the damping factor allows to eliminate this degeneracy creating a gap between $\lambda = 1$ and all other eigenvalues. For our DNA Google matrices we find that there is already a significant spectral gap naturally present. In this case the PageRank vector is not sensitive to the damping factor being in the range $0.5 \le \alpha \le 1$ (other eigenvectors are independent of α [3,4], [9]). Due to that in the following we present all results at the value $\alpha = 1$.

The spectrum λ_i and right eigenstates $\psi_i(j)$ are determined by the equation

$$\sum_{j'} G_{jj'} \psi_i(j') = \lambda_i \psi_i(j). \quad (2)$$

The PageRank eigenvector $P(j)$ at $\lambda = 1$ has positive or zero elements which can be interpreted as a probability to find a random surfer on a given site j with the total probability normalized to unity $\sum_j P(j) = 1$. Thus, all sites can be ordered in a decreasing order of probability $P(j)$ that gives us the PageRank order index $K(j)$ with most frequent sites at low values of $K = 1,2,...$.

It is useful to consider the density of matrix elements $G_{KK'}$ in the PagePank indexes K,K' similar to the presentation used in [20,21] for networks of Wikipedia, UK universities, Linux Kernel and Twitter. The image of the DNA Google matrix of HS is shown in Fig. 1 for words of 5 and 6 letters. We see that almost all matrix is full that is drastically different from the WWW and other networks considered in [20] where the matrix G is very sparse. Thus the DNA Google matrix is more similar to the case of Twitter which is characterized by a strong connectivity of top PageRank nodes [21].

It is interesting to analyze the statistical properties of matrix elements G_{ij}. Their integrated distribution is shown in Fig. 2. Here N_g is the number of matrix elements of the matrix G with values $G_{ij} > g$. The data show that the number of nonzero matrix elements G_{ij} is very close to N^2. The main fraction of elements has values $G_{ij} \le 1/N$ (some elements $G_{ij} < 1/N$ since for certain j there are many transitions to some node i' with $T_{i'j} \gg N$ and e.g. only one transition to other i'' with $T_{i''j} = 1$). At the same time there are also transition elements G_{ij} with large values whose fraction decays in an algebraic law $N_g \approx AN/g^{\nu-1}$ with some constant A and an exponent ν. The fit of numerical data in the range $-5.5 < \log_{10} g < -0.5$ of algebraic decay gives for $m=6$: $\nu = 2.46 \pm 0.025$ (BT), 2.57 ± 0.025 (CF), 2.67 ± 0.022 (LA), 2.48 ± 0.024 (HS), 2.22 ± 0.04 (DR). For HS case we find $\nu = 2.68 \pm 0.038$ at $m=5$ and $\nu = 2.43 \pm 0.02$ at $m=7$ with the average $A \approx 0.003$ for $m = 5,6,7$. There are visible oscillations in the algebraic decay of N_g with g but in global we see that on average all species are well described by a universal decay law with the exponent $\nu \approx 2.5$. For comparison we also show the distribution N_g for the WWW networks of University of Cambridge and Oxford in year 2006 (data from [4,20]). In these networks we have $N \approx 2 \cdot 10^5$ and on average 10 links per node. We see that in these cases the distribution N_g has a very short range in which the decay is at least approximately algebraic ($-5.5 < \log_{10}(N_g/N^2) < -6$). In contrast to that for the DNA sequences we have a large range of algebraic decay.

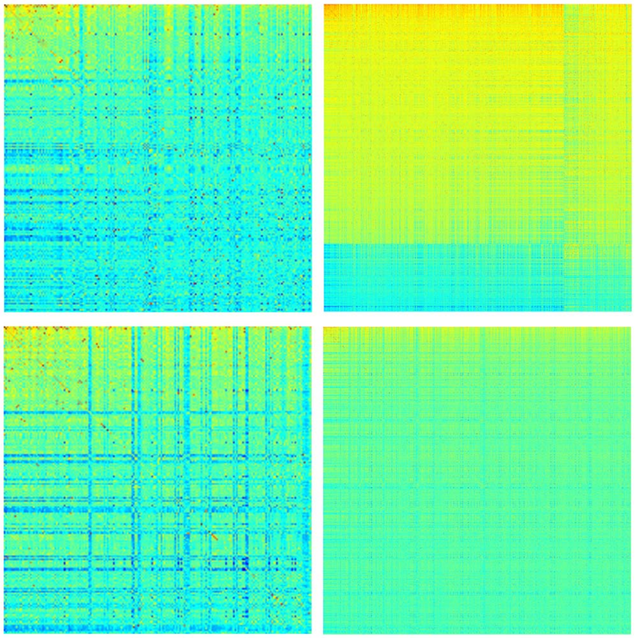

Figure 1. DNA Google matrix of Homo sapiens (HS) constructed for words of 5-letters (top) and 6-letters (bottom) length. Matrix elements $G_{KK'}$ are shown in the basis of PageRank index K (and K'). Here, x and y axes show K and K' within the range $1 \le K,K' \le 200$ (left) and $1 \le K,K' \le 1000$ (right). The element G_{11} at $K = K' = 1$ is placed at top left corner. Color marks the amplitude of matrix elements changing from blue for minimum zero value to red at maximum value.
doi:10.1371/journal.pone.0061519.g001

Since in each column we have the sum of all elements equal to unity we can say that the differential fraction $dN_g/dg \propto 1/g^{\nu}$ gives the distribution of outgoing matrix elements which is similar to the distribution of outgoing links extensively studied for the WWW networks [3,23], [24,25]. Indeed, for the WWW networks all links in a column are considered to have the same weight so that these matrix elements are given by an inverse number of outgoing links [3]. Usually the distribution of outgoing links follows a power law decay with an exponent $\tilde{\nu} \approx 2.7$ even if it is known that this

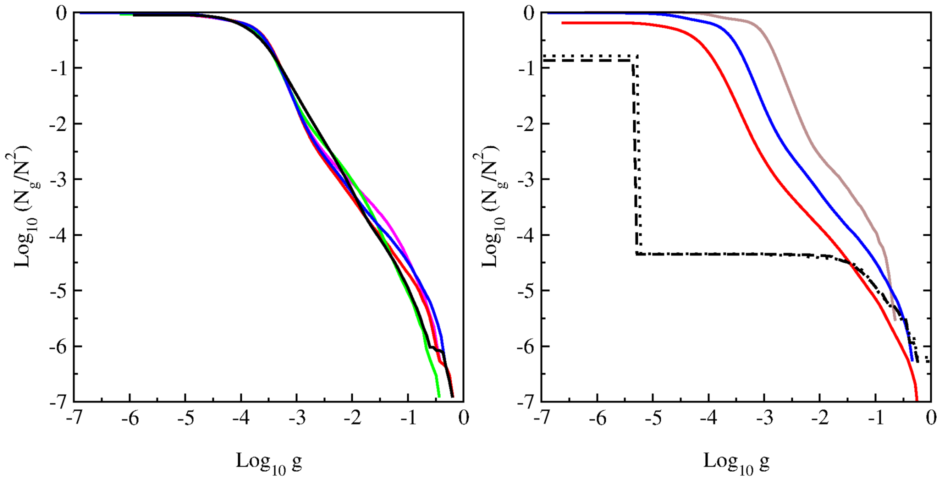

Figure 2. Integrated fraction N_g/N^2 of Google matrix elements with $G_{ij}>g$ as a function of g. *Left panel :* Various species with 6-letters word length: bull BT (magenta), dog CF (red), elephant LA (green), Homo sapiens HS (blue) and zebrafish DR(black). *Right panel :* Data for HS sequence with words of length $m=5$ (brown), 6 (blue), 7 (red). For comparison black dashed and dotted curves show the same distribution for the WWW networks of Universities of Cambridge and Oxford in 2006 respectively.
doi:10.1371/journal.pone.0061519.g002

exponent is much more fluctuating compared to the case of ingoing links. Thus we establish that the distribution of DNA matrix elements is similar to the distribution of outgoing links in the WWW networks with $v\approx\tilde{v}$. We note that for the distribution of outgoing links of Cambridge and Oxford networks the fit of numerical data gives the exponents $\tilde{v}=2.80\pm0.06$ (Cambridge) and 2.51 ± 0.04 (Oxford).

It is known that on average the probability of PageRank vector is proportional to the number of ingoing links [3]. This relation is established for scale-free networks with an algebraic distribution of links when the average number of links per node is about 10 to 100 that is usually the case for WWW, Twitter and Wikipedia networks [4,20], [21,22], [23,24], [25]. Thus in such a case the matrix G is very sparse. For DNA we find an opposite situation where the Google matrix is almost full and zero matrix elements are practically absent. In such a case an analogue of number of ingoing links is the sum of ingoing matrix elements $g_s=\sum_{j=1}^{N}G_{ij}$. The integrated distribution of ingoing matrix elements with the dependence of N_s on g_s is shown in Fig. 3. Here N_s is defined as the number of nodes with the sum of ingoing matrix elements being larger than g_s. A significant part of this dependence, corresponding to large values of g_s and determining the PageRank probability decay, is well described by a power law $N_s\approx BN/g_s^{\mu-1}$. The fit of data at $m=6$ gives $\mu=5.59\pm0.15$ (BT), 4.90 ± 0.08 (CF), 5.37 ± 0.07 (LA), 5.11 ± 0.12 (HS), 4.04 ± 0.06 (DR). For HS case at $m=5{,}7$ we find respectively $\mu=5.86\pm0.14$ and 4.48 ± 0.08. For HS and other species we have an average $B\approx1$.

Usually for ingoing links distribution of WWW and other networks one finds the exponent $\tilde{\mu}\approx2.1$ [23,24], [25]. This value of $\tilde{\mu}$ is expected to be the same as the exponent for ingoing matrix elements of matrix G. Indeed, for the ingoing matrix elements of Cambridge and Oxford networks we find respectively the exponents $\mu=2.12\pm0.03$ and 2.06 ± 0.02 (see curves in Fig. 3). For ingoing links distribution of Cambridge and Oxford networks we obtain respectively $\tilde{\mu}=2.29\pm0.02$ and $\tilde{\mu}=2.27\pm0.02$ which are close to the usual WWW value $\tilde{\mu}\approx2.1$. Thus we can say that for the WWW type networks we have $\mu\approx\tilde{\mu}$. In contrast the exponent μ for DNA Google matrix elements gets significantly larger value $\mu\approx5$. This feature marks a significant difference between DNA and WWW networks.

For DNA we see that there is a certain curvature in addition to a linear decay in log-log scale. From one side, all species are close to a unique universal decay curve which describes the distribution of ingoing matrix elements g_s (there is a more pronounced deviation for DR which does not belong to mammalian species). However, from other side we see visible differences between distributions of various species (e.g. non mammalian DR case has the largest deviation from others mammalian species). We will discuss the links between μ and the exponent β of PageRank algebraic decay $P(K)\propto1/K^{\beta}$ in next sections.

Spectrum of DNA Google matrix

The spectrum of eigenstates of DNA Google matrix G of HS is shown in Fig. 4 for words of $m=5{,}6{,}7$ letters and matrix sizes $N=4^m$. The spectra for DNA sequences of bull BT, dog CF, elephant LA and zebrafish DR are shown in Fig. 5 for words of $m=6$ letters. The spectra and eigenstates are obtained by direct numerical diagonalization of matrix G using LAPACK standard code.

In all cases the spectrum has a large gap which separates eigenvalue $\lambda=1$ and all other eigenvalues with $|\lambda|<0.5$ (only for non mammalian DR case we have a small group of eigenvalues within $0.5<|\lambda|<0.75$). This is drastically different from the spectrum of WWW and other type networks which usually have no gap in the vicinity of $\lambda=1$ (see e.g. [4,21], [22]). In a certain sense the DNA G spectrum is similar to the spectrum of randomized WWW networks and the spectrum of G of the Albert-Baraási network model discussed in [26], but the properties of the PageRank vector are rather different as we will see below.

Visually the spectrum is mostly similar between HS and CF having approximately the same radius of circular cloud $|\lambda|<\lambda_c\approx0.2$. For DR this radius is the smallest with $\lambda_c\approx0.1$. Thus the spectrum of G indicates the difference between mammalian and non mammalian sequences. For HS the increase of the word length $m=5;6;7$ leads to an increase of $\lambda_c\approx0.1;0.2;0.35$. For $m=7$ the number of nonzero matrix elements G_{ij} is close to N^2 and thus on average we have only about $L/(mN^2)\approx8$ transitions per each element. This determines an

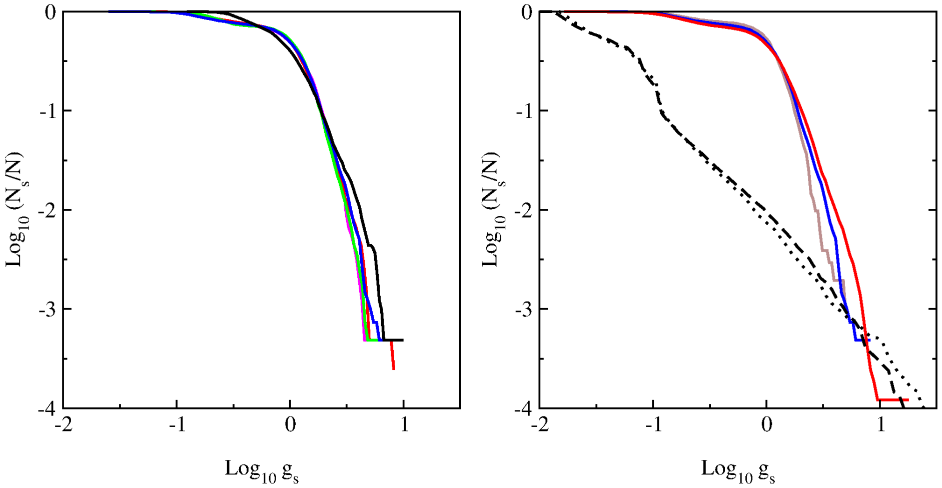

Figure 3. Integrated fraction N_s/N of sum of ingoing matrix elements with $\sum_{j=1}^{N} G_{i,j} \geq g_s$. Left and right panels show the same cases as in Fig. 2 in same colors. The dashed and dotted curves are shifted in x-axis by one unit left to fit the figure scale.
doi:10.1371/journal.pone.0061519.g003

approximate limit of reliable statistical computation of matrix elements G_{ij} for available HS sequence length L. For HS at $m=6$ we verified that two halves of the whole sequence L still give practically the same spectrum with a relative accuracy of $\Delta\lambda/\lambda \approx 0.01$ for eigenvalues in the main part of the cloud at $\lambda_c/3 < |\lambda| < \lambda_c$. This means that the spectrum presented in Figs 4,5 is statistically stable at the values of L used in this work.

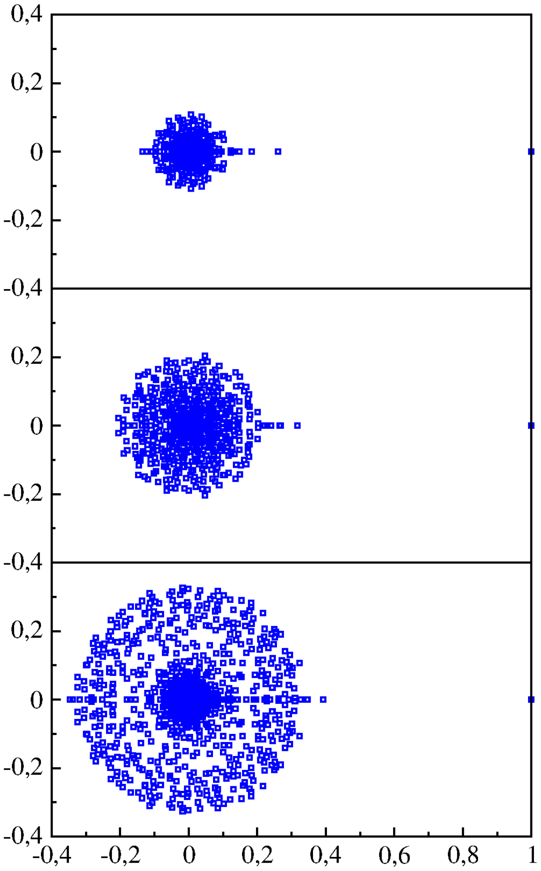

Figure 4. Spectrum of eigenvalues in the complex plane λ for DNA Google matrix of Homo sapiens (HS) shown for words of 5,6,7 letters (from top to bottom).
doi:10.1371/journal.pone.0061519.g004

We also constructed the Google matrix G^* by inverting the direction of transitions $T_{ij} \rightarrow T_{ji}$ and then normalizing sum of all elements in each column to unity. This procedure is also equivalent to moving along the sequence, from word to word, not from left to right but from right to left. We note that for WWW and other networks such a matrix with inverted direction of links was used to obtain the CheiRank vector (which is the PageRank vector of matrix G^*). Due to the inversion of links the CheiRank vector highlights very communicative nodes [4,20], [21,22]. In our case the spectrum of G and G^* are identical. As a result the probability distributions of PageRank and CheiRank vectors are the same. This is due to some kind of detailed balance principle: we count only transitions between nearby words in a DNA sequence and the direction of displacement along the sequence does not affect the average transition probabilities so that $T_{ij} = T_{ji}$ (up to statistical fluctuations). In a certain sense this situation is similar to the case of Ulam networks in symplectic maps where the conservation of phase space area leads to the same properties of G and G^* [7,10].

We tried to test if a random matrix model can reproduce the distribution of eigenvalues in λ plane. With this aim we generated random matrix elements G_{ij} with exactly the same distribution N_g as for HS case at $m=6$ (see Fig. 2). However, in this random model we found all eigenvalues homogeneously distributed in the radius $\lambda_c \approx 0.07$ being significantly smaller compared to the real data. Also in this case the PageRank probability $P(K)$ changes only by 30% in the whole range $1 \leq K \leq N$ being absolutely different from the real data (see next section). Thus the construction of random matrix models which are able to produce results similar to the real data remains as a task for future investigations.

PageRank properties of various species

By numerical diagonalization of the Google matrix we determine the PageRank vector $P(K)$ at $\lambda = 1$ and several other eigenvectors with maximal values of $|\lambda|$. The dependence of probability P on index K is shown in Fig. 6 for various species and different word length m. The probability $P(K)$ describes the steady

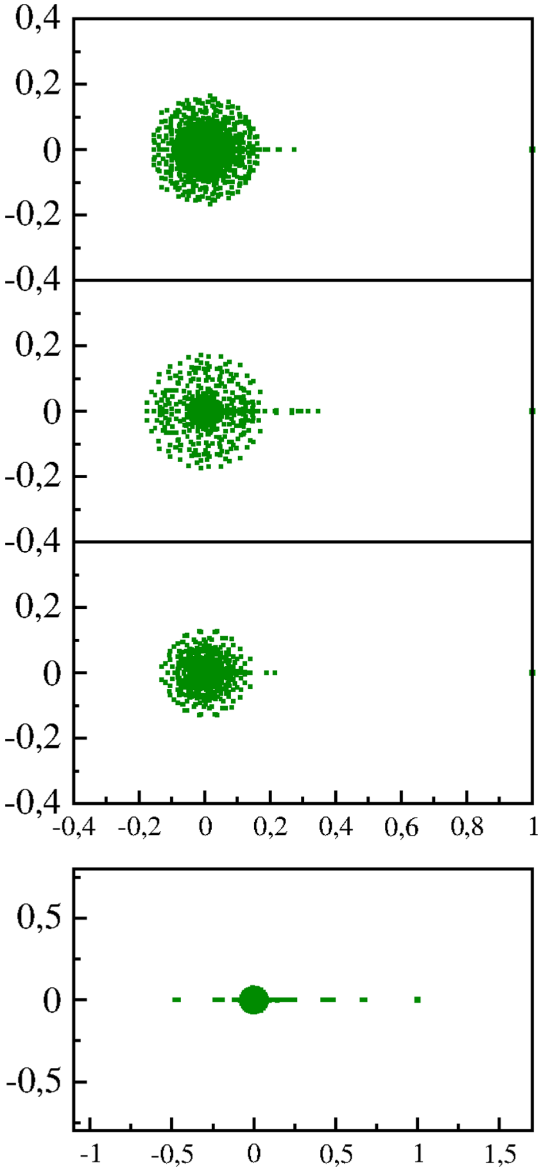

Figure 5. Spectrum of eigenvalues in the complex plane λ for DNA Google matrix of of bull BT, dog CF, elephant LA, zebrafish DR shown for words of 6 **letters (from top to bottom).**
doi:10.1371/journal.pone.0061519.g005

state of random walks on the Markov chain and thus it gives the frequency of appearance of various words of length m in the whole sequence L. The frequencies or probabilities of words appearance in the sequences have been obtained in [13] by a direct counting of words along the sequence (the available sequences L were shorted at that times). Both methods are mathematically equivalent and indeed our distributions $P(K)$ are in a good agreements with those found in [13] even if now we have a significantly better statistics.

The decay of P with K can be approximately described by a power law $P \sim 1/K^{\beta}$. Thus for example for HS sequence at $m=7$ we find $\beta=0.357 \pm 0.003$ for the fit range $1.5 \leq \log_{10} K \leq 3.7$ that is rather close to the exponent found in [13]. Since on average the PageRank probability is proportional to the number of ingoing links, or the sum of ingoing matrix elements of G, one has the relation between the exponent of PageRank β and exponent of ingoing links (or matrix elements): $\beta=1/(\mu-1)$ [3,4], [23,24], [25]. Indeed, for the HS DNA case at $m=7$ we have $\mu=4.48$ that gives $\beta=0.29$ being close to the above value of $\beta=0.357$ obtained from the direct fit of $P(K)$ dependence. We think that the agreement is not so perfect since there is a visible curvature in the log-log plot of N_s vs g_s in Fig. 3. Also due to a small value of β the variation range of P is not so large that reduces the accuracy of the numerical fit even if a formal statistical error is relatively small compared to a visible systematic nonlinear variations. In spite of this only approximate agreement we should say that in global the relation between β and μ works correctly. In average we find for DNA network the value of $\mu \approx 5$ being significantly larger than for the WWW networks with $\tilde{\mu} \approx 2.1$ [3]. This gives a significantly smaller value $\beta \approx 0.25$ for DNA case comparing to the usual WWW value $\beta \approx 0.9$ (we note that the randomized WWW networks and the Albert-Barabási model have $\beta \approx 1$ [26]). The relation between β and μ also works for the DR DNA case at $m=6$ with $\mu=4.04$ that gives $\beta=0.33$ being in a satisfactory agreement with the fit value $\beta=0.426$ found from $P(K)$ dependence of Fig. 6.

At $m=6$ we find for our species the following values of exponent $\beta=0.273 \pm 0.005$ (BT), 0.340 ± 0.005 (CF), 0.281 ± 0.005 (LA), 0.308 ± 0.005 (HS), 0.426 ± 0.008 (DR) in the range $1 \leq \log_{10} K \leq 3.3$. There is a relatively small variation of β between various mammalian species. The data of Fig. 6 for HS show that the value of β remains stable with the increase of word length. These observations are similar to those made in [13].

PageRank proximity between species

The top ten 6-letters words, with largest probabilities $P(K)$, are given for all studied species in Table 1. Two top words are identical for BT, CF, HS. To see a similarity between species on a global scale it is convenient to plot the PageRank index $K_s(i)$ of a given species s versus the index $K_{hs}(i)$ of HS for the same word i. For identical sequences one should have all points on diagonal, while the deviations from diagonal characterize the differences between species. The examples of such PageRank proximity $K-K$ diagrams are shown in Figs. 7,8 for words at $m=6$. A zoom of data on a small scale at the range $1 \leq K \leq 200$ is shown in Fig. 9. A visual impression is that CF case has less deviations from HS rank compared to BT and LA. The non-mammalian DR case has most strong deviations from HS rank. For BT, CF and LA cases we have a significant reduction of deviations from diagonal around $K \approx 3N/4$. This effect is also visible for DR case even if being less pronounced. We do not have explanation for this observation.

The fraction of purine letters A or G in a word of $m=6$ letters is shown by color in Fig. 7 for all words ranked by PageRank index K. We see that these letters are approximately homogeneously distributed over the whole range of K values. In contrast to that the distribution of letters A or T is inhomogeneous in K: their fraction is dominant for $1 \leq K < N/4$, approximately homogeneous for $N/4 \leq K \leq 3N/4$ and is close to zero for $3N/4 < K \leq N$ (see Fig. 8). We find that in the whole HS sequence the fractions $F_{a,c,g,t}$ of A,C,G,T are respectively 0.276596,0.192576,0.192624,0.276892 (and $F_n=0.061312$ for undetermined N_l). Thus we have the fraction of A,G being close to $1/2 \approx (F_a+F_g)/(1-F_n)=0.499867$ and the fraction of A,T being $(F_a+F_t)/(1-F_n)=0.589640>0.5$. Thus it is more probable to have A or T in the whole sequence that can be a possible origin of the inhomogeneous distribution of A or T along K and large fraction of A, T at top PageRank positions.

The whole HS sequence used here is composed from 5 humans with individual length $L_i \approx 3 \cdot 10^9 \approx L/5$. We consider the first and last fifth parts of the whole sequence L separately thus forming two independent sequences HS1 and HS2 of two individuals. We

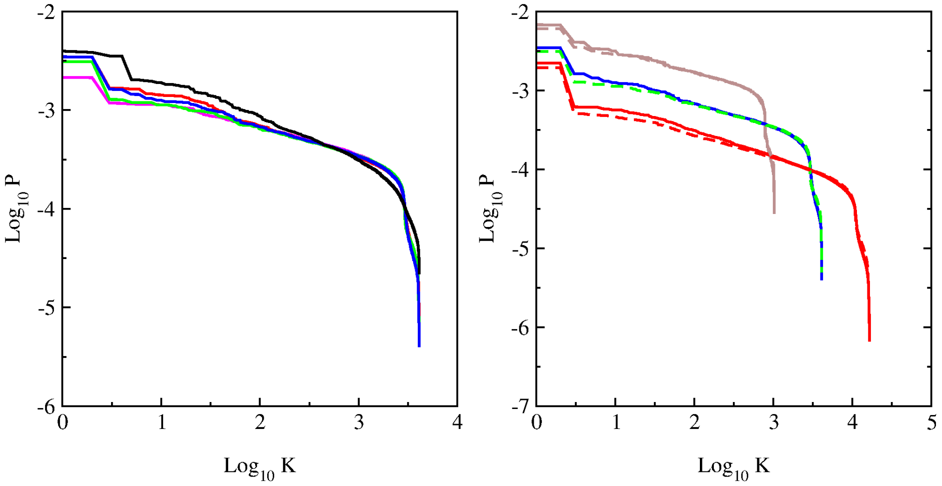

Figure 6. Dependence of PageRank probability $P(K)$ on PageRank index K. *Left panel :* Data for different species for word length of 6-letters: bull BT (magenta), dog CF (red), elephant LA (green), Homo sapiens HS (blue) and zebrafish DR (black). *Right panel :* Data for HS (full curve) and LA (dashed curve) for word length $m=5$ (brown), 6 (blue/green), 7 (red).
doi:10.1371/journal.pone.0061519.g006

determine for the the corresponding PageRank indexes K_{hs1} and K_{hs2} and show their PageRank proximity diagram in Fig. 10. In this case the points are much closer to diagonal compared to the case of comparison of HS with other species.

To characterize the proximity between different species or different HS individuals we compute the average dispersion $\sigma(s_1,s_2)=\sqrt{\sum_{i=1}^{N}(K_{s_1}(i)-K_{s_2}(i))^2)/N}$ between two species (individuals) s_1 and s_2. Comparing the words with length $m=5,6,7$

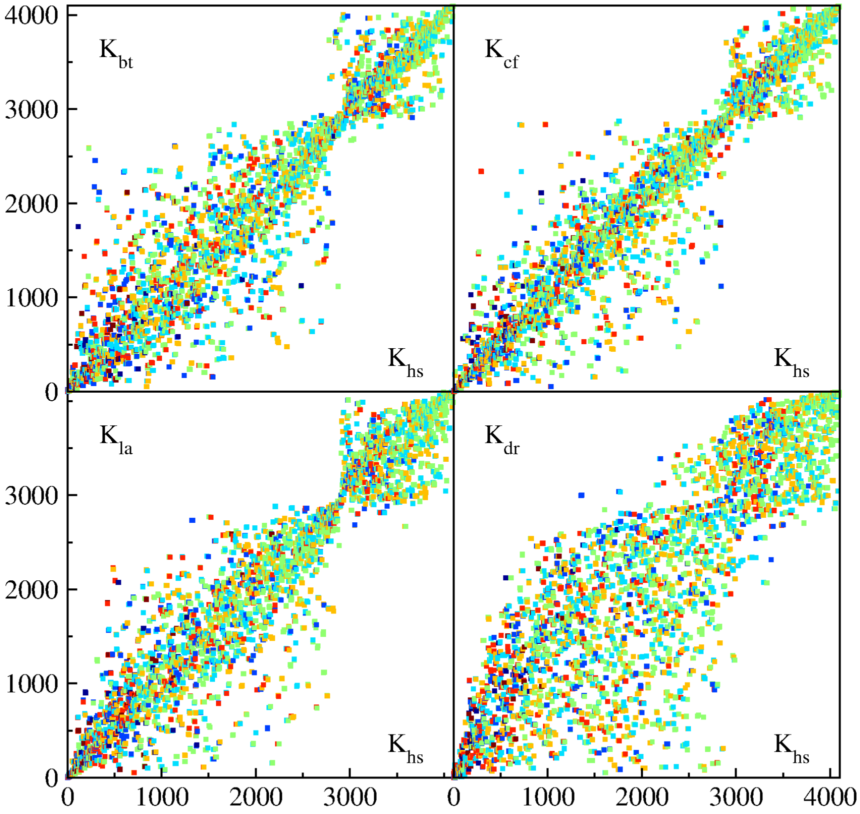

Figure 7. PageRank proximity $K-K$ plane diagrams for different species in comparison with Homo sapiens: x-axis shows PageRank index $K_{hs}(i)$ of a word i and y-axis shows PageRank index of the same word i with $K_{bt}(i)$ of bull, $K_{cf}(i)$ of dog, $K_{la}(i)$ of elephant and $K_{dr}(i)$ of zebrafish; here the word length is $m=6$. The colors of symbols marks the purine content in a word i (fractions of letters A or G in any order); the color varies from red at maximal content, via brown, yellow, green, light blue, to blue at minimal zero content.
doi:10.1371/journal.pone.0061519.g007

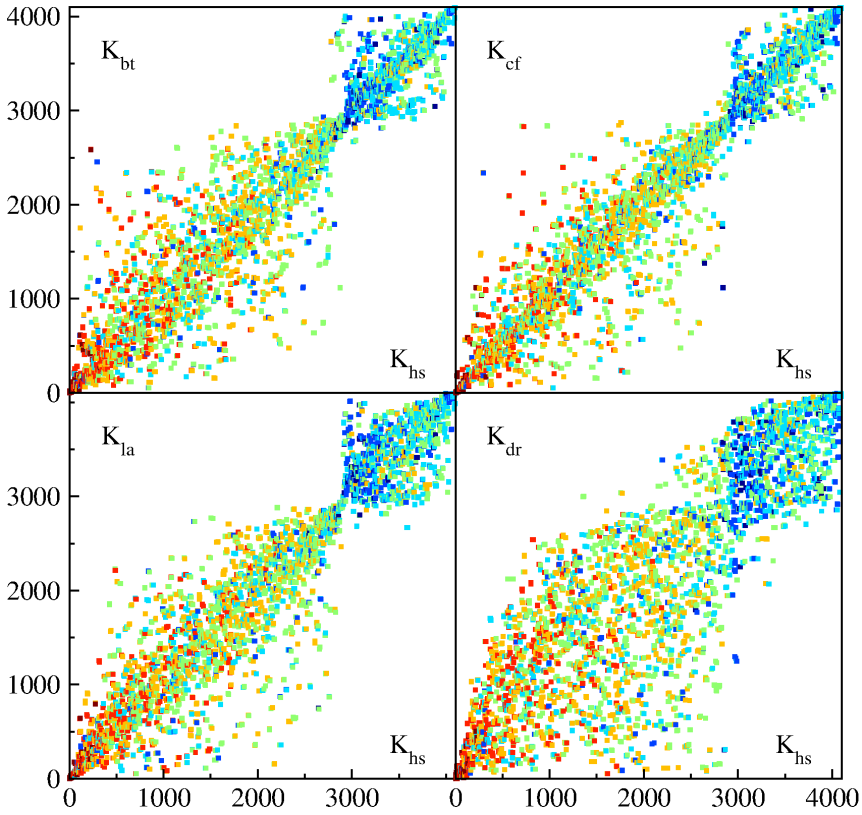

Figure 8. Same as in Fig. 7 but now the color marks the fraction of of letters A or T in any order in a word i with red at maximal content and blue at zero content.
doi:10.1371/journal.pone.0061519.g008

we find that the scaling $\sigma \propto N$ works with a good accuracy (about 10% when N is increased by a factor 16). To represent the result in a form independent of m we compare the values of σ with the corresponding random model value σ_{rnd}. This value is computed assuming a random distribution of N points in a square $N \times N$ when only one point appears in each column and each line (e.g. at $m=6$ we have $\sigma_{rnd} \approx 1673$ and $\sigma_{rnd} \propto N$). The dimensionless dispersion is then given by $\zeta(s_1,s_2)=\sigma(s_1,s_2)/\sigma_{rnd}$. From the ranking of different species we obtain the following values at $m=6$: $\zeta(CF,BT)=0.308$; $\zeta(LA,BT)=0.324$, $\zeta(LA,CF)=0.303$; $\zeta(HS,BT)=0.246$, $\zeta(HS,CF)=0.206$, $\zeta(HS,LA)=0.238$; $\zeta(DR,BT)=0.425$, $\zeta(DR,CF)=0.414$, $\zeta(DR,LA)=0.422$, $\zeta(DR,HS)=0.375$ (other m have similar values). According to this statistical analysis of PageRank proximity between species we find that ζ value is minimal between CF and HS showing that these are two most similar species among those considered here.

For two HS individuals we find $\zeta(HS1,HS2)=0.031$ being significantly smaller then the proximity correlator between different species. We think that this PageRank proximity correlator ζ can be useful as a quantitative measure of statistical proximity between various species.

Finally, in Table 2 we give for all species the words of 6 letters with the 10 minimal PageRank probabilities. Thus for HS the less probable is the word TACGCG corresponding to two amino acids Tyr and Ala. In general the ten last words are mainly composed of C and G even if the letters A and T still have small but nonzero weight. The last two words are the same for mammalian species but they are different for DR sequence.

Other eigenvectors of G

The properties of 10 eigenstates $\psi_i(j)$ of DNA Google matrix with largest modulus of eigenvalues $|\lambda_i|$ are analyzed in Table 3 and Fig. 11. The words W_i at the maximal amplitude $|\psi_i(j)|$ are presented for all species in Table 3. We see that in general these words W_i are rather different from the top PageRank word W_1 (some words appear in pairs since there are pairs of complex conjugated values $\lambda_i=\lambda_i^*$).

The probability of the above top 10 eigenstates as a function of PageRank index K are shown in Fig. 11. We see that the majority of the vectors, different from the PageRank vector, have well localized peaks at relatively large values $K>50$. This shows that in the DNA network there are some modes located on certain specific patterns of words.

To illustrated the localized structure of eigenmodes $\psi_i(j)$ for HS case at $m=6$ we compute the inverse participation ratio $\xi_i=(\sum_j |\psi_i(j)|^2)^2/\sum_j |\psi_i(j)|^4$ which gives an approximate number of nodes on which the main probability of an eigenstate $\psi_i(j)$ is located (see e.g. [4,21,26]). The obtained values are $\xi_i=385.26$, 16.37, 2.07, 1.72, 2.23, 3.19, 77.43, 77.43, 2.33, 2.06 for $i=1,...10$ respectively. We see that for $i>1$ we have significantly smaller ξ values compared to the case of PageRank vector with a large ξ_1. This supports the conclusion about localized structure of a large fraction of eigenvectors of G.

In [22] on an example of Wikipedia network it is shown that the eigenstates with relatively large $|\lambda|$ select specific communities of the network. The detection of communities in complex networks is now an active research direction [27]. We expect that the eigenmodes of G matrix can select specific words of bioinformatic

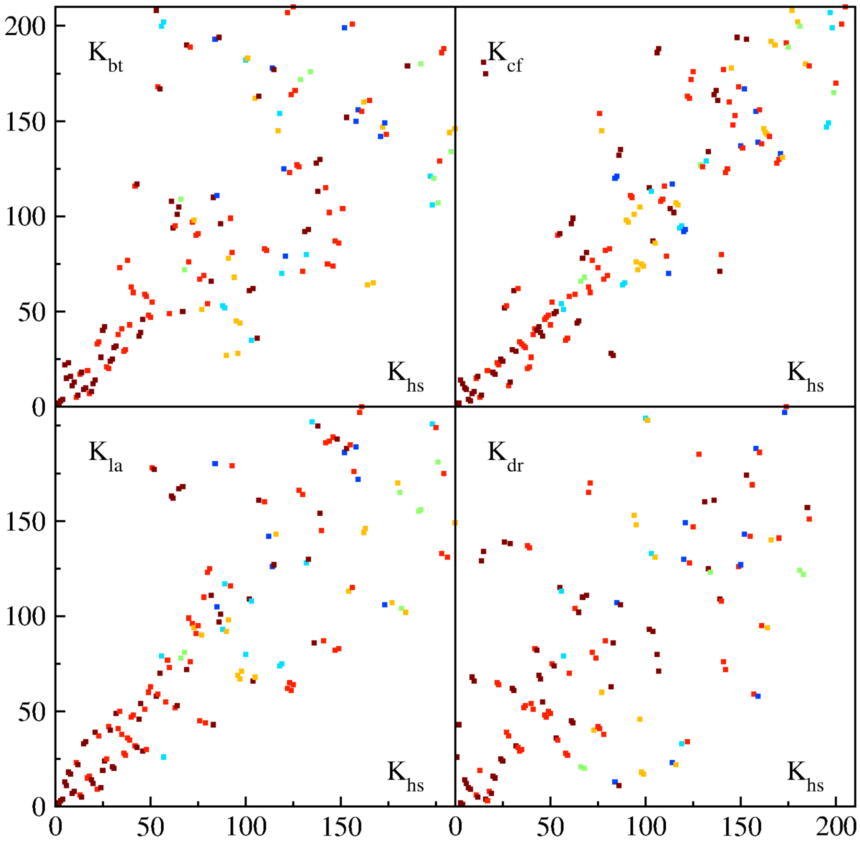

Figure 9. Zoom of the PageRank proximity $K-K$ diagram of Fig. 8 for the range $1 \leq K \leq 200$ with the same color for A or T content.
doi:10.1371/journal.pone.0061519.g009

interest. However, a detailed analysis of words from eigenmodes remains for further more detailed investigations.

Discussion

In this work we used long DNA sequences of various species to construct from them the Markov process describing the probabilistic transitions between words of up to 7 letters length. We construct the Google matrix of such transitions with the size up to 4^7 and analyze the statistical properties of its matrix elements. We show that for all 5 species, studied in this work, the matrix elements of significant amplitude have a power law distribution with the exponent $\nu \approx 2.5$ being close to the exponent of outgoing links distribution typical for WWW and other complex directed networks with $\tilde{\nu} \approx 2.7$. The distribution of significant values of the sum of ingoing matrix elements of G is also described by a power

Table 1. Top ten PageRank entries at DNA word length $m=6$ for species: bull BT, dog CF, elephant LA, Homo sapiens HS and zebrafish DR.

BT	CF	LA	HS	DR
TTTTTT	TTTTTT	AAAAAA	TTTTTT	ATATAT
AAAAAA	AAAAAA	TTTTTT	AAAAAA	TATATA
ATTTTT	AATAAA	ATTTTT	ATTTTT	AAAAAA
AAAAAT	TTTATT	AAAAAT	AAAAAT	TTTTTT
TTCTTT	AAATAA	AGAAAA	TATTTT	AATAAA
TTTTAA	TTATTT	TTTTCT	AAAATA	TTTATT
AAAGAA	AAAAAT	AAGAAA	TTTTTA	AAATAA
TTAAAA	ATTTTT	TTTCTT	TAAAAA	TTATTT
TTTTCT	TTTTTA	TTTTTA	TTATTT	CACACA
AGAAAA	TAAAAA	TAAAAA	AAATAA	TGTGTG

doi:10.1371/journal.pone.0061519.t001

Table 2. Ten words with minimal PageRank probability given at $m=6$ for species: bull BT, dog CF, elephant LA, Homo sapiens HS and zebrafish DR.

BT	CF	LA	HS	DR
CGCGTA	TACGCG	CGCGTA	TACGCG	CCGACG
TACGCG	CGCGTA	TACGCG	CGCGTA	CGTCGG
CGTACG	TCGCGA	ATCGCG	CGTACG	CGTCGA
CGATCG	CGTACG	TCGCGA	TCGACG	TCGACG
ATCGCG	CGATCG	CGCGAT	CGTCGA	TCGTCG
CGCGAT	CGAACG	GTCGCG	CGATCG	CCGTCG
TCGACG	CGTTCG	CGATCG	CGTTCG	CGACGG
CGTCGA	TCGACG	CGCGAC	CGAACG	CGACCG
CGTTCG	CGTCGA	TCGCGC	CGACGA	CGGTCG
TCGTCG	ACGCGA	ACGCGA	CGCGAA	CGACGA

Here the top row is the last PageRank entry, bottom is the tenth one from the end of PageRank.
doi:10.1371/journal.pone.0061519.t002

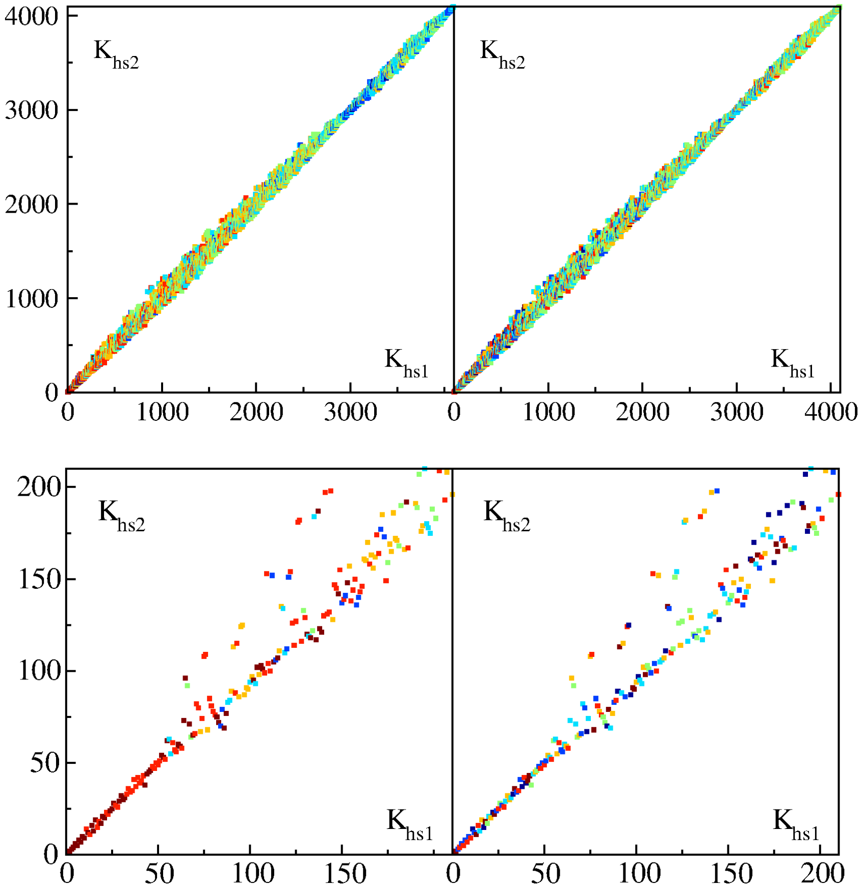

Figure 10. PageRank proximity $K-K$ diagram of Homo sapiens $HS2$ versus Homo sapiens $HS1$ at $m=6$ (see text for details). Top panels show the content of A,T (left) and A,G (right) in the same way as in Fig. 8 and Fig. 7 respectively. Bottom panels show zoom of top panels.
doi:10.1371/journal.pone.0061519.g010

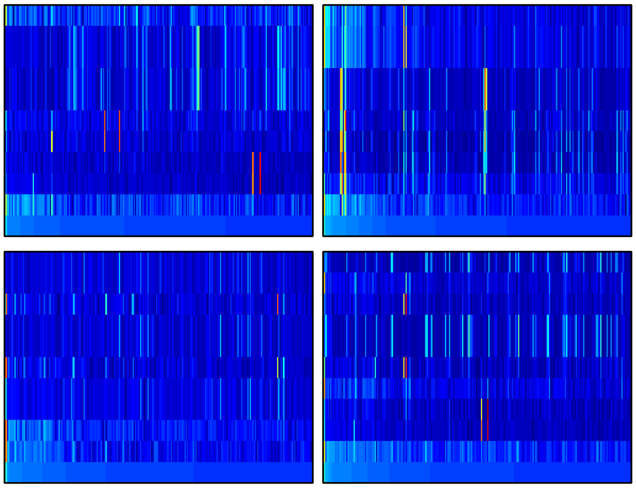

Figure 11. Dependence of eigenstates amplitude $|\psi_i(K)|$ on PageRank index K in x-axis and eigenvalue index i in y-axis for largest ten eigenvalues $|\lambda_i|$ counted by i from $i=1$ at $|\lambda_1|=1$ to $i=10$ at $|\lambda_{10}|\approx 0.2$. The range $1\leq K\leq 250$ is shown with PageRank vector for a given species at the bottom line of each panel. For each species in each panel the color is proportional to $\sqrt{|\psi_i(j)|}$ changing from blue at zero to red at maximal amplitude value which is close to unity in each panel. The panels show the species: bull BT (top left), dog CF (top right), elephant LA (bottom left), Homo sapiens HS (bottom right).
doi:10.1371/journal.pone.0061519.g011

Table 3. Words W_i corresponding to the maximum value of eigenvector modulus $w_i=max_j(|\psi_i(j)|)$ for species bull BT, dog CF, elephant LA, Homo sapiens HS and zebrafish DR, which are shown in dark red in Fig. 11.

i	BT	CF	LA	HS	DR
1	TTTTTT	TTTTTT	AAAAAA	TTTTTT	ATATAT
2	TTTTTT	AAAAAA	AAAAAA	TTTTTT	TATATA
3	ACACAC	CTCTCT	AAAAAA	ACACAC	ATATAT
4	ACACAC	AGAGAG	AAAAAA	ACACAC	TAGATA
5	CACACA	CTCTCT	AAAAAA	TTTTTT	ATAGAT
6	CACACA	TCTCTC	AAAAAA	CACACA	TATCTA
7	CCAGGC	AGAGAG	TATGAG	TGGGAG	ATCTAT
8	CCAGGC	AGAGAG	TATGAG	TGGGAG	TAGATA
9	CCCATG	TGTGTG	TTTTTT	CACACA	ATAGAT
10	CCCATG	TGTGTG	AGAGTA	TTTTTT	TATCTA

The eigenvectors at $i=1,\ldots,10$ correspond to the ten largest eigenvalues $|\lambda_1|,\ldots,|\lambda_{10}|$ of the DNA Google matrix for DNA word length $m=6$. The first row $i=1$ corresponds to top PageRank entries.
doi:10.1371/journal.pone.0061519.t003

law with the exponent $\mu \approx 5$ which is significantly larger than the corresponding exponent for WWW networks with $\tilde{\mu} \approx 2.1$. We show that similar to the WWW networks the exponent μ determines the exponent $\beta = 1/(\mu-1) \approx 0.25$ of the algebraic PageRank decay which is significantly smaller then its value for WWW networks with $\beta \approx 0.9$. The PageRank decay is similar to the frequency decay of various words studied previously in [13]. It is interesting to note that the value $\mu - 1$ is close to the exponent of Poincaré recurrences decay which has a value close to 4 [12] (even if we cannot derive a direct mathematical relation between them).

Using PageRank vectors of various species we introduce the PageRank proximity correlator ζ which allows to measure in a quantitative way the proximity between different species. This parameter remains stable in respect to variation of the word length.

The spectrum of the Google matrix is determined and it is shown that it is characterized by a significant gap between $\lambda = 1$ and other eigenvalues. Thus, this spectrum is qualitatively different from the WWW case where the gap is absent at the damping factor $\alpha = 1$. We show that the eigenmodes with largest values of $|\lambda| < 1$ are well localized on specific words and we argue that the words corresponding to such localized modes can play an interesting role in bioinformatic properties of DNA sequences.

Finally we would like to trace parallels between the Google matrix analysis of words in DNA sequences and the small world properties of human language. Indeed, it is known that the frequency of words in natural languages follows a power law Zipf distribution with the exponent $\beta \approx 1$ [28]. The parallels between words distributions in DNA sequences and statistical linguistics were already pointed in [13]. The analysis of degree distributions of undirected networks of words in natural languages was found to follow a power law with an exponent $\nu_l \approx 1.5 - 2.7$ [29] being not so far from the one found here for the matrix elements distribution. It is argued that the language evolution plays an important role in the formation of such a distribution in languages [30]. The parallels between linguistics and DNA sequence complexity are actively discussed in bioinformatics [31,32]. We think that the Google matrix analysis can provide new insights in the construction and characterization of information flows on DNA sequence networks extending recent steps done in [33].

In summary, our results show that the distributions of significant matrix elements are similar to those of the scale-free type networks like WWW, Wikipedia and linguistic networks. In analogy with lingusitic networks it can be useful to go from words network analysis to a more advanced functional level of links inside sentences that may be viewed as a network of links between amino acids or more complex biological constructions.

Supporting Information

Supporting Information S1. Supplementary methods, references, tables, sequences data and figures are available at: http://www.quantware.ups-tlse.fr/QWLIB/dnagooglematrix/.

Acknowledgments

We thank K.M.Frahm for useful discussions and help in collection of DNA sequences from [12] which are studied here.

Author Contributions

Conceived and designed the experiments: DLS. Performed the experiments: VK. Analyzed the data: VK DLS. Contributed reagents/materials/analysis tools: VK DLS. Wrote the paper: VK DLS.

References

1. Markov AA (1906) Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, 15: 135 (in Russian) [English trans.: *Extension of the limit theorems of probability theory to a sum of variables connected in a chain* reprinted in Appendix B of Howard RA *Dynamic Probabilistic Systems*, volume 1: *Markov models*, Dover Publ. (2007)].
2. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine Computer Networks and ISDN Systems 30: 107.
3. Langville AM, Meyer CD (2006) Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, Princeton.
4. Frahm KM, Georgeot B, Shepelyansky DL (2011) Universal emergence of PageRank, J Phys. A: Math. Theor. 44: 465101.
5. Brin M, Stuck G (2002) Introduction to dynamical systems, Cambridge Univ. Press, Cambridge, UK.
6. Ulam SM (1960) A Collection of mathematical problems, Interscience tracs in pure and applied mathematics 8: 73, Interscience, New York.
7. Frahm KM, Shepelyansky DL (2010) Ulam method for the Chirikov standard map Eur. Phys J B 76: 57.
8. Froyland G, Padberg K (2009) Almost-invariant sets and invariant manifolds connecting probabilistic and geometric descriptions of coherent structures in flows, Physica D 238: 1507.
9. Shepelyansky DL, Zhirov OV (2010) Google matrix, dynamical attractors and Ulam networks, Phys. Rev E 81: 036213.
10. Ermann L, Shepelyansky DL (2012) The Arnold cat map, the Ulam method and time reversal, Physica D 241: 514.
11. Ensembl Genome Data Base. Available: http://www.ensembl.org/ and ftp://ftp.ensembl.org/pub/release-62/genbank/.
12. Frahm KM, Shepelyansky DL (2012) Poincaré recurrences of DNA sequences, Phys. Rev E 85: 016214.
13. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, et al. (1995) Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics, Phys. Rev E 52: 2939.
14. Robin S, Rodolphe F, Schbath S (2005) DNA, words and models, Cambridge Univ. Press, Cambridge.
15. Halperin D, Chiapello H, Schbath S, Robin S, Hennequet-Antier C, et al. (2007) Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling, PLoS Genetics 3(9): e153.
16. Dai Q, Yang Y, Wang T (2008) Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison, Bioinformatics 24(20): 2296.
17. Reinert G, Chew D, Sun D, Waterman MS (2009) Alignment-free sequence comparison (I): statistics and power, J Comp. Biology 16(12): 1615.
18. Burden CJ, Jing J, Wilson SR (2012) Alignment-free sequence comparison for biologically realistic sequences of moderate length, Stat. Appl. Gen. Mol. Biology 11(1) 3.
19. Sequences Data Quantware Web Site. Available: www.quantware.ups-tlse.fr/QWLIB/dnagooglematrix/.
20. Ermann L, Chepelianskii AD, Shepelyanksy DL (2012) Toward two-dimensional search engines, J Phys. A: Math. Theor. 45:275101.
21. Frahm KM, Shepelyansky DL (2012) Google matrix of Twitter, Eur. Phys J B 85:355.
22. Ermann L, Frahm KM, Shepelyansky DL (2012) Spectral properties of Google matrix of Wikipedia and other networks, arXiv:1212.1068 [cs.IR].
23. Donato D, Laura L, Leonardi S, Millozzi S (2004) Large scale properties of the Webgraph, Eur. Phys J B 38: 239.
24. Pandurangan G, Raghavan P, Upfal E (2005) Using PageRank to characterize Web structure, Internet Math. 3: 1.
25. Zhirov AO, Zhirov OV, Shepelyansky DL (2010) Two-dimensional ranking of Wikipedia articles, Eur. Phys J B 77: 523.
26. Giraud O, Georgeot B, Shepelyansky DL (2009) Delocalization transition for the Google matrix, Phys. Rev E 80: 026107.
27. Fortunato S (2010) Community detection in graphs, Phys. Rep. 486: 75.
28. Zipf GK (1949) Human behavior and the principle of least effort, Addison-Wesley, Boston.
29. Cancho RFi, Sole RV (2001) The small world of human language, Proc R Soc. Lond B 268: 2261.
30. Dorogovtsev SN, Mendes JFF (2001) Language as an evolving word web, Proc R Soc. Lond B 268: 2603.
31. Brendel V, Beckmann JS, Trifonov EN (1986) Linguistics of nucleotide sequences: morphology and comparison of vocabularies, J Boimolecular Structure Dynamics 4: 11.
32. Popov O, Segal DM, Trifonov EN (1996) Linguistic complexity of protein sequences as compared to texts of human languages, Biosystems 38: 65.
33. Frenkel Zakharia M, Frenkel Zeev M, Trifonov EN, Snir S (2009) Structural relatedness via flow networks in protein sequence space, J Theor. Biology 260: 438.

EPJ manuscript No.
(will be inserted by the editor)

Time evolution of Wikipedia network ranking

Young-Ho Eom[1], Klaus M. Frahm[1], András Benczúr[2], and Dima L. Shepelyansky[1]

[1] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

[2] Informatics Laboratory, Institute for Computer Science and Control,
Hungarian Academy of Sciences (MTA SZTAKI), Pf. 63, H-1518 Budapest, Hungary

Received: April 24, 2013

Abstract. We study the time evolution of ranking and spectral properties of the Google matrix of English Wikipedia hyperlink network during years 2003 - 2011. The statistical properties of ranking of Wikipedia articles via PageRank and CheiRank probabilities, as well as the matrix spectrum, are shown to be stabilized for 2007 - 2011. A special emphasis is done on ranking of Wikipedia personalities and universities. We show that PageRank selection is dominated by politicians while 2DRank, which combines PageRank and CheiRank, gives more accent on personalities of arts. The Wikipedia PageRank of universities recovers 80 percents of top universities of Shanghai ranking during the considered time period.

PACS. 89.75.Fb Structures and organization in complex systems – 89.75.Hc Networks and genealogical trees – 89.20.Hh World Wide Web, Internet

1 Introduction

At present Wikipedia [1] became the world largest Encyclopedia with open public access to its contain. A recent review [2] represents a detailed description of publications and scientific research of this modern Library of Babel, which stores an enormous amount of information, approaching the one described by Jorge Luis Borges [3]. The hyperlinks of citations between Wikipedia articles represent a directed network which reminds the structure of the World Wide Web (WWW). Hence, the mathematical tools developed for WWW search engines, based on the Markov chains [4], Perron-Frobenius operators [5] and the PageRank algorithm of the corresponding Google matrix [6,7], give solid mathematical grounds for analysis of information flow on the Wikipedia network. In this work we perform the Google matrix analysis of Wikipedia network of English articles extending the results presented in [8,9],[10,11]. The main new element of this work is the study of time evolution of Wikipedia network during the years 2003 to 2011. We analyze how the ranking of Wikipedia articles and the spectrum of the Google matrix G of Wikipedia are changed during this period.

The directed network of Wikipedia articles is constructed in a usual way: a directed link is formed from an article j to an article i when j quotes i and an element A_{ij} of the adjacency matrix is taken to be unity when there is such a link and zero in absence of link. Then the matrix S_{ij} of Markov transitions is constructed by normalizing elements of each column to unity ($\sum_j S_{ij} = 1$) and replacing columns with only zero elements (*dangling nodes*) by $1/N$, with N being the matrix size. Then the Google matrix of the network takes the form [6,7]:

$$G_{ij} = \alpha S_{ij} + (1-\alpha)/N \quad . \tag{1}$$

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 $\alpha \approx 0.85$ [7]. The matrix G belongs to the class of Perron-Frobenius operators [5,7], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \le \alpha$. 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. It is possible to rank all nodes in a decreasing order of PageRank probability $P(K(i))$ so that the PageRank index $K(i)$ counts all N nodes i according their ranking, placing the most popular articles or nodes at the top values $K = 1, 2, 3....$

Due to the gap $1-\alpha \approx 0.15$ between the largest eigenvalue $\lambda = 1$ and other eigenvalues the PageRank algorithm permits an efficient and simple determination of the PageRank by the power iteration method [7]. It is also possible to use the powerful Arnoldi method [12,13],[14] to compute efficiently the eigenspectrum λ_i of the Google matrix:

$$\sum_{k=1}^{N} G_{jk}\psi_i(k) = \lambda_i \psi_i(j) \quad . \tag{2}$$

The Arnoldi method allows to find a several thousands of eigenvalues λ_i with maximal $|\lambda|$ for a matrix size N as large as a few tens of millions [10,11], [14,15]. Usually, at $\alpha = 1$ the largest eigenvalue $\lambda = 1$ is highly degenerate [15] due to many invariant subspaces which define many independent Perron-Frobenius operators providing (at least) one eigenvalue $\lambda = 1$.

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. This time inversion approach was used in [16,17] but the statistical properties and correlations between direct and inversed ranking were not analyzed there. In [18], on an example of the Linux Kernel network, it was shown thus this approach allows to obtain an additional interesting characterization of information flow on directed networks. Indeed, the right eigenvector of G^* at eigenvalue $\lambda = 1$ gives a probability $P^*(i)$, called CheiRank vector [8]. It determines a complementary rank index $K^*(i)$ of network nodes in a decreasing order of probability $P^*(K^*(i))$ [8, 9],[10,18]. It is known that the PageRank probability is proportional to the number of ingoing links characterizing how popular or known is a given node. In a similar way the CheiRank probability is proportional to the number of outgoing links highlighting the node communicativity (see e.g. [7,19], [20,21],[8,9]). The statistical properties of distribution of indexes $K(i), K^*(i)$ on the PageRank-CheiRank plane are described in [9].

In this work we apply the above mathematical methods to the analysis of time evolution of Wikipedia network ranking using English Wikipedia snapshots dated by December 31 of years 2003, 2005, 2007, 2009, 2011. In addition we use the snapshot of August 2009 (200908) analyzed in [8]. The parameters of networks with the number of articles (nodes) N, number of links N_ℓ and other information are given in Tables 1,2 with the description of notations given in Appendix.

The paper is composed as following: the statistical properties of PageRank and CheiRank are analyzed in Section 2, ranking of Wikipedia personalities and universities are considered in Sections 3, 4 respectively, the properties of spectrum of Google matrix are considered in Section 5, the discussion of the results is presented in Section 6, Appendix Section 7 gives network parameters.

2 CheiRank versus PageRank

The dependencies of PageRank and CheiRank probabilities $P(K)$ and $P^*(K^*)$ on their indexes K, K^* at different years are shown in Fig. 1. The top positions of K are occupied by countries starting from *United States* while at the top positions of K^* we find various listings (e.g. geographical names, prime ministers etc.; in 2011 we have appearance of listings of listings). Indeed, the countries accumulate links from all types of human activities and nature, that make them most popular Wikipedia articles, while listings have the largest number of outgoing links making them the most communicative articles.

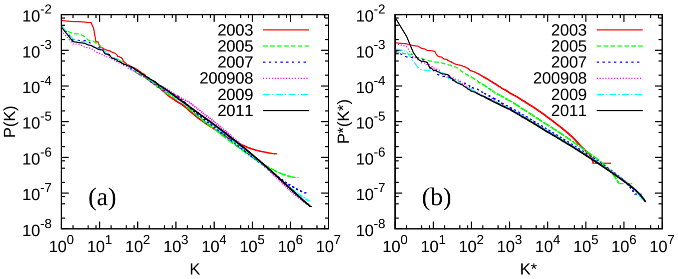

Fig. 1. PageRank probability $P(K)$ (left panel) and CheiRank probability $P^*(K^*)$ (right panel) are shown as a function of the corresponding rank indexes K and K^* for English Wikipedia articles at years 2003, 2005, 2007, 200908, 2009, 2011; here the damping factor is $\alpha = 0.85$.

The data of Fig. 1 show that the global behavior of $P(K)$ remains stable from 2007 to 2011. The probability $P^*(K^*)$ is stable in the time interval 2007 - 2009 while at 2011 we see the appearance of peak at $1 \leq K^* < 10$ that is related to introduction of listings of listings which were absent at earlier years. At the same time the behavior of $P^*(K^*)$ in the range $10 \leq K^* \leq 10^6$ remains stable for 2007 - 2011.

Each article i has its PageRank and CheiRank indexes $K(i)$, $K^*(i)$ so that all articles are distributed on two-dimensional plane of PageRank-CheiRank indexes. Following [8,9] we present the density of articles in the 2D plane (K, K^*) in Fig. 2. The density is computed for 100×100 logarithmically equidistant cells which cover the whole plane (K, K^*) for each year. The density distribution is globally stable for years 2007-2011 even if there are articles which change their location in 2D plane. We see an appearance of a mountain like ridge of probability along a line $\ln K^* \approx \ln K + 4.6$ that indicate the presence of correlation between $P(K(i))$ and $P^*(K^*(i))$. Following [8,9, 18] we characterize the interdependence of PageRank and CheiRank vectors by the correlator

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

We find the following values of the correlator at various time slots: $\kappa = 2.837(2003), 3.894(2005), 4.121(2007), 4.084(200908), 6.629(2009), 5.391(2011)$. During that period the size of the network increased almost by 10 times while κ increased less than 2 times. This confirms the stability of the correlator κ during the time evolution of the Wikipedia network.

In the next two Sections we analyze the time variation of ranking of personalities and universities.

3 Ranking of personalities

To analyze the time evolution of ranking of Wikipedia personalities (persons or humans) we chose the top 100 persons appearing in the ranking list of Wikipedia 200908

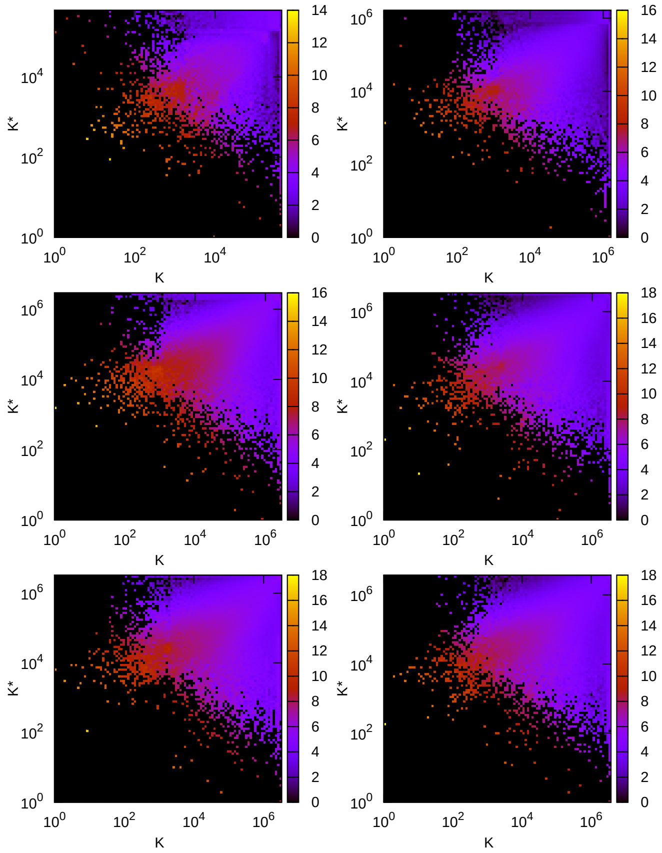

Fig. 2. Density of Wikipedia articles in the CheiRank versus PageRank plane at different years. Color is proportional to logarithm of density changing from minimal nonzero density (dark) to maximal one (white), zero density is shown by black (distribution is computed for 100×100 cells equidistant in logarithmic scale; bar shows color variation of natural logarithm of density); left column panels are for years 2003, 2007, 200908 and right column panels are for 2005, 2009, 2011 (from top to bottom).

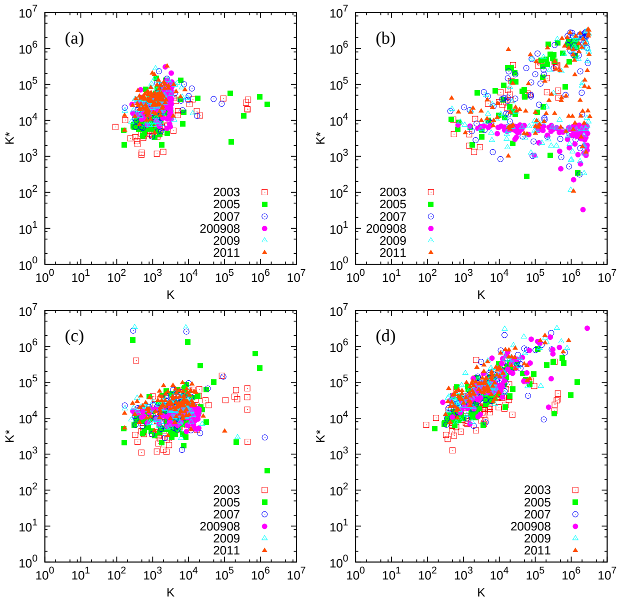

Fig. 3. Change of locations of top-rank persons of Wikipedia in K-K* plane. Each list of top ranks is determined by data of top 100 personalities of time slot 200908 in corresponding rank. Data sets are shown for (a) PageRank, (b) CheiRank, (c) 2DRank, (d) rank from Hart [22].

given in [8] in order of PageRank, CheiRank and 2DRank. We remind that 2DRank K_2 is obtained by counting nodes in order of their appearance on ribs of squares in (K, K^*) plane with their size growing from $K = 1$ to $K = N$ [8].

The distributions of personalities in PageRank-CheiRank plane is shown at various time slots in Fig. 3. There are visible fluctuations of distribution of nodes for years 2003, 2005 when the Wikipedia size has rapid growth. For other years the distribution of top 100 nodes of PageRank and 2DRank is stable even if individual nodes change their ranking. For top 100 of CheiRank the fluctuations remain strong during all years. Indeed, the number of outgoing links is more easy to be modified by authors writing a given article, while a modification of ingoing links depends on authors of other articles.

In Fig. 3 we also show the distribution of top 100 personalities from Hart's book [22] (the list of names is also available at the web page [8]). This distribution also remains stable in years 2007-2011. It is interesting to note that while top PageRank and 2DRank nodes form a kind of droplet in (K, K^*) plane, the distribution of Hart's personalities approximately follows the ridge along the line $\ln K^* \approx \ln K + 4.6$.

The time evolution of top 10 personalities of slot 200908 is shown in Fig. 4 for PageRank and 2DRank. For PageRank the main part of personalities keeps their rank position in time, e.g. G.W.Bush remains at first-second position. B.Obama significantly improves his ranking as a result of president elections. There are strong variations for Elizabeth II which we relate to modification of article name during the considered time interval. We also see a steady improvement of ranking of C.Linnaeus that we attribute to a growth of various botanic descriptions and listings at Wikipedia articles which quote his name. For 2DRank we observe stronger variations of K_2 index with time. Such a politician as R.Nixon has increasing K_2 index with time since the period of his presidency goes in the past. At the same time singers and artists remain at approximately constant level of K_2.

In [8] it was pointed out that the top personalities of PageRank are dominated by politicians while for 2DRank the dominant component of human activity is represented by artists. We analyze the time evolution of the distribution of top 30 personalities over 6 categories of human activity (*politics, arts, science, religion, sport and etc (or others)*). The category *etc* contains only C.Columbus. The results are presented in Fig. 5. They clearly show that the PageRank personalities are dominated by politicians whose percentage increases with time, while the percent of arts decreases. For 2DRank we see that the arts are dominant even if their percentage decreases with time. We also see the appearance of sport which is absent in

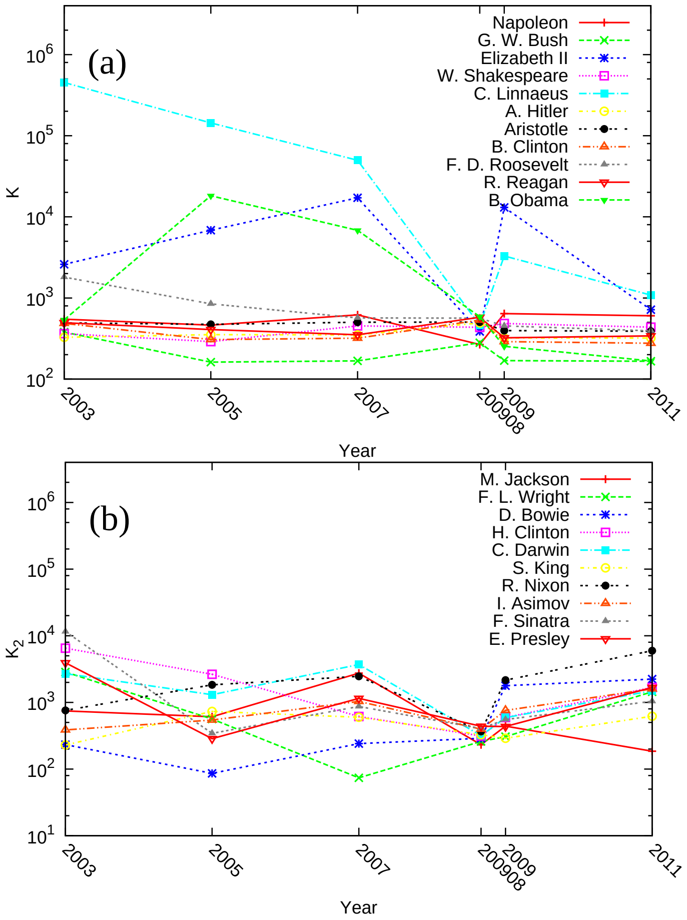

Fig. 4. Time evolution of top 10 personalities of year 200908 in indexes of PageRank K (a) and 2DRank K_2 (b); B.Obama is added in panel (a).

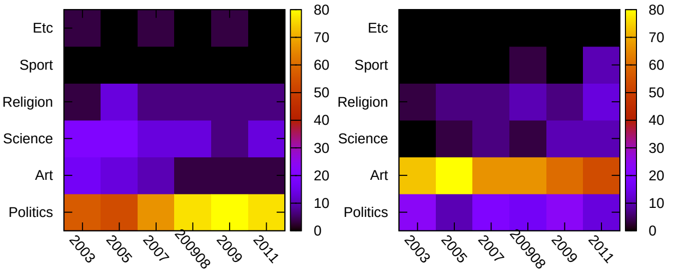

Fig. 5. Left panel: distribution of top 30 PageRank personalities over 6 activity categories at various years of Wikipedia. Right panel: distribution of top 30 2DRank personalities over the same activity categories at same years. Categories are politics, art, science, religion, sport, etc (other). Color shows the number of personalities for each activity expressed in percents.

PageRank. The mechanism of the qualitative ranking differences between two ranks is related to the fact that 2DRank takes into account via CheiRank a contribution of outgoing links. Due to that singers, actors, sportsmen increase their ranking since they are listed in various music albums, movies sport competition results. Due to that the component of arts gets higher positions in 2DRank in contrast to politics dominance in PageRank. Thus the two-dimensional ranking on PageRank-CheiRank plane allows to select qualities of nodes according to their popularity and communicativity.

4 Ranking of universities

The local ranking of top 100 universities is shown in Fig. 6 for years 2003, 2005, 2007 and in Fig. 7 for 2009, 200908, 2011. The local ranking is obtained by selecting top 100 universities appearing in PageRank listing so that they get their university ranking K from 1 to 100. The same procedure is done for CheiRank listing of universities obtaining their local CheiRank index K^* from 1 to 100. Those universities which enter inside 100×100 square on the local index plane (K, K^*) are shown in Figs. 6, 7.

The data show that the top PageRank universities are rather stable in time, e.g. U Harvard is always on the first top position. At the same time the positions in K^* are strongly changing in time. To understand the origin of this variations in CheiRank we consider the case of U Cambridge. Its Wikipedia article in 2003 is rather short but it contains the list of all 31 Colleges with direct links to their corresponding articles. This leads to a high position of U Cambridge with university $K^* = 4$ in 2003 (Fig. 8). However, with time the direct links remain only to about 10 Colleges while the whole number of Colleges are presented by a list of names without links. This leads to a significant increase of index up to $K^* \approx 40$ at Dec 2009. However, at Dec 2011 U Cambridge again improves significantly its CheiRank obtaining $K^* = 2$. The main reason of that is the appearance of section of "Notable alumni and academics" which provides direct links to articles about outstanding scientists studied and/or worked at U Cambridge that leads to second position at $K^* = 2$ among all universities. We note that in 2011 the top CheiRank University is George Mason University with university $K^* = 1$. The main reason of this high ranking is the presence of detailed listings of alumni in politics, media, sport with direct links to articles about corresponding personalities (including former director of CIA). These two examples show that the links, kept with a large number of university alumni, significantly increase CheiRank position of university. We note that artistic and politically oriented universities usually preserve more links with their alumni.

The time evolution of global ranking of top 10 universities of year 200908 for PageRank and 2DRank is shown in Fig. 8. The results show the stability of PageRank order with a clear tendency of top universities (e.g. Harvard) to go with time to higher and higher top positions of K. Thus for U Harvard the global value of K changes from $K \approx 300$ in 2003 to $K \approx 100$ in 2011, while the whole size N of the Wikipedia network increases almost by a factor 10 during this time interval. Since Wikipedia ranks all human knowledge, the stable improvement of PageRank indexes of universities reflects the global growing impor-

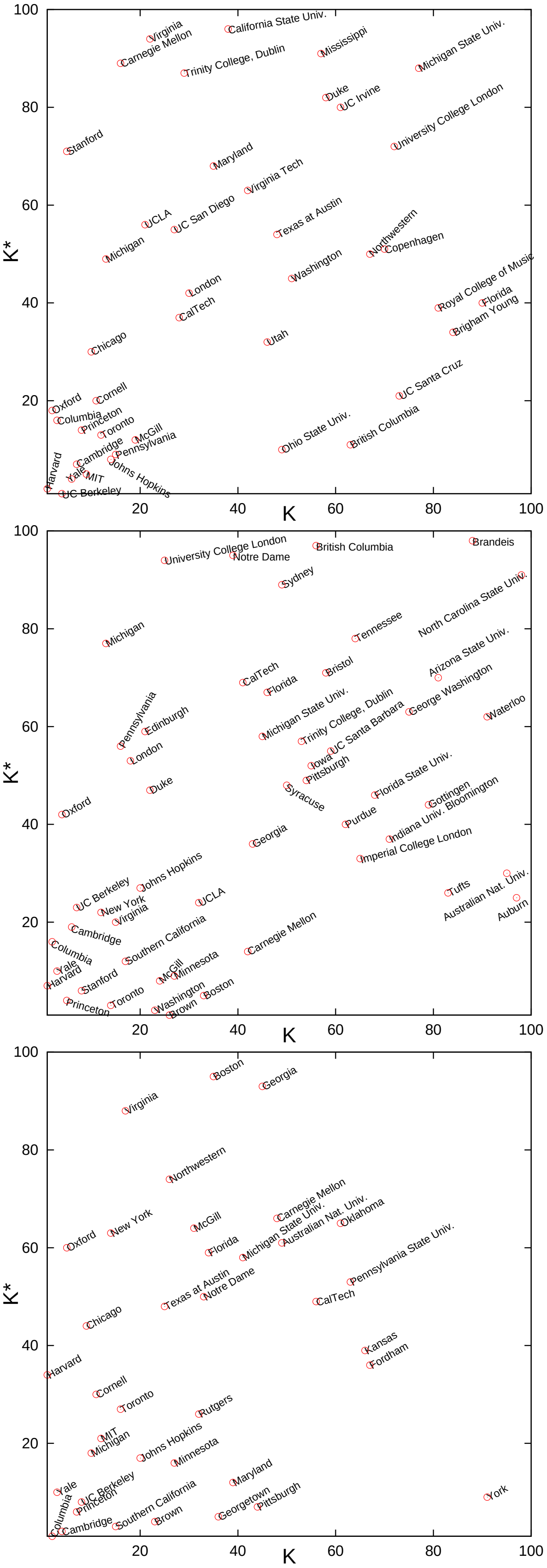

Fig. 6. University of Wikipedia articles in the local CheiRank versus PageRank plane at different years; panels are for years 2003, 2005, 2007 (from top to bottom).

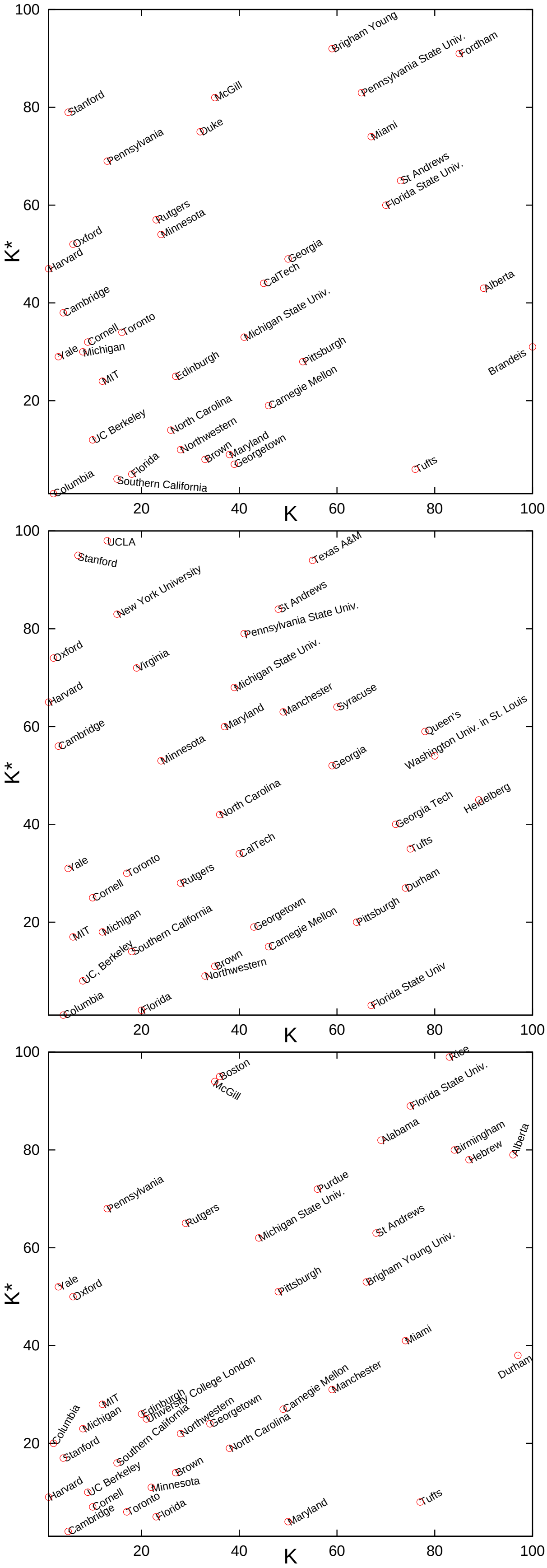

Fig. 7. Same as in Fig. 6 for years 2009, 200908, 2011 (from top to bottom).

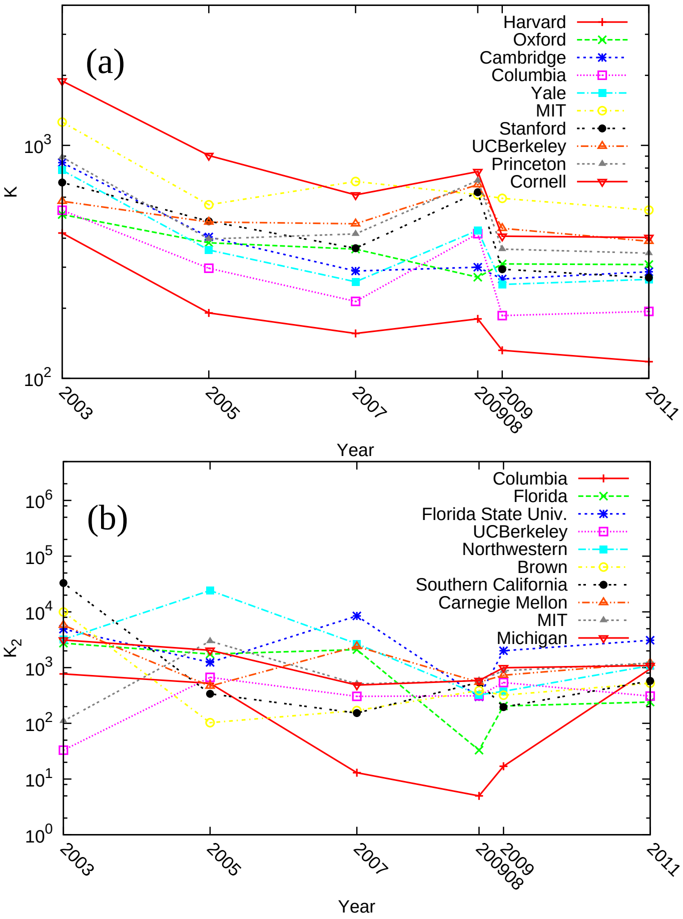

Fig. 8. Time evolution of global ranking of top 10 Universities of year 200908 in indexes of PageRank K (a) and 2DRank K_2 (b).

tance of universities in the world of human activity and knowledge.

The time evolution of the same universities in 2DRank remains stable in time showing certain interchange of their ranking order. We think that an example of U Cambridge considered above explains the main reasons of these fluctuations. In view of 10 times increase of the whole network size during the period 2003 - 2011 the average stability of 2DRank of universities also confirms the significant importance of their place in human activity.

Finally we compare the Wikipedia ranking of universities in their local PageRank index K with those of Shanghai university ranking [23]. In the top 10 of Shanghai university rank the Wikipedia PageRank recovers 9 (2003), 9 (2005), 8 (2007), 7 (2009), 7 (2011). This shows that the Wikipedia ranking of universities gives the results being very close to the real situation. A small decrease of overlap with time can be attributed to earlier launched activity of leading universities on Wikipedia.

5 Google matrix spectrum

Finally we discuss the time evolution of the spectrum of Wikipedia Google matrix taken at $\alpha = 1$. We perform the numerical diagonalization based on the Arnoldi method [12,13] using the additional improvements described in [14,15] with the Arnold dimension $n_A = 6000$. The Google matrix is reduced to the form

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix} \qquad (4)$$

where S_{ss} describes disjoint subspaces V_j of dimension d_j invariant by applications of S; S_{cc} depicts the remaining part of nodes forming the wholly connected *core space*. We note that S_{ss} is by itself composed of many small diagonal blocks for each invariant subspace and hence those 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 (See also Appendix). 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. Here we restrict ourselves to the statistical analysis of the spectrum λ_i. The analysis of eigenstates ψ_i ($G\psi_i = \lambda_i\psi_i$), which has been done in [11] for the slot 200908, is left for future studies.

The spectrum for all Wikipedia time slots is shown in Fig. 9 for G and in Fig. 10 for G^*. We see that the spectrum remains stable for the period 2007 - 2001 even if there is a small difference of slot 200908 due to a slightly different cleaning link procedure (see Appendix). For the spectrum of G^* in 2007 - 2001 we observe a well pronounced 3-6 arrow star structure. This structure is very similar to those found in random unistochastic matrices of side 3-4 [24] (see Fig.4 therein). This fact has been pointed in [11] for the slot 200908. Now we see that this is a generic phenomenon which remains stable in time. This indicates that there are dominant groups of 3-4 nodes which have structure similar to random unistochastic matrices with strong ties between 3-4 nodes and various random permutations with random hidden complex phases. The spectral arrow star structure is significantly more pronounce for the case of G^* matrix. We attribute this to more significant fluctuations of outgoing links that probably makes sectors of G^* to be more similar to elements of unistochastic matrices. A further detailed analysis will be useful to understand these arrow star structure and its links with various communities inside Wikipedia.

As it is shown in [11] the eigenstates of G and G^* select certain well defined communities of the Wikipedia network. Such an eigenvector detection of the communities provides a new method of communities detection in addition to more standard methods developed in network science and described in [25]. However, the analysis of eigenvectors represents a separate detailed research and in this work we restrict ourselves to PageRank and CheiRank vectors.

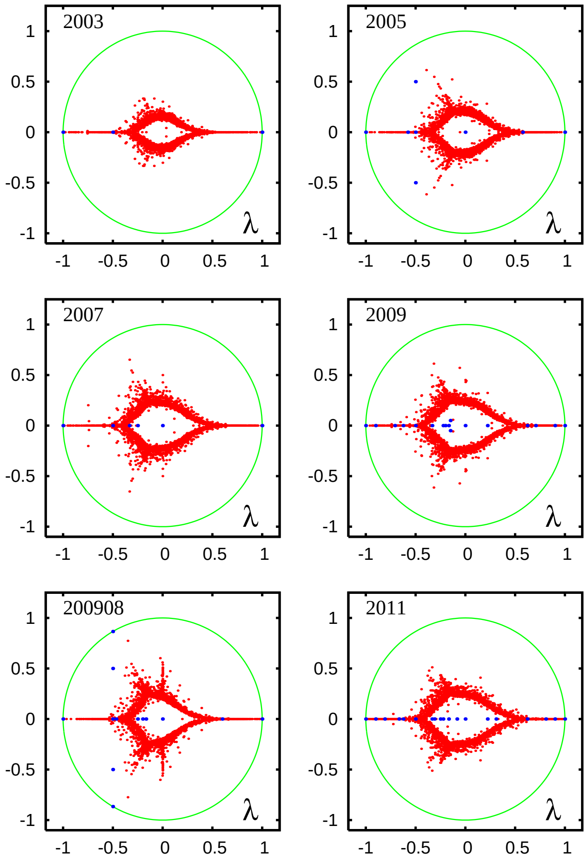

Fig. 9. Spectrum of eigenvalues λ of the Google matrix G of Wikipedia at different years. Red dots are core space eigenvalues, blue dots are subspace eigenvalues and the full green curve shows the unit circle. The core space eigenvalues were calculated by the projected Arnoldi method with Arnoldi dimensions $n_A = 6000$.

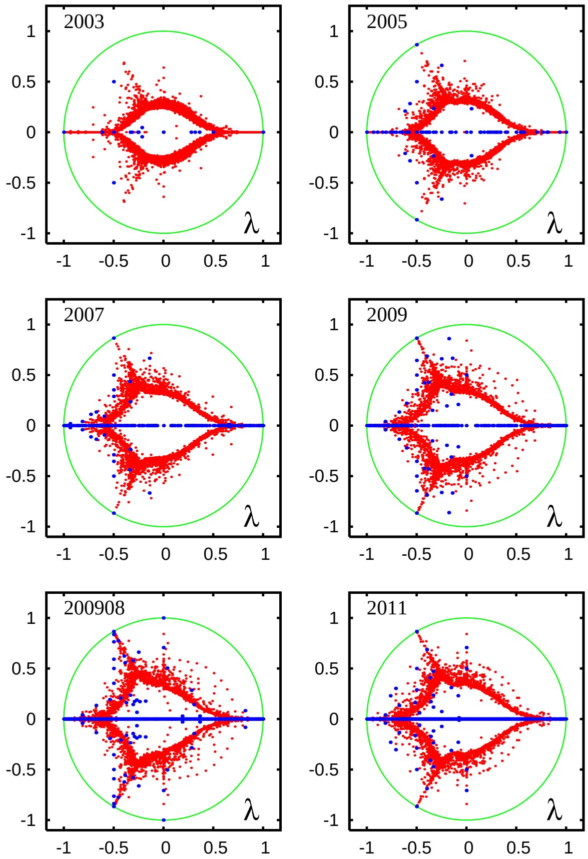

Fig. 10. Same as in Fig. 9 but for the spectrum of matrix G^*.

Finally we note that the fraction of isolated subspaces is very small for G matrix. It is increased approximately by a factor of order 10 for G^* but still it remains very small compared to the networks of UK universities analyzed in [15]. This fact reflects a strong connectivity of network of Wikipedia articles.

6 Discussion

In this work we analyzed the time evolution of ranking of network of English Wikipedia articles. Our study demonstrates the stability of such statistical properties as PageRank and CheiRank probabilities, the article density distribution in PageRank-CheiRank plane during the period 2007 - 2011. The analysis of human activities in different categories shows that PageRank gives main accent to politics while the combined 2DRank gives more importance to arts. We find that with time the number of politicians in the top positions increases. Our analysis of ranking of universities shows that on average the global ranking of top universities goes to higher and higher positions. This clearly marks the growing importance of universities for the whole range of human activities and knowledge. We find that Wikipedia PageRank recovers 70 - 80 % of top 10 universities from Shanghai ranking [23]. This confirms the reliability of Wikipedia ranking.

We also find that the spectral structure of the Wikipedia Google matrix remains stable during the time period 2007 -2011 and show that its arrow star structure reflects certain features of small size unistochastic matrices.

Acknowledgments: 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 allocations 2012-P0110, 2013-P0110. We also acknowledge the France-Armenia collaboration grant CNRS/SCS *No* 24943 (IE-017) on “Classical and quantum chaos”.

7 Appendix

The tables with all network parameters used in this work 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

	N	N_ℓ	n_A
2003	455436	2033173	6000
2005	1635882	11569195	6000
2007	2902764	34776800	6000
2009	3484341	52846242	6000
200908	3282257	71012307	6000
2011	3721339	66454329	6000

Table 1. Parameters of all Wikipedia networks at different years considered in the paper.

	N_s	N_d	$d_{\max}$	$N_{\text{circ.}}$	N_1
2003	15	7	3	11	7
2003*	940	162	60	265	163
2005	152	97	4	121	97
2005*	5966	1455	1997	2205	1458
2007	261	150	6	209	150
2007*	10234	3557	605	5858	3569
2009	285	121	8	205	121
2009*	11423	4205	134	7646	4221
200908	515	255	11	381	255
200908*	21198	5355	717	8968	5365
2011	323	131	8	222	131
2011*	14500	4637	1323	8591	4673

Table 2. G and G^* eigespectrum parameters for all Wikipedia networks, year marks spectrum of G, year with star marks spectrum of G^*.

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 $N_s \geq d_{\max}$. The data for G are marked by the corresponding year of the time slot, the data for G^* are marked by the year with a star. Links cleaning procedure eliminates all redirects (nodes with one outgoing link), this procedure is slightly different from the one used for the slot 200908 in [8]. All data sets and high resolution figures are available at the web page [26].

References

1. Wikipedia, *Wikipedia*, `en.wikipedia.org/wiki/Wikipedia`
2. F.A. Nielsen, *Wikipedia research and tools: review and comments*, (2012), available at SSRN: `dx.doi.org/10.2139/ssrn.2129874`
3. J.L. Borges, *The Library of Babel* in *Ficciones* (Grove Press. N.Y. 1962).
4. A.A. Markov, *Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga*, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, **15** (1906) 135 (in Russian) [English trans.: *Extension of the limit theorems of probability theory to a sum of variables connected in a chain* reprinted in Appendix B of: R.A. Howard *Dynamic Probabilistic Systems*, volume 1: *Markov models*, Dover Publ. (2007)]
5. M. Brin and G. Stuck, *Introduction to dynamical systems*, Cambridge Univ. Press, Cambridge, UK (2002)
6. S. Brin and L. Page, Computer Networks and ISDN Systems **30**, 107 (1998)
7. A. M. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton (2006)
8. A.O.Zhirov, O.V.Zhirov and D.L.Shepelyansky, Eur. Phys. J. B **77**, 523 (2010); `www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/`
9. L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, J. Phys. A: Math. Theor. **45**, 275101 (2012); `www.quantware.ups-tlse.fr/QWLIB/dvvadi/`
10. K.M.Frahm and D.L.Shepelyansky, Eur. Phys. J. B **85**, 355 (2012); `www.quantware.ups-tlse.fr/QWLIB/twittermatrix/`
11. L.Ermann, K.M. Frahm and D.L.Shepelyansky, *Spectral properties of Google matrix of Wikipedia and other networks*, arXiv:1212.1068 [cs.IR] (2012) (Eur. Phys. J. B in press).
12. G.W. Stewart, *Matrix Algorithms Eigensystems*, (SIAM, 2001), Vol. II
13. G.H. Golub and C. Greif, *BIT Num. Math.* **46**, 759 (2006)
14. K.M. Frahm and D.L. Shepelyansky, *Eur. Phys. J. B* **76**, 57 (2010)
15. K.M.Frahm, B.Georgeot and D.L.Shepelyansky, J. Phys, A: Math. Theor. **44**, 465101 (2011)
16. D. Fogaras, *Lect. Notes Comp. Sci.* **2877**, 65 (2003)
17. V. Hrisitidis, H. Hwang and Y. Papakonstantino, *ACM Trans. Database Syst.* **33**, 1 (2008)
18. A. D. Chepelianskii, *Towards physical laws for software architecture*, arXiv:1003.5455[cs.SE] (2010); `www.quantware.ups-tlse.fr/QWLIB/linuxnetwork/`
19. D. Donato, L. Laura, S. Leonardi and S. Millozzi, Eur. Phys. J. B **38**, 239 (2004)
20. G. Pandurangan, P. Raghavan and E. Upfal, Internet Math. **3**, 1 (2005)
21. N. Litvak, W.R.W. Scheinhardt, and Y. Volkovich, Lecture Notes in Computer Science, **4936**, 72 (2008).
22. M.H. Hart, *The 100: ranking of the most influential persons in history*, Citadel Press, N.Y. (1992).
23. `www.shanghairanking.com/`
24. K. Zyczkowski, M. Kus, W. Slomczynski and H.-J. Sommers, J. Phys. A: Math. Gen. **36**, 3425 (2003)
25. S. Fortunato, Phys. Rep. *486*, 75 (2010)
26. `www.quantware.ups-tlse.fr/QWLIB/wikirankevolution/`

Eur. Phys. J. B (2013) 86: 322
DOI: 10.1140/epjb/e2013-40120-6

THE EUROPEAN
PHYSICAL JOURNAL B

Regular Article

Poincaré recurrences and Ulam method for the Chirikov standard map

K.M. Frahm[a] and D.L. Shepelyansky[a,b]

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

Received 12 February 2013 / Received in final form 15 April 2013
Published online 15 July 2013 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2013

Abstract. We study numerically the statistics of Poincaré recurrences for the Chirikov standard map and the separatrix map at parameters with a critical golden invariant curve. The properties of recurrences are analyzed with the help of a generalized Ulam method. This method allows us to construct the corresponding Ulam matrix whose spectrum and eigenstates are analyzed by the powerful Arnoldi method. We also develop a new survival Monte Carlo method which allows us to study recurrences on times changing by ten orders of magnitude. We show that the recurrences at long times are determined by trajectory sticking in a vicinity of the critical golden curve and secondary resonance structures. The values of Poincaré exponents of recurrences are determined for the two maps studied. We also discuss the localization properties of eigenstates of the Ulam matrix and their relation with the Poincaré recurrences.

1 Introduction

The interest to understanding of transition from dynamical to statistical description of motion had started from the dispute between Loschmidt and Boltzmann, which is now known as the Loschmidt paradox [1,2]. The two-dimensional (2D) symplectic maps represent an excellent laboratory for investigation of how statistical laws appear in dynamical, fully deterministic systems. Their properties have been studied in great detail during last decades both on mathematical (see e.g. [3,4] and references therein) and physical (see e.g. [5–7] and references therein) levels of rigor. The case of completely chaotic behavior, appearing e.g. in Anosov systems, is now well understood [3,4] but a generic case of maps with divided phase space, where islands of stability are surrounded by chaotic components, still preserves its puzzles. A typical example of such a map is the Chirikov standard map [5,6] which often gives a local description of dynamical chaos in other dynamical maps and describes a variety of physical systems (see e.g. [8]). This map has the form:

$$\bar{y} = y + \frac{K}{2\pi}\sin(2\pi x), \quad \bar{x} = x + \bar{y} \pmod 1. \tag{1}$$

Here x, y are canonical conjugated variables of generalized phase and action, bars mark the variables after one map iteration and we consider the dynamics to be periodic on a torus so that $0 \le x \le 1$, $0 \le y \le 1$. The dynamics is characterized by one dimensionless chaos parameter K.

For small values of K the phase space is covered by invariant Kolmogorov-Arnold-Moser (KAM) curves which restrict dynamics in the action variable y. For $K > K_g$ the last invariant golden curve with the rotation number $r = r_g = \langle (x_t - x_0)/t\rangle = (\sqrt{5}-1)/2$ is destroyed [9,10] and it is believed that for $K > K_g$ the dynamics in y becomes unbounded [11,12]. A renormalization technique developed by Greene [9] and MacKay [10] allowed to determine $K_g = 0.971635406$ with enormous precision (due to symmetry there is also a symmetric critical curve at $r = 1 - r_g$ at K_g). The properties of the critical golden curve on small scales are universal for all critical curves with the golden tail of the continuous fraction expansion of r for all smooth 2D symplectic maps [10]. Here and below the time t is measured in number of map iterations. For $K > K_g$ the golden KAM curve is replaced by a cantorus [13] which can significantly affect the diffusive transport through the chaotic part of the phase space [14,15]. There are numerical and analytical indications that at any K there are some chaotic regions in the phase space bounded by internal invariant curves; at $K < K_g$ there are isolating invariant curves.

The dynamics inside a chaotic component of the phase space (x, y) is characterized by correlation functions whose decay ensures a transition from dynamical to statistical description. The decay of correlations is related to the probability to stay in a given region of phase space since for a trajectory remaining in a small region the dynamical variables are strongly correlated. This probability in its own turn is related to the statistics of Poincaré recurrences. Indeed, according to the Poincaré recurrence theorem [16] a volume preserving dynamical flow with only bounded orbits has for each open set orbits that intersect the set infinitely often. Such orbits return, after a certain time, to a close vicinity of an initial state. However, the statistics of these recurrences depends on dynamical

[a] http://www.quantware.ups-tlse.fr
[b] e-mail: dima@irsamc.ups-tlse.fr

properties of the system. For a fully chaotic phase space a probability to stay in a certain part of a phase space decays exponentially with time being similar to a random coin flipping [3,4]. However, in dynamical maps with divided phase space, like the Chirikov standard map, the extensive numerical simulations show that the decay of probability of Poincaré recurrences $P(t)$ is characterized by a power law decay $P(t) \propto 1/t^\beta$ has $\beta \approx 1.5$ whose properties still remain poorly understood.

One of the first studies of Poincaré recurrences in dynamical Hamiltonian systems with two degrees of freedom was done in reference [17] where an algebraic decay with an exponent $\beta = 1/2$ was found. This exponent corresponds to an unlimited diffusion on an infinite one-dimensional line which is in contrast to a bounded phase space. This strange observation was explained in references [18,19] as a diffusion in a chaotic separatrix layer of a nonlinear resonance which takes place on relatively short diffusion times. On larger times, which were not accessible to the computations presented in reference [17], this diffusion becomes bounded by a finite width of the separatrix layer and a universal algebraic decay takes place with the exponent $\beta \approx 1.5$ corresponding to a finite chaos measure [18,19]. This algebraic decay of $P(t)$ has been confirmed by various groups in various Hamiltonian systems [20–31].

One can argue that such a slow algebraic decay with $\beta \approx 1.5$ appears due to trajectory sticking near stable islands and critical invariant curves and leads to an even slower correlation function decay $C(t) \sim tP(t)$ with a divergence of certain second moments. A sticking in a vicinity of the critical golden curve [10] is expected to give $\beta \approx 3$ [24,25], being significantly larger than the average value $\beta \approx 1.5$. A certain numerical evidence is presented in reference [27] showing that long time sticking orbits can be trapped not only in a vicinity of a critical golden curve but also in internal chaotic layers of secondary resonances.

Theoretical attempts to describe trapping in secondary resonances as renormalization dynamics on some Cayley type tree was started in references [22,23] with recent extensions done in references [28,32,33]. However, a detailed understanding of the intriguing features of Poincaré recurrences in the Chirikov standard map and other similar maps is still missing.

In this work we use a generalized Ulam method developed in references [34,35] and combine it with a new survival Monte Carlo method trying to reach larger time scales and to obtain a better understanding of statistics of Poincaré recurrences in the Chirikov standard map and the separatrix map.

The paper is composed as follows: in Section 2 we construct the Ulam matrix based on the generalized Ulam method and study the properties of its spectrum, eigenstates and corresponding time evolution for the case of the Chirikov standard map. The survival Monte Carlo method is introduced in Section 3 and the properties of the Poincaré recurrences are studied with its help comparing results with the Ulam method. In Section 4 we apply the above methods to the separatrix map and in Section 5 the localization properties of the eigenstates of the Ulam matrix are analyzed. The discussion of the results is presented in Section 6.

2 Generalized Ulam method with absorption

The Ulam method was proposed in 1960 [36]. In the original version of this method a 2D phase space is divided in $N_d = M \times M$ cells and n_c trajectories are propagated on one map iteration from each cell j. Then the matrix S_{ij} is defined by the relation $S_{ij} = n_{ij}/n_c$ where n_{ij} is the number of trajectories arriving from a cell j to a cell i. By construction we have $\sum_i S_{ij} = 1$ and hence the matrix S_{ij} belongs to the class of the Perron-Frobenius operators (see e.g. [37]). This Ulam matrix can be considered as a discrete Ulam approximate of the Perron-Frobenius operator (UPFO) of the continuous dynamics.

According to the Ulam conjecture [36] the UPFO converges to the continuous limit at large M. Indeed, this conjecture was proven for 1D homogeneously chaotic maps [38]. Various properties of the UPFO for 1D and 2D maps are analyzed in references [39–42]. Recent studies [43,44] demonstrated similarities between the UPFO, the corresponding Ulam networks and the properties of the Google matrix of the world wide web networks. It was shown that in maps with absorption or dissipation the spectrum of the UPFO is characterized by the fractal Weyl law [45].

The coarse-grained cell structure of the original Ulam method corresponds to an effective noise and in case of a divided phase space the noise induces an artificial diffusion between chaotic and regular regions. In reference [34] this problem was solved by replacing the random initial points by a very long chaotic trajectory and the transitions between cells are accumulated along the chaotic trajectory that keeps the invariant curves and stable islands even in presence of the effective noise. Furthermore, the matrix size is also reduced since only cells which are visited at least once by the trajectory are kept. Here we use this approach for the analysis of the Poincaré recurrences keeping the same notations as in reference [34]. In particular, as in reference [34], we exploit the parity symmetry $x \to 1 - x$ and $y \to 1-y$ allowing to limit the effective phase space to $0 \le x \le 1$, $0 \le y \le 0.5$ and therefore reducing the number of cells at a given cell size by a factor of two. In x direction we use therefore M cells and in y direction $M/2$ cells with $M \in \{25, 35, \ldots, 1120, 1600\}$ and the intermediate values are multiples of 25 or 35 by powers of 2.

To study the Poincaré recurrences within the Ulam method we introduce absorption of all trajectories with $y < y_{cut} = 0.05$. The measure of the phase space where the absorption takes place is relatively small (only a few percents of the whole phase space). Thus the absorption does not significantly affect the dynamics of trajectories sticking for long times. Indeed, we will see that the probability decay due to absorption reproduces the decay of Poincaré recurrences in a closed system. At the same time this absorption leads to a survival probability decay and allows us to use efficiently the Ulam method for the analysis of Poincaré recurrences. We generate the matrix S

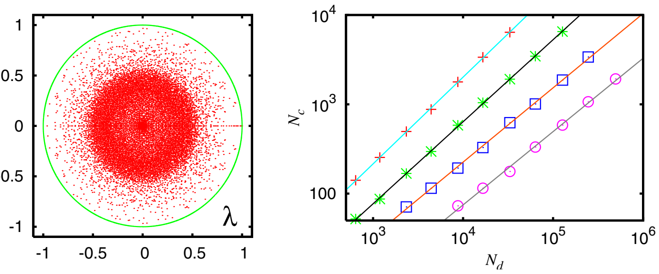

Fig. 1. The left panel shows the eigenvalue spectrum λ_j for the projected case of the UPFO of map (1) at $K = K_g$ in the complex plane for $M = 280$ and $N_d = 16609$ by red/gray dots (projected matrix dimension $N_p = 15\,457$). The green/gray curve represents the circle $|\lambda| = 1$. The right panel shows the number N_c of eigenvalues, with modulus larger than λ_c, versus N_d in a double logarithmic representation for $\lambda_c = 0.5$ (crosses), $\lambda_c = 0.66$ (stars), $\lambda_c = 0.8$ (open squares) and $\lambda_c = 0.9$ (open circles). The straight lines correspond to the power law fits $N_c \sim N_d^\nu$ with exponents $\nu = 0.971 \pm 0.006$ ($\lambda_c = 0.5$), $\nu = 0.919 \pm 0.005$ ($\lambda_c = 0.66$), $\nu = 0.832 \pm 0.010$ ($\lambda_c = 0.8$) and $b = 0.821 \pm 0.021$ ($\lambda_c = 0.9$). The fits are done for the data with $N_c > 50$, $M > 35$ and $M \leq 400$ ($\lambda_c = 0.5$), $M \leq 800$ ($\lambda_c = 0.66$), $M \leq 1120$ ($\lambda_c = 0.8$), $M \leq 1600$ ($\lambda_c = 0.9$), since the Arnoldi method provides only a partial spectrum of the eigenvalues with largest modulus for large values of M.

using one trajectory iterated by the map up to the iteration time $t = 10^{12}$ (as in Ref. [34]; this corresponds to the closed system without absorption and we call this the symplectic case). After that the matrix size N_d is simply reduced only to those cells with $y \geq y_{cut}$ that gives the projected matrix dimension N_p and matrix S_p. The matrix size of this projected case is smaller approximately by 7%. We find, for $M \leq 1600$, an approximate dependence $N_d \approx 0.39 M^2/2$ and $N_p \approx 0.36 M^2/2$. This corresponds to the usual estimate of the chaos measure being around 39% in agreement with the results of Chirikov [6] (see also [14]). For the symplectic case we have the maximal eigenvalue $\lambda = 1$ while in the projected case with absorption we are getting $|\lambda| < 1$.

The spectrum λ_j of the projected case with matrix S_p is shown in Figure 1. The spectrum is obtained by the direct diagonalization of the matrix S_p that can be done numerically up to $M = 280$. It can be compared with the corresponding spectrum of the symplectic system shown in Figure 2 of [34]. The global spectrum structure of S for the symplectic case is similar to the projected case. Indeed, the absorption is relatively weak and does not affect the global properties of motion. However, with absorption the measure is not conserved and the remaining non-escaping set forms a fractal set with the fractal dimension $d < 2$ (see e.g. [45,46]).

In the case of Ulam networks on fractal chaotic repellers the spectrum of UPFO S_p is characterized by the fractal Weyl law with the number of states N_c in the ring $\lambda_c < |\lambda| \leq 1$ growing with the matrix size N_d as $N_c \propto N_d^\nu$ (here for simplicity we use the size N_d of the symplectic case, for the projected case we have simply to change $N_p \approx 0.93 N_d$). It can be argued that the fractal dimension d_0 of the invariant repeller set determines the exponent $\nu = d_0/2$ [45]. Examples of dependencies N_c vs. N_d are given in Figure 1 for various values of λ_c. Definitely we have $\nu < 1$ but there is an evident dependence on λ_c with a decreasing value of ν at $\lambda_c \to 1$. We attribute this to the fact that at $\lambda_c \to 1$ we are dealing with long sticking trajectories whose measure decreases with time.

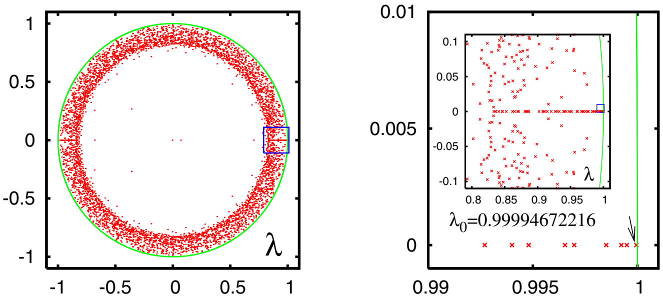

Fig. 2. Partial spectrum λ_j for the projected case of the UPFO of the map (1) at $K = K_g$ for $M = 1600$. The left panel shows all eigenvalues obtained by the Arnoldi method with $n_\mathrm{A} = 5000$. The insert of the right panel shows the blue/black square of the first zoomed range of the left panel; the blue/black square here is the second zoomed range shown in the main figure of the right panel. The eigenvalue with largest modulus $\lambda_0 = 0.99994672216$ is indicated by an arrow. The green/gray curve represents in all cases the circle $|\lambda| = 1$.

Here we should point out that the data for $M \geq 400$ corresponding to $N_d > 30\,000$ are obtained from the Arnoldi method [47] which allows us to find the eigenvalues for matrix sizes up to $N_d \sim 10^6$. However, only a finite number of eigenvalues with largest $|\lambda|$ can be determined numerically using $n_A = 12\,000, 8000, 8000, 6000, 5000$ (for $M = 400, 560, 800, 1120, 1600$, respectively and with n_A being the used Arnoldi dimension). A more detailed description of the Arnoldi method for the UPFO is given in reference [34]. An example of the spectrum λ obtained with the Arnoldi method at the largest value of $M = 1600$ is shown in Figure 2. Here $N_d = 49\,4964$ and $N_p = 45\,8891$. We find that the maximal eigenvalue for the projected case is $\lambda_0 = 0.99994672216$ corresponding to a slow escape rate at large times. As in reference [34] for the symplectic case without absorption, we obtain also for the case with absorption two type of eigenmodes: "diffusion modes" with real eigenvalues close to 1 and whose eigenvectors are rather extended in phase space (with some decay for cells close to the absorption border) and "resonant modes" with complex or real negative eigenvalues and which are quite well localized around a chain of stable islands close to an invariant curve. It turns out that many of the resonant modes (those "far" away from the absorption border), coincide numerically very well with corresponding resonant modes for the case without absorption already found in reference [34].

The dependence of the density of eigenvalues $\rho(|\lambda|)$ on $|\lambda|$ is shown in Figure 3. We see the proximity between the symplectic and projected cases not only in density ρ but also in a slow relaxation of the diffusion modes with relaxation rates $\gamma_j \approx \gamma_1 j^2$ ($\gamma_j = -2 \ln |\lambda_j|$) provided we identify γ_{j+1} of the symplectic case with γ_j of the projected

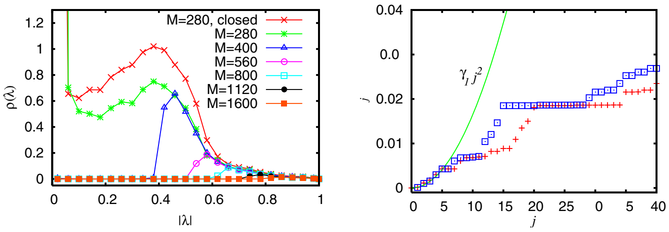

Fig. 3. The left panel shows the density $\rho(\lambda)$ of eigenvalues, being normalized by $\int \rho(\lambda)\, d^2\lambda = 1$, of the UPFO for the map (1) at $K = K_g$ in the complex plane as a function of the modulus $|\lambda|$ for $M = 280$ for the symplectic case (upper curve, crosses) and the projected case (lower green curve, stars). The other curves are partial (non-normalized) densities for the projected case and the values $M = 400$, 560, 800, 1120, 1600 and the number of used eigenvalues (obtained by the Arnoldi method) is $n_A = 12\,000$, 8000, 8000, 6000, 5000, respectively. The right panel shows the decay rates $\gamma_j = -2\ln(|\lambda_j|)$ versus level number j for the UPFO eigenvalues λ_j, with $M = 1600$ and $N_d = 494\,964$. The red/gray crosses correspond to the UPFO of symplectic case and the blue/black squares correspond to the projected case (data points for this case are shifted to one position to the right). The green curve corresponds to the quadratic dispersion law $\gamma_j \approx \gamma_1\, j^2$ which is approximately valid for the diffusion modes with $0 \le j \le 5$ and where γ_1 is taken from the UPFO of the symplectic case.

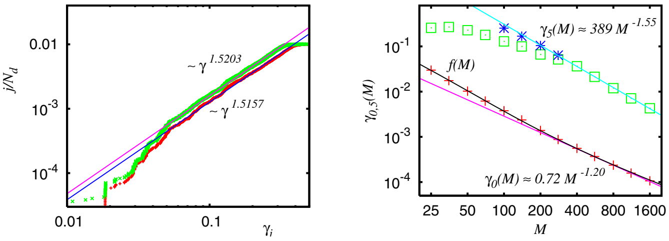

Fig. 4. The left panel shows the rescaled level number j/N_d versus the decay rate γ_j, in a double logarithmic scale, for the map (1) at K_g with $M = 1600$ and $N_d = 494\,964$. Red/lower data points correspond to the UPFO projected case and green/upper data points correspond to the UPFO symplectic case. The two straight lines correspond to the power law fits $j/N_d \approx 0.052745\,\gamma^{1.5203}$ (symplectic case) and $j/N_d \approx 0.041570\,\gamma^{1.5157}$ (projected case) for the data in the range $0.04 \le \gamma \le 0.3$. The statistical error bound of the exponents obtained from the fits is close to 0.1% in both cases. The right panel shows the decay rates $\gamma_j(M)$ for $j = 0$ (red crosses), $j = 5$ (green open squares) of the UPFO projected case in a double logarithmic scale. The lower/pink straight line corresponds to the power law fit $\gamma_0(M) \approx 0.72M^{-1.20}$ and the upper/light blue straight line to the fit $\gamma_5(M) \approx 389M^{-1.55}$ (both fits obtained for the range $400 \le M \le 1600$). The black/curved line corresponds to the other fit $\gamma_0(M) = f(M) = \frac{D}{M}\,\frac{1+C/M}{1+B/M}$ with $D = 0.162$, $C = 165$ and $B = 17.0$ (fit obtained for the range $25 \le M \le 1600$). We mention that γ_5 corresponds for $M \ge 400$ to a resonant mode whose eigenvector is strongly localized close to the three stable islands of the resonance 1/3. However, for $M \le 280$ γ_5 corresponds to a different mode and the resonant mode at 1/3 is associated to γ_7 ($M = 280$), γ_{13} ($M = 200$), γ_{17} ($M = 140$) and γ_{23} ($M = 100$) which are shown as four additional data points (blue stars).

case because γ_0 of the symplectic case is simply zero and the relaxation rate γ_1 to the ergodic state of the symplectic case corresponds roughly to the exponential long time escape rate γ_0 of the projected case. The proximity of the two cases is also well seen in the dependence of integrated density of states $\rho_\Sigma(\gamma) = j/N_d$ on γ_j shown in Figure 4 (here j is a number of eigenvalues with $\gamma \le \gamma_j$). In both cases we have the algebraic dependence $\rho_\Sigma(\gamma) \propto \gamma^\beta$ with $\beta \approx 1.5$. In reference [34] it was argued that this exponent is the same as for the exponent of decay of Poincaré recurrences $P(t)$. These data show that an introduction of small absorption at $y < y_{cut}$ does not produce significant modification for trajectories trapped for long times in a vicinity of the critical golden curve or other secondary islands located far away from the absorption band.

The slowest decay rates, such like γ_0 and γ_5, decrease algebraically with the increase of M as it is shown in right panel of Figure 4. In the fit range $400 \le M \le 1600$ we have a power law $\gamma_0(M) \approx 0.72\ M^{-1.20}$ but taking into account the curvature for the interval $25 \le M \le 1600$ the modified fit $\gamma_0(M) = \frac{D}{M}\,\frac{1+C/M}{1+B/M}$ with $D = 0.162$, $C = 165$ and $B = 17.0$ seems to indicate a behavior $\gamma_0(M) \propto M^{-1}$ in the limit $M \to \infty$. This behavior is similar to the one found in reference [34] for γ_1 in the symplectic case (where γ_0 is simply 0). On the other hand the resonant mode γ_5 obeys the power law $\gamma_5(M) \approx 389\, M^{-1.55}$ which is valid for the interval $100 \le M \le 1600$ if we use for the smaller values of M not γ_5 but the resonant mode localized to the same chain of resonant islands which may have a different eigenvalue index (see Fig. 4 for details). The comparison of these decays indicate that eventually at very large values of M, far outside the range numerically accessible by the Arnoldi method, the resonant modes become dominant over the diffusion modes. The limit $\gamma \to 0$ for $M \to \infty$ is related to long sticking trajectories near critical invariant curves which restrict the chaos component and whose phase space structure can be better resolved with decreasing cell size $1/M$. As in reference [34] we argue that these lowest modes are affected by the effective noise present in the Ulam method. Due to that we do not have a clear explanation for this algebraic decay. However, the fact that γ_j (at fixed value of j) vanishes with increasing M indicates that the limit $t_{\rm exp}$ in time, when the statistics of Poincaré recurrences $P(t)$ obtained from the UPFO becomes exponential, increases as well according to $t_{\rm exp} \propto \gamma_0^{-1}$ and therefore we expect to recover the power law decay of $P(t)$ for $M \to \infty$ (see below in Sect. 3).

With the help of the Arnoldi method we find certain eigenstates corresponding to eigenvalues of the matrix S_p and satisfying the equation

$$\sum_{i=0}^{N_p-1} (S_p)_{mi}\psi_j(i) = \lambda_j\psi_j(m). \tag{2}$$

Examples of two eigenmodes $|\psi_0|$ and $|\psi_{29}|$ are shown in Figure 5. The state $|\psi_0|$ corresponds to the first diffusive mode mainly located in a vicinity of the critical golden curve while $|\psi_{29}|$ corresponds to the mode located near a resonant chain with rotation number $r = 2/7$.

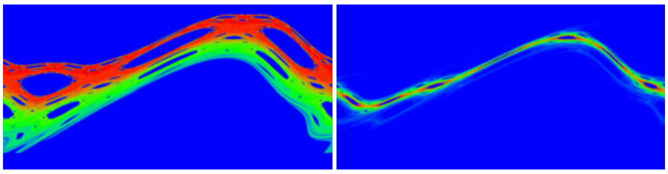

Fig. 5. Density plot of the modulus of the eigenvector components $|\psi|$ of the UPFO projected case of map (1) at K_g with $M = 1600$ for the two modes with eigenvalues $\lambda_0 = 0.99994672$ (left panel) and $\lambda_{29} = -0.22008951 + i\,0.96448508 \approx |\lambda_{29}|\,e^{i2\pi(2/7)}$ (right panel). The density is shown by color with red/gray for maximum and blue/black for zero.

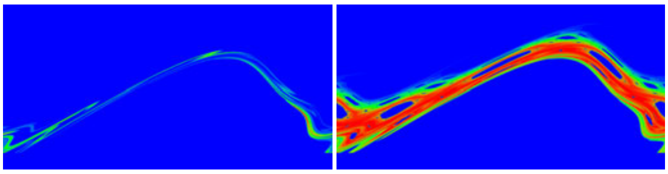

Fig. 6. Time dependent probability density calculated by $\psi(t) = (S_p)^t\,\psi(0)/\parallel (S_p)^t\,\psi(0)\parallel_1$ where S_p is the UPFO for the projected case for $M = 1600$, $\psi(0)$ an initial vector with $\psi_l(0) = \delta_{l,\ell_0}$ and ℓ_0 being the index of the cell at $x_0 = y_0 = 0.0625$ and $\parallel \ldots \parallel_1$ is the 1-norm defined by $\parallel \psi \parallel_1 = \sum_l |\psi_l|$. The densities are shown for $t = 40$ (left panel) and $t = 400$ (right panel). In the limit $t \to \infty$ the vector $\psi(t)$ converges to the eigenvector of maximal eigenvalue λ_0 shown in the left panel of Figure 5. The full convergence is achieved for $t \geq 40\,000$ so that for these times the density plot of $\psi(t)$ remains unchanged at the given color-resolution.

It is also interesting to follow how the probability initially placed in one cell ℓ_0 evolves with time. Of course, the total probability starts to decay due to absorption but by renormalizing the total probability back to unity after each map iteration we obtain its evolution in phase space. At large times we have convergence to the state ψ_0 with maximal λ_0 but at intermediate times we see the regions of phase space which contribute to long time sticking and long Poincaré recurrences. Two snapshots are shown in Figure 6. The videos of such an evolution for the maps (1) and (3) are available at [35].

3 Poincaré recurrences by survival Monte Carlo method

The numerical computation of the Poincaré recurrences counting the number of crossing of a given line (e.g. $y = 0$) in the phase space is known to be a very stable numerical method since the integrated probability of recurrences on a line at times larger than t is positively defined (see e.g. [18,19,24,28]). However, at large times the direct numerical computation becomes time consuming.

With the aim to reach larger times we present here a new method to calculate the statistics of Poincaré recurrences of map such as the Chirikov standard map (1). We will call this method the *Survival Monte Carlo method* (SMCM). The idea of this method is to chose a certain, quite large number $N_i \gg 1$, of initial conditions randomly chosen in some small cell close to an unstable fix point and to calculate in parallel the time evolution of these trajectories. At the initial time $t = 0$ we put the Poincaré return probability to $P(0) = 1$ and the number of trajectories to $N(0) = N_i$. At each time t_k, when a given trajectory escapes in the absorption region $y < y_{cut} = 0.05$ of the phase space, we put $P(t_k + 1) = P(t_k)\,(N(t_k) - 1)/N(t_k)$ and $N(t_k + 1) = N(t_k) - 1$, otherwise we simply keep $P(t_k + 1) = P(t_k)$ and $N(t_k + 1) = N(t_k)$. When the number of remaining trajectories $N(t_k)$ drops below a certain threshold value N_f (typically chosen such that $N_i \gg N_f \gg 1$) we *reinject* a new trajectory close to one of the other remaining trajectories with a small random deviation: $x_{\text{new}}(t) \in [x_i(t) - \varepsilon/2,\, x_i(t) + \varepsilon/2]$ and $y_{\text{new}}(t) \in [y_i(t) - \varepsilon/2,\, y_i(t) + \varepsilon/2]$. The main idea is to keep a typical statistics of trajectories at a given time t and to concentrate the computational effort on the very long and rare trajectories without wasting resources on the more probable trajectories with short times of Poincaré recurrences.

In this method the proper choice of ε is important. On one hand ε should not be too small in order to avoid too strong correlations between the trajectories and on the other hand it should be very small in order to avoid an uncontrolled too strong diffusion into regions too close to stable islands where the trajectories may be trapped stronger and longer as they should be without the random deviations. Fortunately in the chaotic region even a modest Lyapunov exponent ensures exponential separation of trajectories and choosing a very small value of ε one may hope to reduce the correlation between the injected trajectory and its reference trajectory after a modest number of iterations. Furthermore at longer times the average time between the escape of two trajectories becomes very large that helps to reduce these correlations.

We have chosen the parameters $\varepsilon = 10^{-14}$, $N_i = 10^6$ and the two cases $N_f = 100$ and $N_f = 1000$. For $N_f = 100$ we have been able to iterate up to times 10^{11} and for $N_f = 1000$ up to times 10^{10}. We mention that at the values $\varepsilon = 10^{-10}, 10^{-14}$, we observe sticking of certain trajectories for very long times while other trajectories escape more rapidly (see Fig. 7). For $N_f = 10$ these fluctuations become enormously large. Examples of the survival probability $P(t)$ obtained for 10 different realizations with $N_f = 100$ (left panel) and $N_f = 1000$ (right panel) are shown in Figure 7. Of course the fluctuations appear for $N_f = 100$ at shorter times ($t \sim 10^5$–10^6) as compared to $N_f = 1000$ ($t \sim 10^6$–10^7). For $N_f = 10$ the fluctuations appear even at shorter times.

We calculate in parallel different realizations of $P(t)$ with respect of the random variables (for the initial conditions, for the random deviations of the reinjected trajectories and for the random choice at which remaining trajectory the reinjection happens). The comparison of obtained data shows that the distribution $P(t)$ is stable at small and large times. But at very large times it turns out that the fluctuations become quite strong.

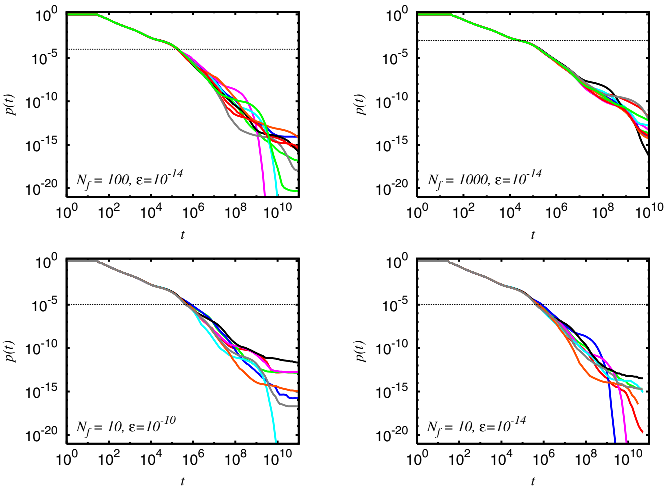

Fig. 7. Statistics of Poincaré recurrences $P(t)$ of the map (1) calculated by the SMCM as survival probability after times larger than t (data are shown in double logarithmic scale). Top panels: the number of initial trajectories is $N_i = 10^6$ and the number of final trajectories is $N_f = 100$ (left panel) or $N_f = 1000$ (right panel). The initial positions are randomly chosen in a cell of size $(1600)^{-1} \times (1600)^{-1}$ at the position $x_0 = y_0 = 0.0625$, here the small random deviation for reinjected trajectories is $\sim \varepsilon = 10^{-14}$. In both panels the results for $P(t)$ are shown for 10 realizations with different random seeds. The horizontal dotted line indicates the limit probability $N_f/N_i = 10^{-4}$ (left panel) or $N_f/N_i = 10^{-3}$ (right panel) below which the reinjection of trajectories is applied. The two realizations in the left panel which drop below the shown range (of $P(t) \geq 10^{-21}$) "saturate" eventually at the values $P(10^{11}) \approx 2 \times 10^{-36}$ or $P(10^{11}) \approx 10^{-35}$. Bottom panels: same as in top panels but with $N_f = 10$ at $\varepsilon = 10^{-10}$ (left panel) and $\varepsilon = 10^{-14}$ (right panel).

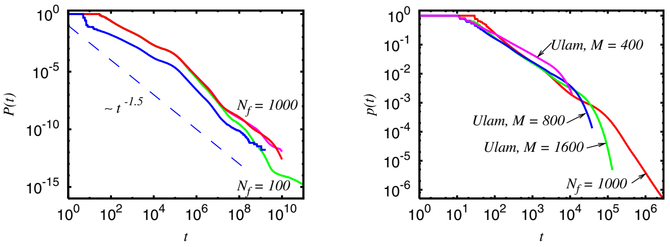

Fig. 8. The left panel shows the average over 10 random realizations of the statistics of Poincaré recurrences $P(t)$ of the map (1), obtained by the SMCM for survival probability and shown in Figure 7 at $N_f = 1000$, $\varepsilon = 10^{-14}$ (red curve); for comparison we also show data obtained by the same type of averaging but at $\varepsilon = 10^{-10}$ (magenta curve having a strong overlap with the red curve). The next lower/green curve, for $t \leq 10^{11}$, corresponds to $N_f = 100$, $\varepsilon = 10^{-14}$. The lowest blue curve, for $t \leq 1.7 \times 10^9$, corresponds to the data of references [25,26] obtained by a direct computation of the statistics of Poincaré recurrences. The dashed straight line indicates a power law behavior $P \propto t^{-1.5}$. The right panel compares the statistics of Poincaré recurrences $P(t)$, obtained by the SMCM for $N_f = 1000, \varepsilon = 10^{-14}$, to $P(t)$ obtained by the Ulam method for $M = 400$, 800, 1600. At large times $t > t_{\text{exp}} \sim 10^4$–$10^5$ the curves obtained by the Ulam method show an exponential behavior $P(t) \sim \lambda_0^t$ determined by the largest eigenvalue of the UPFO for the projected case.

We note that the SMCM allows us to determine the survival probability $P(t)$. Its comparison with the statistics of Poincaré recurrences computed by the usual method [18,24–26] is shown in Figure 8. We see that both methods give the same behavior $P(t)$ with a small shift in time related to different initial conditions. The equivalence of both methods is rather clear: in both methods the probability is determined by long sticking trajectories; both methods consider the recurrences to the lines $y = 0$ or $y = 0.05$ which are close to each other.

The decay of $P(t)$ averaged over 10 random realizations is shown in Figure 8. In general we see that the SMCM allows to reach extremely long times with $t = 10^{11}$ for $N_f = 100$ and $t = 10^{10}$ for $N_f = 1000$. For $N_f = 100$ we see that the fluctuations start to be important for $t > 10^9$ while the case with $N_f = 1000$ remains stable up to $t = 10^{10}$. This allows us to obtain the behavior of $P(t)$ for times being about one order of magnitude larger compared to previous numerical simulations.

We argue that these fluctuations appear not due to different values of $\varepsilon = 10^{-10}, 10^{-14}$ but due to enormous "spin glass" like fluctuations due to sticking in different regions of chaotic phase space. Indeed, according to the arguments presented in reference [24] at a recurrence time t a trajectory reaches a chaotic layer measure at a resonance q being $\mu_q \sim tP(t) \sim 1/q^2$. According to the data of Figure 8 at $N_f = 10^3$ and $t = 10^{10}$ with $P(t) \sim 10^{-13}$ we have $\mu_q \sim 10^{-3}$ and $q \sim 30$. Thus this chaos measure is very large compared to the displacement amplitude $\mu_q \sim 10^{-3} \gg \varepsilon \geq 10^{-10}$. Thus, these displacements generally should not move trajectories from chaotic to integrable components. In fact the strong fluctuations of various groups of orbits at $N_f = 10$ originate from sticking of orbits for very different time scales in various parts of phase space. At large values of $N_f = 1000$ the statistical averaging reduces these fluctuations but at larger times at fixed N_f the fluctuations become more and more pronounced. Our direct comparison of $P(t)$ for $N_f = 1000$ at $\varepsilon = 10^{-14}$ and $\varepsilon = 10^{-10}$ (see Fig. 8) show that the fluctuations remain small up to $t = 10^{10}$. For $N_f = 10$ this time is reduced down to $t \sim 10^8$. This comparison of data at two values of ε confirms that the chosen values of ε do not affect the averaged values of $P(t)$ on time scales considered in Figures 7, 8. We also note that the curve of Poincaré recurrences decay $P(t)$, computed in a standard way as in references [24,28], as well as $P(t)$ computation described here, is not affected by a change of the computational precision from a single to a double one (up to statistical fluctuations at the tail of $P(t)$). This is related to the above argument that μ_q measure is rather large at the times reached in numerical simulations.

According to the empirical data in Figure 8 at right panel for the Ulam method we see that the time t_{cel}, during which the computations of $P(t)$ with a finite size cell of size $\varepsilon_{cel} = 1/M$ are correct, scales approximately as $t_{cel} \sim 10/\varepsilon_{cel}$. In a similar way we find that for $\varepsilon = 10^{-6}$ and single precision computations the curve $P(t)$ obtained at $N_f = 1000$, $\varepsilon = 10^{-14}$ is reproduced up to a time $t_{cel} \sim 10/\varepsilon$. It may be interesting to analyze the dependence of t_{cel} on ε in more detail but we leave this for further studies.

For the case $N_f = 1000$, $\varepsilon = 10^{-14}$ in Figure 8 the algebraic fit of data in the range $10^6 \leq t \leq 10^{10}$ gives the Poincaré exponent $\beta = 1.587 \pm 0.009$. For $N_f = 100$ case we find $\beta = 1.710 \pm 0.017$ for the range $10^6 \leq t \leq 10^{11}$. The formal statistical error is rather small in both cases but it is clear that for $N_f = 100$ we start to have an effect of strong fluctuations due to long sticking around islands and thus the reliable value of β is given by the case with $N_f = 1000$.

The survival probability $P(t)$ can be also computed using the Ulam method at various sizes of discrete cells determined by M. The results obtained by the generalized Ulam method and by the SMCM are shown in the right panel of Figure 8. The comparison shows that both methods give the same results but the SMCM is much more efficient allowing to follow the decay $P(t)$ up to significantly larger times since for the Ulam method we expect the decay $P(t)$ only to be accurate for $t < t_{\rm exp} \sim \gamma_0^{-1}$ because for $t > t_{\rm exp}$ it becomes exponential $P(t) \propto \lambda_0^t = \exp(-\gamma_0\, t/2)$. The data of Figure 8 clearly shows that $t_{\rm exp}$ increases with M in accordance with the decay of γ_0 obtained from Figure 4.

Using the SMCM we can follow the evolution of the survival probability as a function of time showing the density plot of long sticking trajectories. Examples of such distributions are shown in Figure 9. These figures show that at short times $t < 100$ the trajectories are not yet able to cross the cantori barriers and remain relatively far from the golden curve, at larger times $t = 10^4, 10^6, 10^8$ the probability becomes concentrated close to the golden curve. But at very larger times $t = 10^{10}$ we find trajectories sticking in a vicinity of the golden curve or other secondary resonances. Thus we see that at long time $P(t)$ has contributions not only from the vicinity of the critical golden curve but also from other secondary resonances. In this respect, our conclusion confirms a similar one expressed in reference [27] obtained from simulations on shorter time scales.

4 Separatrix map with critical golden curve

To show that the previous case of the Chirikov standard map represents a generic situation we also study the UPFO of the projected case for the separatrix map [6], defined by:

$$\bar{y} = y + \sin(2\pi x), \bar{x} = x + \frac{\Lambda}{2\pi}\ln(|\bar{y}|) \pmod 1 \,. \qquad (3)$$

This map can be locally approximated by the Chirikov standard map by linearizing the logarithm near a certain y_0 that leads after rescaling to the map (1) with an effective parameter $K_{\rm eff} = \Lambda/|y_0|$ [6]. As in reference [34] we study the map (3) at $\Lambda_c = 3.1819316$ with the critical golden curve at the rotation number $r = r_g = (\sqrt{5}-1)/2 = 0.618\ldots$ The construction of the matrix S is described in reference [34], its size is given by an approximate relation $N_d \approx 0.78M^2/2$ for the phase space region $0 < x \leq 1$, $0 \leq y \leq 4$ (symplectic case and using the symmetry: $x \to x + 1/2 \pmod 1$, $y \to -y$). The absorption is done for $y < y_{cut} = 0.4$ corresponding to 10% of the maximal possible value of y. Thus for the UPFO for the projected case we have $N_p \approx 0.68M^2/2$. In fact we have $2(N_d - N_p)/M^2 = 0.1$ since all part of the phase space is chaotic at $0 < y < y_{cut}$ and all cells in this region were occupied by the Ulam method. Thus for $M = 1600$ we have $N_d = 997\,045$, $N_p = 869\,045$.

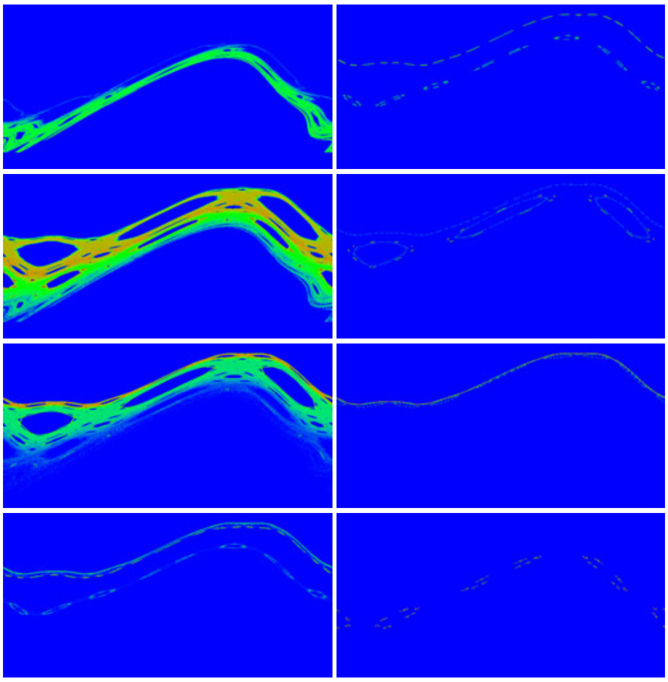

Fig. 9. Density plots of the trajectories of the SMCM (with $N_f = 1000$) for the map (1) for various times t and random realizations. All density plots are obtained from a histogram of 10^7 data points and using a resolution of 800×400 cells for the phase space $0 \leq x < 1$ and $0 \leq y < 0.5$. The data points are obtained by iterating $N(t)$ trajectories (with $N(t) = P(t)\, N_i$ for $P(t) \geq 10^{-3}$ and $N(t) = N_f$ for $P(t) < 10^{-3}$) from t to $t + \Delta t$ with $\Delta t = 10^7/N(t)$. The left four panels and the upper right panel correspond to one particular random realization at $t = 10^2$, 10^4, 10^6, 10^8, 10^{10} and the three lower right panels correspond to three other random realizations at $t = 10^{10}$. For short times $t < 10^5$ there is no significant difference between the density plots for different random realizations at a given time. More detailed density plots for intermediate times and higher resolution figures are available at reference [35].

In Figure 10, in analogy to Figure 4, we show the dependence of integrated number of eigenvalues j/N_d on $\gamma_j = -2\ln|\lambda_j|$ for the symplectic and projected cases of the UPFO of the map (3). In both cases we have approximately the same dependence with the algebraic exponent $\beta \approx 1.5$ which works for the range $0.04 \leq \gamma \leq 0.3$. The minimal values of γ (e.g. γ_0 and γ_2) drop approximately inversely proportionally to M. As for symplectic case [34] we attribute this decrease with M to a finite size

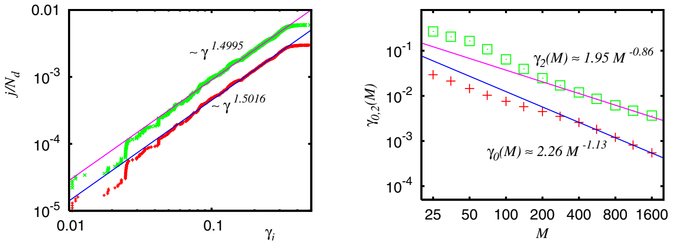

Fig. 10. The left panel shows the rescaled level number j/N_d versus the decay rate γ_j, in a double logarithmic scale, for the separatrix map (3) at Λ_c with $M = 1600$ and $N_d = 997\,045$. Red/lower data points correspond to the UPFO for the projected case and green/upper data points correspond to the symplectic case. For the symplectic case the data points are shifted up by a factor 2 to separate the two data sets. The two straight lines show the power law fits $j/N_d \approx 0.014173\,\gamma^{1.4995}$ (symplectic case) and $j/N_d \approx 0.014207\,\gamma^{1.5016}$ (projected case) for the range $0.04 \leq \gamma \leq 0.3$. The statistical error of the exponents is close to 0.2% in both cases. The right panel shows the decay of $\gamma_j(M)$ with M for $j = 0$ (red crosses), $j = 2$ (green open squares) for the UPFO for the projected case of map (3). The lower/blue straight line corresponds to the power law fit $\gamma_0(M) \approx 2.26M^{-1.13}$ and the upper/pink straight line to the fit $\gamma_2(M) \approx 1,95M^{-0.86}$ (for the range $400 \leq M \leq 1600$). The eigenvector corresponding to γ_2 is localized near the two stable islands of the resonance 1/2.

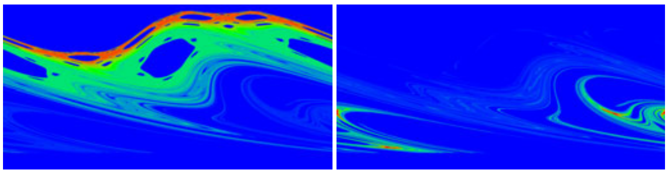

Fig. 11. Density plot of the modulus of the eigenvector components of the UPFO for the projected case of the map (3) at $M = 1600$ for the two modes with $\lambda_0 = 0.99972660$ (left panel) and $\lambda_{77} = -0.49158775 + i\,0.85153885 \approx |\lambda_{77}|\,e^{i\,2\pi(1/3)}$ (right panel).

coarse-graining effect of the Ulam method. As in reference [34], we argue that the exponent β for a more physical intermediate range of γ is directly related to the Poincaré exponent.

Examples of two eigenmodes at λ_0 and λ_{77} are shown in Figure 11. In the first case we have an eigenmode of diffusive type similar to Figure 5 while in the latter case we have an eigenmode concentrated around unstable fix points of resonance 1/3 (see corresponding state of symplectic case in bottom left panel of Fig. 11 in Ref. [34]).

The comparison of the statistics of Poincaré recurrences obtained from the map (3) by the SMCM and the usual method are shown in Figure 12. The data of the usual method obtained in reference [25] allows us to follow the decay of $P(t)$ up to $t = 2 \times 10^8$, while with the SMCM we reach times $t = 10^{10}$ with $N_f = 1000$ and $t = 10^{11}$ with $N_f = 100$. We have a good agreement between three curves for the range $100 \leq t \leq 10^8$ with a certain constant displacement in $\log_{10} t$ of data from the usual method compared to the SMCM data. This shift appears due to different initial conditions but apart of this shift all oscillations of $P(t)$ curve are well reproduced. This shows that both methods works correctly. However, with the SMCM we are able to reach times being by one to two orders of magnitude larger than previously.

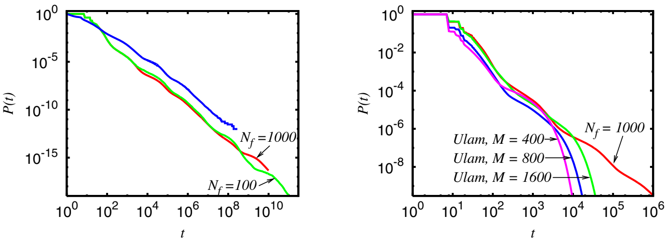

Fig. 12. The left panel shows the average over 10 random realizations of the statistics of Poincaré recurrences $P(t)$ of the map (3), obtained by the SMCM. The red curve, for $t \leq 10^{10}$, corresponds to $N_f = 1000$. The green curve, for $t \leq 10^{11}$, corresponds to $N_f = 100$. The upper/blue curve, for $t \leq 2.8 \times 10^8$, corresponds to the data shown in reference [25] using a direct computation of the statistics of Poincaré recurrences. The right panel compares $P(t)$ SMCM data for $N_f = 1000$ (red curve in left and right panels) with $P(t)$ obtained by the Ulam method for $M = 400, 800, 1600$. At large times $t > t_{\exp} \sim 2 \times 10^3 – 2 \times 10^4$ the Ulam method leads to an exponential decay $P(t) \sim \lambda_0^t$ determined by the largest eigenvalue of the UPFO for the projected case.

The algebraic fit of SMCM data in Figure 12 gives $\beta = 1.855 \pm 0.004$ for $N_f = 100$ (range $10^4 \leq t \leq 10^{11}$) and $\beta = 1.706 \pm 0.004$ for $N_f = 1000$ (range $10^4 \leq t \leq 10^{10}$). In both cases the statistical error is rather small but there are visible fluctuations which become to be significant at $t > 10^9$ for $N_f = 100$ even if they are smaller compared to the similar case of map (1) shown in Figure 8. Due to that one should take as the reliable value $\beta = 1.706$ that shows a noticeable difference from the value $\beta = 1.587$ found above for the Chirikov standard map at $K = K_g$.

The comparison of the SMCM data for $P(t)$ with the results of the Ulam method are shown in the right panel of Figure 12. As it was the case for the similar comparison shown in Figure 8 we find that both methods give the same results but the Ulam method works only for time scales being significantly smaller than those reached with the SMCM.

Finally, as in Figure 9, we show in Figure 13 the density distribution obtained for various realizations and various times of the map (3). The situation is similar to Figure 9: at short times the density is bounded by cantori barriers, at large times it reaches the critical golden curve and at even larger times we see that the density is located near the critical golden curve or other secondary resonances depending on the realization.

5 Properties of eigenstates of Ulam matrix

Let us now try to analyze how the decay of Poincaré recurrences is related to the properties of the (right) eigenvectors $\psi(x, y)$ of the UPFO for the projected

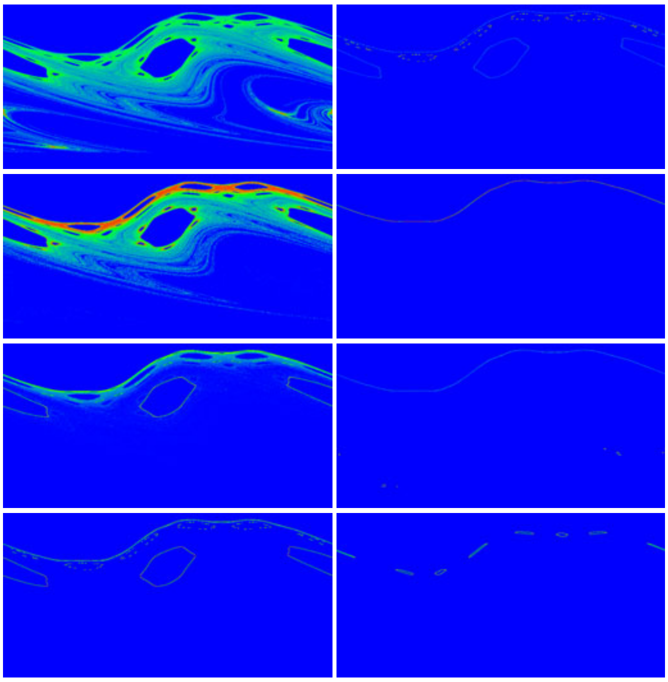

Fig. 13. Density plots of the trajectories of the SMCM with $N_f = 1000$ for the map (3) for various times t and various realizations. All density plots are obtained by a histogram of 10^7 data points with a resolution of 800×400 cells for the phase space $0 \le x < 1$ and $0 \le y < 4$. The data points are obtained by iterating the $N(t)$ trajectories (with $N(t) = P(t)\, N_i$ for $P(t) \ge 10^{-3}$ and $N(t) = N_f$ for $P(t) < 10^{-3}$) from t to $t + \Delta t$ with $\Delta t = 10^7/N(t)$. The left four panels and the upper right panel correspond to one particular random realization at $t = 10^2$, 10^4, 10^6, 10^8, 10^{10} and the three lower right panels correspond to three other random realizations at $t = 10^{10}$. For short times $t < 10^5$ there is no significant difference between the density plots for different random realizations at a given time.

case. For this we determine the x-average of the eigenvector amplitude around a given position x_0 over a band of 1% width of the whole x-range: $\langle|\psi(y)|\rangle = 100\ M^{-1} \sum_{|\Delta x|<0.005} |\psi(x_0 + \Delta x, y)|$. The y-dependence of this average allows to visualize the localization properties of the eigenstate in y-direction. In Figure 14 we show this quantity for two examples for each of the maps (1) and (3) and for different values of M between 400 and 1600.

For the case of the map (1), shown in the left column of Figure 14, we see a clear evidence of exponential localization of eigenstates. In fact the average amplitude in a vicinity of $y \approx 0$, where the initial state is taken and where the absorption happens, has enormously small values being of the order of 10^{-15}. These amplitudes on the tail drop significantly with an increase of M. For the map (3) the decay of eigenstates is more irregular since the band at $x \approx x_0$ crosses some secondary islands thus leading to appearance of a plateau in the decay with y. But in global we can still say that there is an exponential decay of eigenstates. This exponential localization of eigenstates reminds the Anderson localization in disordered solid state systems (see e.g. [48]).

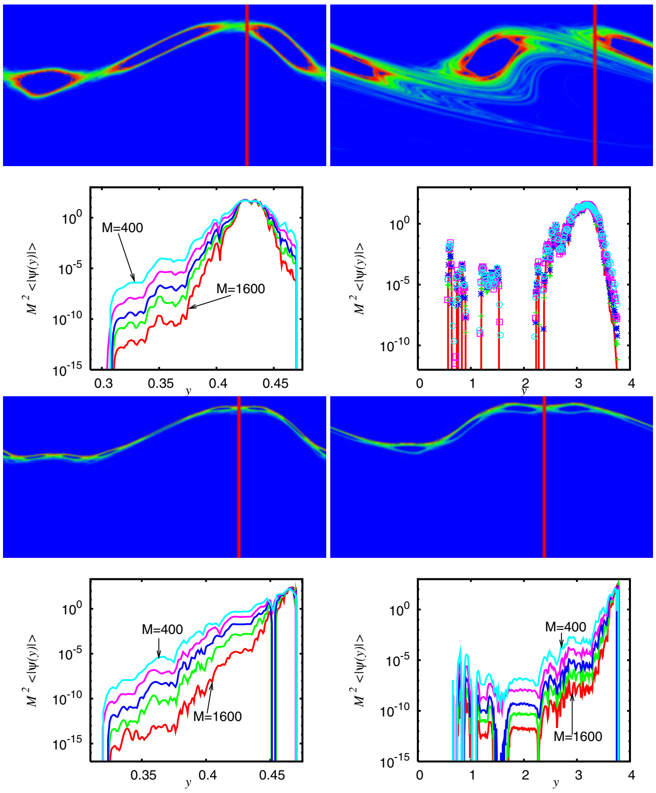

Fig. 14. The localization properties in y-direction for certain eigenvectors of the UPFO for the projected case for the maps (1) (left column) and (3) (right column). The panels in the second and fourth row show the averaged modulus $\langle|\psi(y)|\rangle$ of the eigenvector components within a band of 1% width of the whole x-range at a certain $x = x_0$. The global structure of the corresponding eigenstates is shown in the corresponding first and third panels (counting from the top; the red vertical thick line indicates the range of x-values where the average has been performed for each y-value, $M = 1600$). Data are shown for $M = 400$ (cyan/highest curve), $M = 560$ (pink/second curve), $M = 800$ (blue/third curve), $M = 1120$ (green/fourth curve) and $M = 1600$ (red/lowest curve). In the right panel of the second row the data for different values of M approximately coincide and only the data for $M = 1600$ are shown by a full (red) curve; other M values are shown as isolated data points for $M = 1120$ (green crosses), $M = 800$ (blue stars), $M = 560$ (pink squares) and $M = 400$ (cyan circles). For $M = 1600$ the eigenvectors, shown in the density plots of the first and third row, correspond to the modes λ_4 and λ_{31} of the map (1) (left column) and to the modes λ_2 and λ_{17} of the map (3) (right column); for other M we show corresponding eigenvector located at the same resonances.

We can also consider the projection of our initial state taken in a cell ℓ_0 on the eigenstates. Indeed, this initial state can be expressed as $\psi_{\rm init} = \sum_j \mu_j\, \psi_j^R$ where μ_j are expansion amplitudes and ψ_j^R the right eigenvectors defined by equation (2). To determine the values of μ_j we need first to compute the left eigenvectors ψ_j^L of the Ulam matrix S_p which are biorthogonal to the right eigenvectors

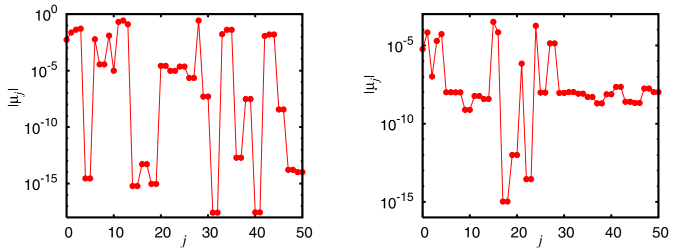

Fig. 15. Modulus of the projection coefficients μ_j of the initial density vector $\psi_{\rm init}$, localized in one cell at $x_0 = y_0 = 0.0625$, with respect to the right eigenvectors ψ_j^R (of the UPFO projected case for $M = 1600$) versus level number j. These coefficients appear in the expansion $\psi_{\rm init} = \sum_j \mu_j \, \psi_j^R$ (see text). The left and right panels represent data for the maps (1) and (3) respectively. The cases with $|\mu_j| = |\mu_{j+1}|$ correspond to pairs of complex conjugated modes with $\mu_{j+1} = \mu_j^*$.

ψ_j^R and provide the expansion amplitudes by the identity: $\mu_j = \langle \psi_j^L | \psi_{\rm init} \rangle / \langle \psi_j^L | \psi_j^R \rangle$. Note that this expression does not depend on the chosen normalization of the eigenvectors and it requires only that $\langle \psi_j^L | \psi_j^R \rangle \neq 0$. However, for convenience, we have normalized both type of eigenvectors by the L_1-norm such that $\sum_{x,y} |\psi_j^{R,L}(x,y)| = 1$. We have numerically determined the first 51 left eigenvectors with the help of the Arnoldi method applied to the transpose of S_p and therefore obtained the corresponding expansion amplitudes.

The dependence of μ_j on j is shown in Figure 15. We see that there are enormously large fluctuations of μ_j which are in a range of 10 orders of magnitude. In particular the amplitudes corresponding to resonant modes are very small which is easy to understand if the resonant mode is localized far away from the initial state and does therefore not contribute to the expansion. We think that these fluctuations are at the origin of the slow algebraic decay of Poincaré recurrences $P(t)$ (see below).

In Figure 16 we show the contribution of the largest N_m eigenmodes to the statistics of Poincaré recurrences (for $M = 1600$) given by the formula: $p(t) = \sum_{j=0}^{N_m-1} p_j \, \lambda_j^t$ with $p_j = \mu_j \sum_{x,y} \psi_j^R(x,y)$ and the eigenvalues ordered as $|\lambda_0| > |\lambda_1| > |\lambda_2| > \ldots$

For $N_m = N_p$, we have the statistics of Poincaré recurrences obtained from the iteration of the UPFO and already shown in Figures 8 and 12. For $N_m = 51$ we have evaluated the sum using the expansion coefficients shown in Figure 15. Both curves coincide at $t > 10^2$ for the map (1) or at $t > 3 \times 10^2$ for the map (3) showing that the largest eigenmodes determine the long time behavior. For large times ($t > 10^4 – 10^5$) only the first eigenmode contributes and the decay is purely exponential. It turns out that in the sum for $N_m = 51$ the terms arising from the resonant modes can be omitted without changing the curve up to graphical precision since these modes contribute only very weakly in the expansion. In general, the partial sum $p(t)$ converges to the actual statistics of Poincaré recurrences $P(t)$ with increasing N_m and at given value of N_m one expects that $p(t)$ and $P(t)$ coincide for $t \gg 2\,\gamma_{N_m}^{-1}$.

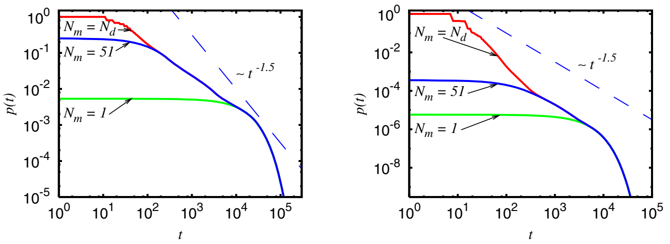

Fig. 16. Contributions of the largest eigenmodes of the UPFO projected case at $M = 1600$ to the statistics of Poincaré recurrences for the maps (1) (left panel) and (3) (right panel). Here, we show the probability $p(t)$ obtained from the expansion over eigenvectors given by the formula $p(t) = \sum_{j=0}^{N_m-1} p_j \, \lambda_j^t$ with $p_j = \mu_j \sum_{x,y} \psi_j^R(x,y)$, N_m being the number of used modes and the eigenvalues being ordered as $|\lambda_0| > |\lambda_1| > |\lambda_2| > \ldots$ (see text). The upper red curve is obtained from the direct iteration of the UPFO (see green curve in the right panels of Figs. 8 and 12) and corresponds to the contribution of the full spectrum of all eigenvalues with $N_m = N_p$. The middle blue curve corresponds to $N_m = 51$ with the same μ_j values as those shown in Figure 15. The main contributions to this curve arise from the diffusion modes (with real positive eigenvalues $\lambda_j > 0$), the other resonant modes with complex or real negative eigenvalues give only a small contribution which does not modify the curve up to graphical precision. The bottom green curve corresponds to $N_m = 1$, i.e. the contribution $\mu_0 \, \lambda_0^t$ of the largest λ eigenmode. In both panels the dashed line indicates for comparison a power law decay $P(t) \propto t^{-1.5}$.

The data of Figures 14–16 illustrate the nontrivial link between the localized eigenstates of the Ulam matrix and the decay of Poincaré recurrences. The eigenmodes are exponentially localized and for many of them their projection on the initial state is very small but at some large times their contribution can become very important since the modes with large projections decay more rapidly.

6 Discussion

Our studies show that the generalized Ulam method reproduces well the decay of Poincaré recurrences $P(t)$ in 2D symplectic maps with divided phase space. At the same time the computation of $P(t)$ is obtained in a more efficient way by the proposed SMCM allowing to reach time scales of the order of $t = 10^{10}$. We find that at these large times the Poincaré exponent has values $\beta = 1.58$ for the Chirikov standard map at K_g and $\beta = 1.70$ for the separatrix map at Λ_c. The recurrences at large times are dominated by sticking of trajectories not only in a vicinity of the critical golden curve but also in a vicinity of secondary resonance structures. This confirms earlier numerical observations obtained on shorter time scales [27].

The sticking around various different resonant structures on smaller and smaller scales of phase space leads to nontrivial oscillations of the Poincaré exponent. The values of β found here are not so far from the average values found previously by averaging over maps at different parameters with $\beta \approx 1.5$ [18,19], $\beta \approx 1.57$ [28]. In agreement with the data presented here and in reference [34], we find that the above value of β is close to the exponent of integrated density of states of the Ulam matrix which has $\beta \approx 1.5$. At the same time we see that at $t = 10^{10}$ the fluctuations in the Chirikov standard map at various N_f and various random realizations are significantly stronger as compared to the separatrix map.

We attribute these fluctuations to a localization of eigenstates of the Ulam matrix which gives very nontrivial properties of eigenstates projection on an initial state. The properties of these eigenstates are still poorly understood. We think that the further developments of analytical models of renormalization on Cayley type tree [22,23,28,32,33] and their applications to the puzzle of statistics of Poincaré recurrences should develop a more detailed analysis of localization of eigenstates of the Ulam matrix.

This research 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. We also acknowledge the France-Armenia collaboration grant CNRS/SCS No. 24943 (IE-017) on "Classical and quantum chaos". We note that the high resolution figures and video of dynamics are available at the web page [35]. We dedicate this work to the memory of Boris Chirikov (06.06.1928–12.02.2008).

References

1. J. Loschmidt, *Über den Zustand des Wärmegleichgewichts eines Systems von Körpern mit Rcksicht auf die Schwerkraft* (Sitzungsberichte der Akademie der Wissenschaften, Wien, 1876), Vol. II 73, p. 128
2. L. Boltzmann, *Über die Beziehung eines allgemeine mechanischen Satzes zum zweiten Hauptsatze der Wärmetheorie* (Sitzungsberichte der Akademie der Wissenschaften, Wien, 1877), Vol. II 75, p. 67
3. V.I. Arnold, A. Avez, *Ergodic Problems of Classical Mechanics* (Benjamin, Paris, 1968)
4. I.P. Cornfeld, S.V. Fomin, Y.G. Sinai, *Ergodic Theory* (Springer, New York, 1982)
5. B.V. Chirikov, *Research Concerning the Theory of Nonlinear Resonance and stochasticity*, Preprint No. 267 (Institute of Nuclear Physics, Novosibirsk, 1969) (in Russian) [Engl. Transl., CERN Trans. 71–40, Geneva, October (1971)]
6. B.V. Chirikov, Phys. Rep. **52**, 263 (1979)
7. A.J. Lichtenberg, M.A. Lieberman, *Regular and Chaotic Dynamics* (Springer, Berlin, 1992)
8. B. Chirikov, D. Shepelyansky, Scholarpedia **3**, 3550 (2008)
9. J.M. Greene, J. Math. Phys. **20**, 1183 (1979)
10. R.S. MacKay, Physica D **7**, 283 (1983)
11. R.S. MacKay, I.C. Percival, Comm. Math. Phys. **94**, 469 (1985)
12. B.V. Chirikov, Critical Perturbation in Standard map: a better approximation `arXiv:nlin/0006021[nlin.CD]` (2000)
13. S. Aubry, Physica D **7**, 240 (1983)
14. R.S. MacKay, J.D. Meiss, I.C. Percival, Physica D **13**, 55 (1984)
15. J.M. Greene, R.S. MacKay, J. Stark, Physica D **21**, 267 (1986)
16. H. Poincaré, Acta Mathematica **13**, 1 (1890)
17. S.R. Chanon, J.L. Lebowitz, Ann. N.Y. Acad. Sci. **357**, 108 (1980)
18. B.V. Chirikov, D.L. Shepelyansky, Preprint 81-69 Inst. Nuclear Physics, Novosibirsk, 1981 [English translation, Princeton Univ. Report No. PPPL-TRANS-133, (1983)]
19. B.V. Chirikov, D.L. Shepelyansky, in *Proceedings IX Int. Conf. on Nonlinear Oscillations Kiev, 1981*, Naukova Dumka **2**, 420 (1984)
20. C.F.F. Karney, Physica D **8**, 360 (1983)
21. B.V. Chirikov, D.L. Shepelyansky, Physica D **13**, 395 (1984)
22. J. Meiss, E. Ott, Phys. Rev. Lett. **55**, 2741 (1985)
23. J. Meiss, E. Ott, Physica D **20**, 387 (1986)
24. B.V. Chirikov, D.L. Shepelyansky, Phys. Rev. Lett. **82**, 528 (1999)
25. B.V. Chirikov, D.L. Shepelyansky, Phys. Rev. Lett. **89**, 239402 (2002)
26. M. Weiss, L. Hufnagel, R. Ketzmerick, Phys. Rev. Lett. **89**, 239401 (2002)
27. M. Weiss, L. Hufnagel, R. Ketzmerick, Phys. Rev. E **67**, 046209 (2003)
28. G. Cristadoro, R. Ketzmerick, Phys. Rev. Lett. **100**, 184101 (2008)
29. R. Artuso, C. Manchein, Phys. Rev. E **80**, 036210 (2009)
30. R. Venegeroles, Phys. Rev. Lett. **102**, 064101 (2009)
31. I.I. Shevchenko, Phys. Rev. E **81**, 066216 (2010)
32. V.A. Avetisov, S.K. Nechaev, Phys. Rev. E **81**, 046211 (2010)
33. R. Ceder, O. Agam, Phys. Rev. E **87**, 012918 (2013)
34. K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **76**, 57 (2010)
35. *Quantware Library*, edited by K. Frahm, D.L. Shepelyansky, Section QNR16 at `http://www.quantware.ups-tlse.fr/QWLIB/ulammethod/`
36. S.M. Ulam, A Collection of Mathematical Problems, *Interscience tracs in pure and applied mathematics*, (Interscience, New York, 1960), Vol. 8, p. 73
37. M. Brin, G. Stuck, *Introduction to Dynamical Systems* (Cambridge University Press, Cambridge, 2002)
38. T.-Y. Li, J. Approx. Theory **17**, 177 (1976)
39. Z. Kovács, T. Tél, Phys. Rev. A **40**, 4641 (1989)
40. Z. Kaufmann, H. Lustfeld, J. Bene, Phys. Rev. E **53**, 1416 (1996)
41. G. Froyland, R. Murray, D. Terhesiu, Phys. Rev. E **76**, 036702 (2007)
42. G. Froyland, K. Padberg, Physica D **238**, 1507 (2009)
43. D.L. Shepelyansky, O.V. Zhirov, Phys. Rev. E **81**, 036213 (2010)
44. L. Ermann, D.L. Shepelyansky, Phys. Rev. E **81**, 036221 (2010)
45. L. Ermann, D.L. Shepelyansky, Eur. Phys. J. B **75**, 299 (2010)
46. D.L. Shepelyansky, Phys. Rev. E **77**, 015202(R) (2008)
47. G.W. Stewart, *Matrix Algorithms: Eigensystems* (SIAM, 2001), Vol. II
48. F. Evers, A.D. Mirlin, Rev. Mod. Phys. **80**, 1355 (2008)

Google matrix of the citation network of Physical Review

Klaus M. Frahm,[1] Young-Ho Eom,[1] and Dima L. Shepelyansky[1]

[1] *Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France*

(Dated: October 21, 2013)

We study the statistical properties of spectrum and eigenstates of the Google matrix of the citation network of Physical Review for the period 1893 - 2009. The main fraction of complex eigenvalues with largest modulus is determined numerically by different methods based on high precision computations with up to $p = 16384$ binary digits that allows to resolve hard numerical problems for small eigenvalues. The nearly nilpotent matrix structure allows to obtain a semi-analytical computation of eigenvalues. We find that the spectrum is characterized by the fractal Weyl law with a fractal dimension $d_f \approx 1$. It is found that the majority of eigenvectors are located in a localized phase. The statistical distribution of articles in the PageRank-CheiRank plane is established providing a better understanding of information flows on the network. The concept of ImpactRank is proposed to determine an influence domain of a given article. We also discuss the properties of random matrix models of Perron-Frobenius operators.

PACS numbers: 89.75.Hc, 89.20.Hh, 89.75.Fb

I. INTRODUCTION

The development of Internet led to emergence of various types of complex directed networks created by modern society. The size of such networks grows rapidly going beyond ten billions in last two decades for the World Wide Web (WWW). Thus the development of mathematical tools for the statistical analysis of such networks becomes of primary importance. In 1998, Brin and Page proposed the analysis of WWW on the basis of PageRank vector of the associated Google matrix constructed for a directed network [1]. The mathematical foundations of this analysis are based on Markov chains [2] and Perron-Frobenius operators [3]. The PageRank algorithm allows to compute the ranking of network nodes and is known to be at the heart of modern search engines [4]. However, in many respects the statement of Brin and Page that *"Despite the importance of large-scale search engines on the web, very little academic research has been done on them"* [1] still remains valid at present. In our opinion, this is related to the fact that the Google matrix G belongs to a new class of operators which had been rarely studied in physical systems. Indeed, the physical systems are usually described by Hermitian or unitary matrices for which the Random Matrix Theory [5] captures many universal properties. In contrast, the Perron-Frobenium operators and Google matrix have eigenvalues distributed in the complex plane belonging to another class of operators.

The Google matrix is constructed from the adjacency matrix A_{ij} which has unit elements if there is a link pointing from node j to node i and zero otherwise. Then the matrix of Markov transitions is constructed by normalizing elements of each column to unity ($S_{ij} = A_{ij}/\sum_i A_{ij}$, $\sum_j S_{ij} = 1$) and replacing columns with only zero elements (*dangling nodes*) by $1/N$, with N being the matrix size. After that the Google matrix of the network takes the form [1, 4]:

$$G_{ij} = \alpha S_{ij} + (1-\alpha)/N \quad . \tag{1}$$

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 $\alpha \approx 0.85$ [4]. The PageRank vector P_i is the right eigenvector of G at $\lambda = 1$ ($\alpha < 1$). According to the Perron-Frobenius theorem [3], P_i components are positive and represent the probability to find a random surfer on a given node i (in the stationary limit) [4]. All nodes can be ordered in a decreasing order of probability $P(K_i)$ with highest probability at top values of PageRank index $K_i = 1, 2, ...$.

The distribution of eigenvalues of G can be rather nontrivial with appearance of the fractal Weyl law and other unusual properties (see e.g. [6, 7]). For example, a matrix G with random positive matrix elements, normalized to unity in each column, has $N-1$ eigenvalues λ concentrated in a small radius $|\lambda| < 1/\sqrt{3N}$ and one eigenvalue $\lambda = 1$ (see below in section VII). Such a distribution is drastically different from the eigenvalue distributions found for directed networks with algebraic distribution of links [8] or those found numerically for other directed networks including WWW of universities [9, 10], Linux Kernel and Twitter networks [11, 12], Wikipedia networks [13, 14]. In fact even the Albert-Barabási model of preferential attachment [16] still generates the complex spectrum of λ with a large gap ($|\lambda| < 1/2$) [8] being very different from the gapless and strongly degenerate G spectrum of WWW of British universities [10] and Wikipedia [13, 14]. Thus it is useful to get a deeper understanding of the spectral properties of directed networks and to develop more advanced models of complex networks which have a spectrum similar to such networks as British universities and Wikipedia.

With the aim to understand the spectral properties of Google matrix of directed networks we study here the Citation Network of Physical Review (CNPR) for the whole period up to 2009 [15]. This network has $N = 463348$ nodes (articles) and $N_\ell = 4691015$ links. Its network structure is very similar to the tree network since the citations are time ordered (with only a few exceptions of mutual citations of simultaneously published articles). As a result we succeed to develop powerful tools which allowed us to obtain the spectrum of G in semi-analytical way. These results are compared with the spectrum obtained numerically with the help of the powerful Arnoldi method (see its description in [17, 18]). Thus we are able to get a better understanding of the spectral properties of this network. Due to time ordering of article citations there are strong similarities between the CNPR and the network of integers studied recently in [19].

We note that the PageRank analysis of the CNPR had been performed in [20, 21],[22] showing its efficiency in determining the influential articles of Physical Review. The citation networks are rather generic (see e.g. [23]) and hence the extension of PageRank analysis of such networks is an interesting and important task. Here we put the main accent on the spectrum and eigenstates properties of the Google matrix of the CNPR but we also discuss the properties of two-dimensional (2D) ranking on PageRank-CheiRank plane developed recently in [24, 25],[26]. We also analyze the properties of ImpactRank which shows a domain of influence of a given article.

In addition to the whole CNPR we also consider the CNPR without Rev. Mod. Phys. articles which has $N = 460422$, $N_\ell = 4497707$. If in the whole CNPR we eliminate future citations (see description below) then this triangular CNPR has $N = 463348$, $N_\ell = 4684496$. Thus on average we have approximately 10 links per node. The network includes all articles of Physical Review from its foundation in 1893 till the end of 2009.

The paper is composed as follows: in Section II we present a detailed analysis of the Google matrix spectrum of CNPR, the fractal Weyl law is discussed in Section III, properties of eigenstates are discussed in Section IV, CheiRank versus PageRank distributions are considered in Section V, properties of impact propagation through the network are studied in Section VI, certain random matrix models of Google matrix are studied in Section VII, the discussion of the results is given in Section VIII.

II. EIGENVALUE SPECTRUM

The Google matrix of CNPR is constructed on the basis of Eq.(1) using citation links from one article to another (see also [22]). The matrix structure for different order representations of articles is shown in Fig. 1. In the top left panel all articles are ordered by time that generates almost perfect triangular structure corresponding to time ordering of citations. Still there are a few cases with joint citations of articles which appear almost at the same time. This breaks the triangular structure but the weight of such cases is small and we will see that with a good approximation one can neglect such links in a first approximation. The triangular matrix structure is also well visible in the middle left panel where articles are time ordered within each Phys. Rev. journal. The left bottom panel shows the matrix elements for each Phys

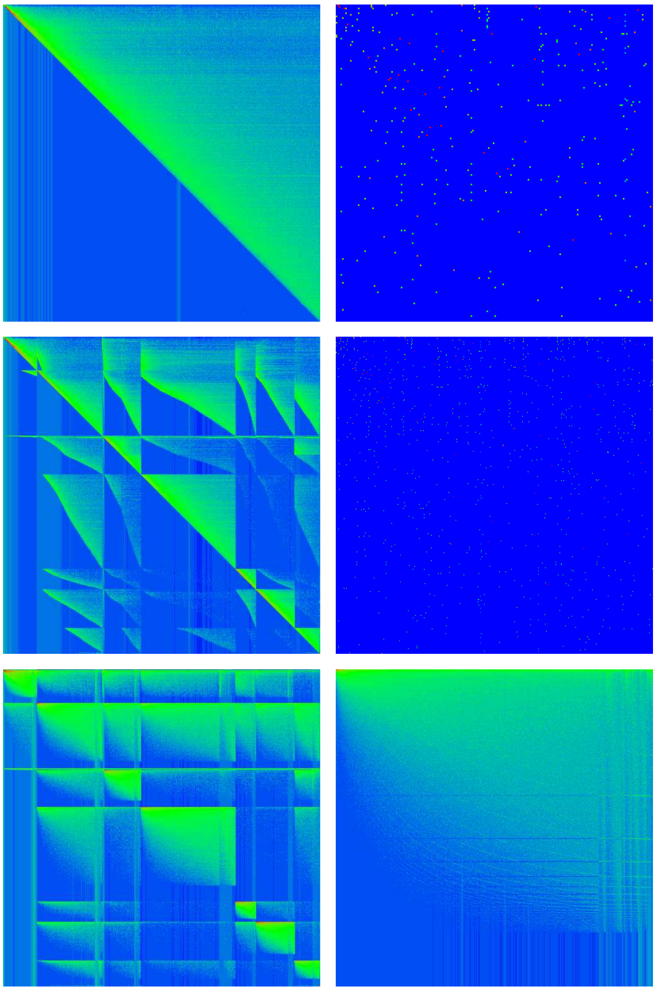

FIG. 1: (Color online) Different order representations of the Google matrix of the CNPR ($\alpha = 1$). *Left column:* The top panel shows the density of matrix elements $G_{tt'}$ in the basis of the publication time index t (and t'). The middle panel shows the density of matrix elements in the basis of journal ordering according to: Phys. Rev. Series I, Phys. Rev., Phys. Rev. Lett., Rev. Mod. Phys., Phys. Rev. A, B, C, D, E, Phys. Rev. STAB and Phys. Rev. STPER with time ordering inside each journal. The bottom panel shows the same as middle panel but with PageRank index ordering inside each journal. Note that the journals Phys. Rev. Series I, Phys. Rev. STAB and Phys. Rev. STPER are not clearly visible due to a small number of published papers. Also Rev. Mod. Phys. appears only as a thin line with 2-3 pixels (out of 500) due to a limited number of published papers. The three left panels and the bottom right panel show the coarse-grained density of matrix elements done on 500 × 500 square cells for the entire network. *Right column:* Matrix elements $G_{KK'}$ are shown in the basis of PageRank index K (and K') with the range $1 \leq K, K' \leq 200$ (top panel); $1 \leq K, K' \leq 400$ (middle panel); $1 \leq K, K' \leq N$ (bottom panel). Color shows the amplitude (or density) of matrix elements G changing from blue for zero value to red at maximum value. The PageRank index K is determined from the PageRank vector at $\alpha = 0.85$.

Rev journal when inside each journal the articles are ordered by their PageRank index K. The right panels show the matrix elements of G on different scales, when all articles are ordered by the PageRank index K.

The dependence of number of no-zero links N_G, between nodes with PageRank index being less than K, on K is shown in Fig. 2 (left panel). We see that compared to the other networks of universities, Wikipedia and Twitter studied in [13] we have for CNPR the lowest values of N_G/K practically for all available K values. This reflects weak links between top PageRank articles of CNPR being in contrast with Twitter which has very high interconnection between top PageRank nodes. Since the matrix elements $G_{KK'}$ are inversely proportional to the number of links we have very strong average matrix elements for CNPR at top K values (see Fig. 2 (right panel)).

In the following we present the results of numerical and analytical analysis of the spectrum of the CNPR matrix G.

A. Nearly nilpotent matrix structure

The triangular structure of the *CNPR* Google matrix in time index (see Fig. 1) has important consequences for the eigenvalue spectrum λ defined by the equation for the eigenstates $\psi_i(j)$:

$$\sum_{j'} G_{jj'}\psi_i(j') = \lambda_i \psi_i(j) \ . \qquad (2)$$

The spectrum of G at $\alpha = 1$, or the spectrum of S, obtained by the Arnoldi method [17, 18] with the Arnoldi dimension $n_A = 8000$, is shown in Fig. 3. For comparison we also show the case of reduced CNPR without Rev. Mod. Phys.. We see that the spectrum of the reduced case is rather similar to the spectrum of the full CNPR.

The matrix S can be decomposed on invariant subspaces S_{ss}, the core space S_{cc} with fully connected nodes, and the coupling block S_{sc}, thus being presented in the form [10]:

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix} \ . \qquad (3)$$

The subspace-subspace block S_{ss} is actually composed of many diagonal blocks for each of the invariant subspaces. Each of these blocks corresponds to a column sum normalized matrix of the same type as G and has therefore at least one unit eigenvalue thus explaining the high degeneracy of S eigenvalue $\lambda = 1$. This structure is discussed in detail in [10].

A network with a similar triangular structure, constructed from factor decompositions of integer numbers, was previously studied in [19]. There it was analytically shown that the corresponding G has only a small number of non-vanishing eigenvalues and that the numerical diagonalization methods, including the Arnoldi method, are facing subtle difficulties of numerical stability due to large Jordan blocks associated to the highly degenerate zero eigenvalue. The numerical diagonalization of these Jordan blocks is highly sensitive to numerical round-off errors. For example a perturbed Jordan block of dimension D associated to the eigenvalue zero and with a perturbation ε in the opposite corner has eigenvalues on a complex circle of radius $\varepsilon^{1/D}$ [19] which may became very large for sufficient large D even for $\varepsilon \sim 10^{-15}$. Therefore in presence of many such Jordan blocks the numerical diagonalization methods create rather big "artificial clouds" of incorrect eigenvalues.

In the examples studied in [19] these clouds extended up to eigenvalues $|\lambda| \approx 0.01$. The spectrum for the Physical Review network shown in Fig. 3 shows also a sudden increase of the density of eigenvalues below $|\lambda| \approx 0.3-0.4$ and one needs to be concerned if these eigenvalues are "real" or only an artifact of the same type of numerical instability. Actually, we find that the eigenvalues of Fig. 3 below $|\lambda| \approx 0.3 - 0.4$ are changed completely in a random way if we apply to the network or the numerical algorithm certain transformations or modifications which are *mathematically neutral* but which have a different effect on the numerical round-off errors (e.g. a permutation of the network nodes, keeping the same network-link structure, or simply changing the evaluation order of the sums used for the scalar products between vectors in the Gram-Schmidt orthogonalization for the Arnoldi method). This clearly indicates that these eigenvalues are not reliable due to problems in the numerical evaluation.

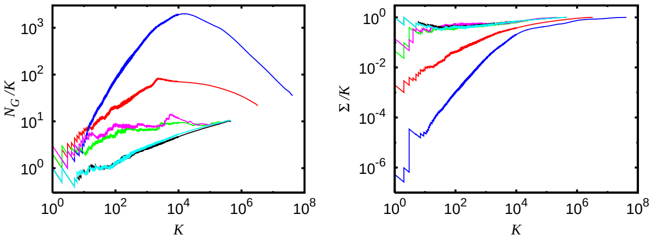

FIG. 2: (Color online) *Left panel:* dependence of the linear density N_G/K of nonzero elements of the adjacency matrix among top PageRank nodes on the PageRank index K for the networks of Twitter (blue curve), Wikipedia (red curve), Oxford University 2006 (magenta curve), Cambridge University 2006 (green curve), with data taken from Ref. [12], and Physical Review all journals (cyan curve) and Physical Review without Rep. Mod. Phys. (black curve) (curves from top to bottom at $K = 100$). *Right panel:* dependence of the quantity Σ/K on the PageRank index K with $\Sigma = \sum_{K_1<K,\,K_2<K} G_{K_1,K_2}$ being the weight of the Google matrix elements inside the $K \times K$ square of top PageRank indexes. The curves correspond to the same networks as in the left panel: Physical Review without Rep. Mod. Phys. (black curve), Physical Review all journals (cyan curve), Oxford University 2006 (magenta curve), Cambridge University 2006 (green curve), Wikipedia (red curve), and Twitter (blue curve) (curves from top to bottom at $K = 1$).

The theory of [19] is based on the exact triangular structure of the matrix S_0 which appears in the representation of $S = S_0 + ed^T/N$ (see also below Eq. 4). In fact the matrix S_0 is obtained from the adjacency matrix by normalizing the sum of the elements in non-vanishing columns to unity and simply keeping at zero vanishing columns. For the network of integers [19] this matrix is nilpotent with $S_0^l = 0$ for a certain modest value of l being much smaller than the network size $l \ll N$. However, for CNPR the matrix S_0 is not exactly nilpotent despite the overall triangular matrix structure visible in Fig. 1. Even though most of the non-vanishing matrix elements $(S_0)_{tt'}$ (whose total number is equal to the number of links $N_\ell = 4691015$) are in the upper triangle $t < t'$ there are a few non-vanishing elements in the lower triangle $t > t'$ (whose number is 12126 corresponding to 0.26 % of the total number of links [27]). The reason is that in most cases papers cite other papers published earlier but in certain situations for papers with close publication date the citation order does not always coincide with the publication order. In some cases two papers even mutually cite each other. In the following we will call these cases “future citations”. The rare non-vanishing matrix elements due to future citations are not visible in the coarse grained matrix representation of Fig. 1 but they are responsible for the fact that S_0 of CMPR is not nilpotent and that there are also a few invariant subspaces. On a purely triangular network one can easily show the absence of invariant subspaces (smaller than the full network size) when taking into account the extra columns due to the dangling nodes.

However, despite the effect of the future citations the matrix S_0 is still partly nilpotent. This can be seen by multiplying a uniform initial vector e (with all components being 1) by the matrix S_0 and counting after each iteration the number N_i of non-vanishing entries [28] in the resulting vector $S_0^i e$. For a nilpotent matrix S_0 with $S_0^l = 0$ the number N_i becomes obviously zero for $i \ge l$. On the other hand, since the components of e and the non-vanishing matrix elements of S_0 are positive, one can easily verify that the condition $S_0^l e = 0$ for some value l also implies $S_0^l \psi = 0$ for an arbitrary initial (even complex) vector ψ which shows that S_0 must be nilpotent with $S_0^l = 0$.

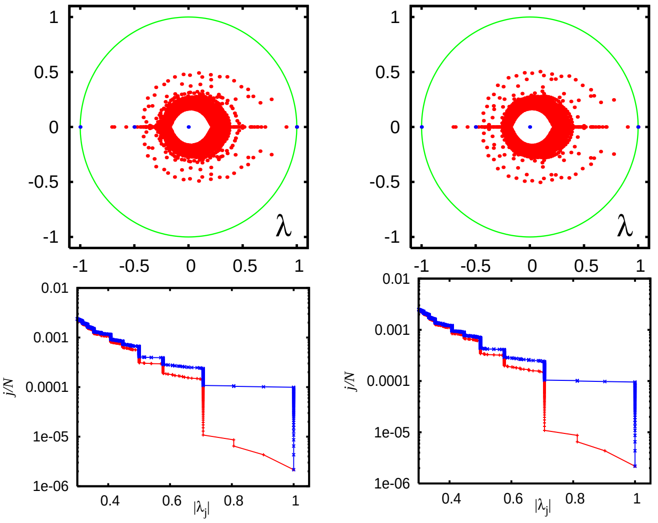

FIG. 3: (Color online) Spectrum of S for CNPR (reduced CNPR without Rev. Mod. Phys.) shown on left panels (right panels). *Top panels:* Subspace eigenvalues (blue dots) and core space eigenvalues (red dots) in λ-plane (green curve shows unit circle); there are 27 (26) invariant subspaces, with maximal dimension 6 (6) and the sum of all subspace dimensions is $N_s = 71$ (75). The core space eigenvalues are obtained from the Arnoldi method applied to the core space subblock S_{cc} of S with Arnoldi dimension $n_A = 8000$ as explained in Ref. [10] and using standard double-precision arithmetic. *Bottom panels:* Fraction j/N of eigenvalues, shown in a logarithmic scale, with $|\lambda| > |\lambda_j|$ for the core space eigenvalues (red bottom curve) and all eigenvalues (blue top curve) from raw data of top panels. The number of eigenvalues with $|\lambda_j| = 1$ is 45 (43) of which 27 (26) are at $\lambda_j = 1$; this number is identical to the number of invariant subspaces which have each one unit eigenvalue.

In Fig. 4 we see that for the CNPR the value of N_i saturates at a value $N_{sat} = 273490$ for $i \ge 27$ which is 59% of the total number of nodes $N = 463348$ in the network. On one hand the (small) number of future citations ensures that the saturation value of N_i is not zero but on the other hand it is smaller than the total number of nodes by a macroscopic factor. Mathematically the first iteration $e \to S_0 e$ removes the nodes corresponding to empty (vanishing) lines of the matrix S_0 and the next iterations remove the nodes whose lines in S_0 have become empty after having removed from the network the non-occupied nodes due to previous iterations. For each node removed during this iteration process one can construct a vector belonging to the Jordan subspace of S_0 associated to the eigenvalue 0. In the following we call this subspace *generalized kernel*. It contains all eigenvectors of S_0^j associated to the eigenvalue 0 where the integer j is the size of the largest 0-eigenvalue Jordan block. Obviously the dimension of this generalized kernel of S_0 is larger or equal than $N - N_{sat} = 189857$ but we will see later that its actual dimension is even larger and quite close to N. We will argue below that most (but not all) of the vectors in the generalized kernel of S_0 also belong to the generalized kernel of S which differs from S_0 by the extra contributions due to the dangling nodes. The high dimension of the generalized kernel containing many large 0-eigenvalue Jordan subspaces explains very clearly the numerical problem due to which the eigenvalues obtained by the double-precision Arnoldi method are not reliable for $|\lambda| < 0.3 - 0.4$.

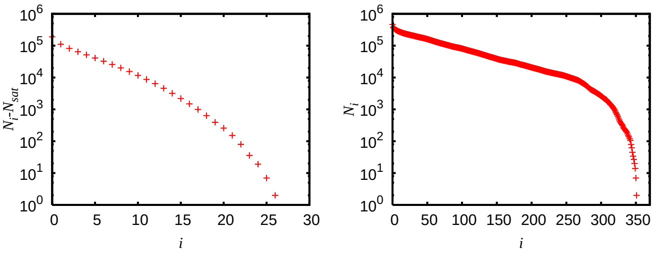

FIG. 4: (Color online) Number of occupied nodes N_i (i.e. positive elements) in the vector $S_0^i\,e$ versus iteration number i (red crosses) for the CNPR (left panel) and the triangular CNPR (right panel). In both cases the initial value is the network size $N_0 = N = 463348$. For the CNPR N_i saturates at $N_i = N_{sat} = 273490 \approx 0.590N$ for $i \geq 27$ while for the triangular CNPR N_i saturates at $N_i = 0$ for $i \geq 352$ confirming the nilpotent structure of S_0. In the left panel the quantity $N_i - N_{sat}$ is shown in order to increase visibility in the logarithmic scale.

B. Spectrum for the triangular CNPR

In order to extend the theory for the triangular matrices developed in [19] we consider the triangular CNPR obtained by removing all future citation links $t' \to t$ with $t \geq t'$ from the original CNPR. The resulting matrix S_0 of this reduced network is now indeed nilpotent with $S_0^{l-1} \neq 0$, $S_0^l = 0$ and $l = 352$ which is much smaller than the network size. This is clearly seen from Fig. 4 showing that N_i, calculated from the triangular CNPR, indeed saturates at $N_i = 0$ for $i \geq 352$. According to the arguments of [19], and additional demonstrations given below, there are at most only $l = 352$ non-zero eigenvalues of the Google matrix at $\alpha = 1$. This matrix has the form

$$S = S_0 + (1/N)\,e\,d^T \tag{4}$$

where d and e are two vectors with $e(n) = 1$ for all nodes $n = 1, \ldots, N$ and $d(n) = 1$ for dangling nodes n (corresponding to vanishing columns in S_0) and $d(n) = 0$ for the other nodes. In the following we call d the dangling vector. The extra contribution $e\,d^T/N$ just replaces the empty columns (of S_0) with $1/N$ entries at each element and d^T is the line vector obtained as the transpose of the column vector d.

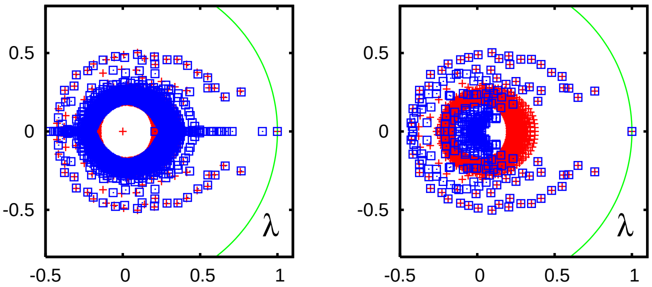

FIG. 5: (Color online) *Left Panel:* Comparison of the core space eigenvalue spectrum of S for CNPR (blue squares) and triangular CNPR (red crosses). Both spectra are calculated by the Arnoldi method with $n_A = 4000$ and standard double-precision. *Right Panel:* Comparison of the numerically determined non-vanishing 352 eigenvalues obtained from the representation matrix (12) (blue squares) with the spectrum of triangular CNPR (red crosses) already shown in the left panel. Numerics is done with standard double-precision.

In the left panel of Fig. 5 we compare the core space spectrum of S for CNPR and triangular CNPR (data are obtained by the Arnoldi method with $n_A = 4000$ and standard double-precision). We see that the largest complex eigenvalues are rather close for both cases but in the full network we have a lot of eigenvalues on the real axis (with $\lambda < -0.3$ or $\lambda > 0.4$) which are absent for the triangular CNPR. Furthermore, both cases suffer from the same problem of numerical instability due to large Jordan blocks.

Let us briefly remind the analytical theory of [19] for pure triangular networks with a nilpotent matrix S_0 such that $S_0^l = 0$. For this we define the coefficients:

$$c_j = d^T S_0^j\,e/N \quad , \quad b_j = e^T S_0^j\,e/N \tag{5}$$

which are non-zero only for $j = 0, 1, \ldots, l-1$. The fact that the non-vanishing columns of S_0 are sum normalized and that the other columns (corresponding to dangling nodes) are zero can be written as: $e^T S_0 = e^T - d^T$ implying $d^T = e^T(\mathbb{1} - S_0)$. Using this identify and the fact that $S_0^k = 0$ for $k \geq l$ we find:

$$\sum_{k=j}^{l-1} c_k = d^T(\mathbb{1} - S_0)^{-1} S_0^j\,e/N = e^T S_0^j\,e/N = b_j \tag{6}$$

and in particular for $j = 0$ we obtain the sum rule $\sum_{k=0}^{l-1} c_k = 1$ and for $j = l-1$ the identity $b_{l-1} = c_{l-1}$.

Consider now a right eigenvector ψ of S with eigenvalue λ. If $d^T\psi = 0$ we find from (4) that ψ is also an eigenvector of S_0 and since S_0 is nilpotent the eigenvalue must be $\lambda = 0$. Therefore for $\lambda \neq 0$ we have necessarily $d^T\psi \neq 0$ and with the appropriate normalization of ψ we have $d^T\psi = 1$ that implies together with the eigenvalue equation: $\psi = (\lambda\mathbb{1} - S_0)^{-1}\,e/N$ where the matrix inverse is well defined for $\lambda \neq 0$. The eigenvalue is determined by the condition:

$$0 = \lambda^l (1 - d^T \psi) = \lambda^l \left(1 - d^T \frac{\mathbb{1}}{\lambda \mathbb{1} - S_0} e/N \right) \quad . \tag{7}$$

Since S_0 is nilpotent we may expand the matrix inverse in a finite series and therefore the eigenvalue λ is the zero of the reduced polynomial of degree l:

$$\mathcal{P}_r(\lambda) = \lambda^l - \sum_{j=0}^{l-1} \lambda^{l-1-j} \, c_j \tag{8}$$

where the coefficients c_j are given by (5). Using $d^T = e^T(\mathbb{1} - S_0)$ we may rewrite (7) in the form:

$$0 = \lambda^l \left(1 - e^T \frac{\mathbb{1} - S_0}{\lambda \mathbb{1} - S_0} e/N \right) = (\lambda - 1)\lambda^l \, e^T \frac{\mathbb{1}}{\lambda \mathbb{1} - S_0} e/N \tag{9}$$

which gives another expression for the reduced polynomial:

$$\mathcal{P}_r(\lambda) = (\lambda - 1) \sum_{j=0}^{l-1} \lambda^{l-1-j} \, b_j \tag{10}$$

using the coefficients b_j and confirming explicitly that $\lambda = 1$ is indeed an eigenvalue of S. The expression (10) can also be obtained by a direct calculation from (6) and (8).

Since the reduced polynomial has at most l zeros λ_j ($\neq 0$ since $c_{l-1} = b_{l-1} \neq 0$) we find that there are at most l non-vanishing eigenvalues of S given by these zeros. They can also be obtained as the eigenvalues of a "small" $l \times l$ matrix. To see this let us define the following set of vectors v_j for $j = 1, \ldots, l$ by $v_j = c_{j-1}^{-1} \, S_0^{j-1} \, e/N$ where we have chosen to apply the prefactor c_{j-1}^{-1} to the vector $S_0^{j-1} \, e/N$ [29]. From (4) and (5) one finds that Sv_j can be expanded in the other vectors v_k as

$$Sv_j = \frac{c_j}{c_{j-1}} \, v_{j+1} + c_0 \, v_1 = \sum_{k=1}^{l} \bar{S}_{kj} \, v_k \tag{11}$$

where $\bar{S}_{kj}$ are the matrix elements of the $l \times l$ representation matrix

$$\bar{S} = \begin{pmatrix} c_0 & c_0 & \cdots & c_0 & c_0 \\ c_1/c_0 & 0 & \cdots & 0 & 0 \\ 0 & c_2/c_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & c_{l-1}/c_{l-2} & 0 \end{pmatrix} \quad . \tag{12}$$

Note that for the last vector v_l we have $Sv_l = c_0 \, v_1$ since $c_l = 0$ and therefore the matrix $\bar{S}$ provides a closed and mathematically exact representation of S on the l-dimensional subspace generated by $v_1, \ldots, v_l$. Furthermore one can easily verify (by a recursive calculation in l) that the characteristic polynomial of $\bar{S}$ coincides with the reduced polynomial (8). Therefore numerical diagonalization of $\bar{S}$ provides an alternative method to compute the non-vanishing eigenvalues of S. In principle one can also determine directly the zeros of the reduced polynomial by the Newton-Maehly method and in [19] this was indeed done for cases with very modest values of $l \leq 29$. However, here for the triangular CNPR we have $l = 352$ and the coefficients c_j become very small, especially: $c_{l-1} \approx 3.6 \times 10^{-352}$ a number which is (due to the exponent) outside the range of 64 bit standard double-precision numbers (IEEE 754) with 52 bits for the mantissa, 10 bits for the exponent (with respect to 2) and two bits for the signs of mantissa and exponent. This exponent range problem is not really serious and can for example be circumvented by a smart reformulation of the algorithm to evaluate the ratio $\mathcal{P}_r(\lambda)/\mathcal{P}_r'(\lambda)$ using only ratios c_j/c_{j-1} which do not have this exponent range problem. However, it turns out that in this approach the convergence of the Newton-Maehly method using double-precision arithmetic is very bad for many zeros and does not provide reliable results. Below we show how this problem can be solved using high precision calculations but for the moment we mention that one may also try another approach by diagonalizing numerically the representation matrix $\bar{S}$ given in (12) which also depends on the ratios c_j/c_{j-1}.

In the right panel of Fig. 5 we compare the numerical double-precision spectra of $\bar{S}$ with the results of the Arnoldi method with double-precision and the uniform initial vector e as start vector for the Arnoldi iterations. We remind that the Arnoldi method determines an orthonormal set of vectors $\zeta_1, \zeta_2, \zeta_3, \ldots, \zeta_{n_A}$ where the first vector ζ_1 is obtained by normalizing a given initial vector and ζ_{j+1} is obtained by orthonormalizing $S\zeta_j$ to the previous vectors determined so far. It is obvious due to (11) that for the initial uniform vector e each ζ_j is given by a linear combination of the vectors v_k with $k = 1, \ldots, j$. Since the subspace of v_k for $k = 1, \ldots, l$ is closed with respect to applications of S the Arnoldi method should, in theory, break off at $n_A = l$ with a zero coupling element. The latter is given as the norm of $S\zeta_l$ othogonalized to $\zeta_1, \ldots, \zeta_l$ and if this norm vanishes the vector ζ_{l+1} cannot be constructed and the Arnoldi method has completely explored an S-invariant subspace of dimension l.

However, due to a strong effect of round-off errors and the fact that the vectors v_j are numerically "nearly" linearly dependent the last coupling element does not vanish numerically (when using double-precision) and the Arnoldi method produces a cloud of numerically incorrect eigenvalues due to the Jordan blocks which are mathematically outside the representation space (defined by the vectors v_j) but which are still explored due to round-off errors and clearly visible in Fig. 5. The double-precision spectrum of $\bar{S}$ seems to provide well defined eigenvalues in the range where the Arnoldi method produces the "Jordan block cloud" but outside this cloud both spectra coincide only partly, mainly for the eigenvalues with largest modulus and positive real part. For the eigenvalues with negative real part there are considerable deviations. As we will see later the eigenvalues produced by the Arnoldi method at double-precision are reliable provided that they are well *outside* the Jordan block cloud of incorrect eigenvalues. Therefore the deviations outside the Jordan block cloud show that the numerical double-precision diagonalization of the representation matrix $\bar{S}$ is not reliable as well but here the effect of numerical errors is quite different as for the Arnoldi method as it is explained below.

We have tried to determine the zeros of the reduced polynomial using higher precision numbers with 80 or even 128 bits (quadruple precision) which helps to solve the (minor) exponent range problem because these formats use more bits for the exponent. However, there are indeed two other serious numerical problems. First it turns out that in a certain range of the complex plane around $\mathrm{Re}(\lambda) \approx -0.1$ to -0.2 and $\mathrm{Im}(\lambda) \leq 0.1$ the numerical evaluation of the polynomial suffers in a severe way from an alternate sign problem with a strong loss of significance. Second the zeros of the polynomial depend in a very sensitive way on the precision of the coefficients c_j (see below). We have found that even 128 bit numbers are not sufficient to obtain all zeros with a reasonable graphical precision.

Therefore we use the very efficient GNU Multiple Precision Arithmetic Library (GMP library) [30]. With this library one has 31 bits for the exponent and one may chose an arbitrary number of bits for the mantissa. We find that using 256 bits (binary digits) for the mantissa the complex zeros of the reduced polynomial can be determined with a precision of 10^{-18}. In this case the convergence of the Newton-Maehly method is very nice and we obtain that the sum (and product) of the complex zeros coincide with a high precision with the theoretical values c_0 (respectively: $(-1)^{l-1}c_{l-1}$) due to (8). We have also tested different ways to evaluate the polynomial, such as Horner scheme versus direct evaluation of the sum and for both methods using both expressions (8) and (10). It turns out that with 256 binary digits during the calculation the zeros obtained by the different variants of the method coincide very well within the required precision of 10^{-18}. Of course the coefficients c_j or b_j given by (5) need also to be evaluated with the precision of 256 binary digits but there is no problem of using high precision vectors since the non-vanishing matrix elements of S_0 are rational numbers that allow to perform the evaluation of the vectors $S_0^j e/N$ with arbitrary precision. We also tested a random modification of c_j according to $c_j \to c_j(1+10^{-16}X)$ where X is a random number in the interval $]-0.5, 0.5[$. This modification gives significant differences of the order of 10^{-2} to 10^{-1} for some of the complex zeros and which are well visible in the graphical representation of the spectra. Therefore, the spectrum depends in a very sensitive way on these coefficients and it is now quite clear that numerical double-precision diagonalization of $\bar{S}$, which depends according to (12) on the values c_j, cannot provide accurate eigenvalues simply because the double-precision round-off errors of c_j imply a sensitive change of eigenvalues. In particular some of the numerical eigenvalues of $\bar{S}$ differ quite strongly from the high precision zeros of the reduced polynomial.

In order to study more precisely the effect of the numerical instability of the Arnoldi method due to the Jordan blocks we also use the GMP library to increase the numerical precision of the Arnoldi method. To be precise we implement the first part of this method, the *Arnold iteration* in which the $n_A \times n_A$ Arnoldi representation matrix is determined by the Gram-Schmidt orthogonalization procedure, using high precision numbers while for the second step, the numerical diagonalization of this representation matrix, we keep the standard double-precision. It turns that only the first step is numerically critical. Once the Arnoldi representation matrix is obtained in a careful and precise way, it is numerically well conditioned and its numerical diagonalization works well with only double-precision.

In Fig. 6 we compare the exact spectrum obtained by the high precision determination of the zeros of the reduced polynomial (using 256 bits) with the spectra of the Arnold method for 52 bits (corresponding to the mantissa of double-precision numbers), 256 bits, 512 bits and 1280 bits. Here we use for the Arnoldi method a uniform initial vector and the Arnold dimension $n_A = l = 352$. In this case, as explained above, in theory the Arnoldi method should provide the exact $l = 352$ non-vanishing eigenvalues (in absence of round-off errors).

However, with the precision of 52 bits we have a considerable number of eigenvalues on a circle of radius ≈ 0.3 centered at 0.05 indicating a strong influence of round-off errors due to the Jordan blocks. Increasing the precision to 256 (or 512) bits implies that the number of correct eigenvalue increases and the radius of this circle decreases to 0.13 (or 0.1) and in particular it does not extend to all angles. We have to increase the precision of the Arnoldi method to 1280 bits to have a perfect numerical confirmation that the Arnoldi method explores the exact invariant subspace of dimension $l = 352$ and generated by the vectors v_j. In this case the eigenvalues obtained from the Arnoldi method and the high-precision zeros of the reduced polynomial coincide with an error below 10^{-14} and in particular the Arnoldi method provides a nearly vanishing coupling matrix element at the last iteration confirming that there is indeed an exact decoupling of the Arnoldi matrix and an invariant closed subspace of dimension 352.

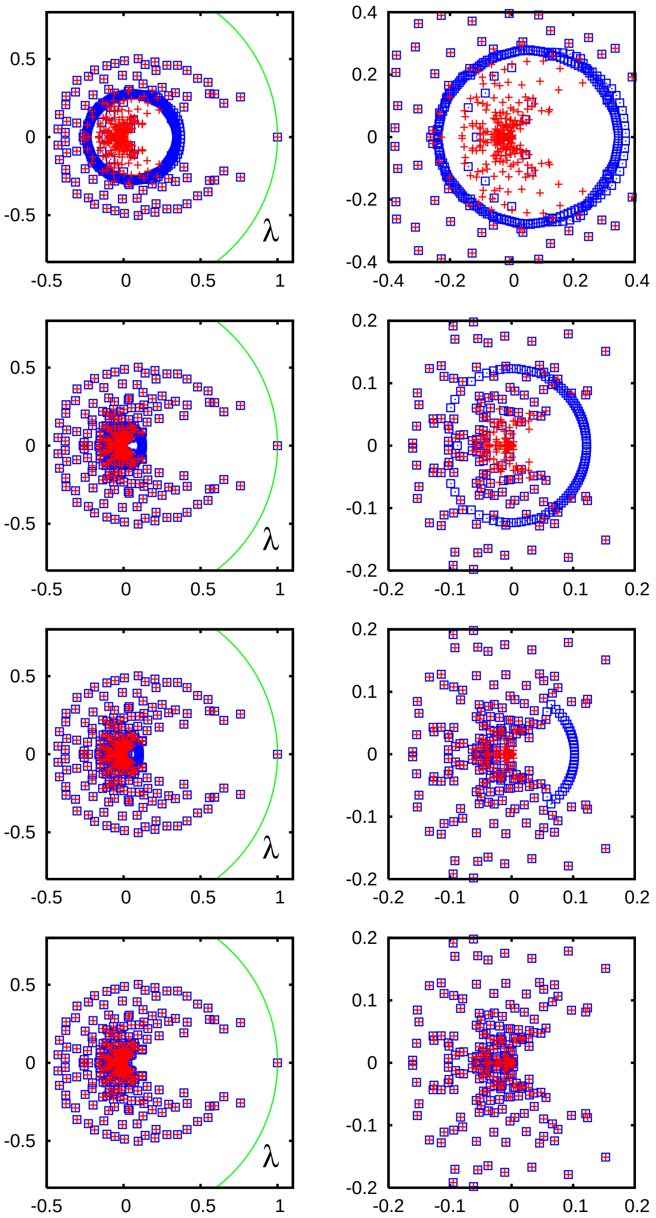

FIG. 6: (Color online) Comparison of the numerically accurate 352 non-vanishing eigenvalues of S matrix of triangular CNPR, determined by the Newton-Maehly method applied to the reduced polynomial (8) with a high-precision calculation of 256 binary digits (red crosses, all panels), with eigenvalues obtained by the Arnoldi method at different numerical precisions (for the determination of the Arnoldi matrix) for triangular CNPR and Arnoldi dimension $n_A = 352$ (blue squares, all panels). The first row corresponds to the numerical precision of 52 binary digits for standard double-precision arithmetic. The second (third, fourth) row corresponds to the precision of 256 (512, 1280) binary digits. All high precision calculations are done with the library GMP [30]. The panels in the left column show the complete spectra and the panels in the right columns show the spectra in a zoomed range: $-0.4 \leq \mathrm{Re}(\lambda), \mathrm{Im}(\lambda) \leq< 0.4$ for the first row or $-0.2 \leq \mathrm{Re}(\lambda), \mathrm{Im}(\lambda) \leq 0.2$ for the second, third and fourth rows.

The results shown in Fig.6 clearly confirm the above theory and the scenario of the strong influence of Jordan blocks on the round-off errors. In particular, we find that in order to increase the numerical precision it is only necessary to implement the first step of the method, the Arnoldi iteration, using high precision numbers while the numerical diagonalization of the Arnoldi representation matrix can still be done using standard double-precision arithmetic. We also observe, that even for the case with lowest precision of 52 bits the eigenvalues obtained by the Arnoldi method are numerically accurate provided that there are well outside the circle (or cloud) of numerically incorrect eigenvalues.

C. High precision spectrum of the whole CNPR

Based on the observation that a high precision implementation of the Arnoldi method is useful for the triangular CNPR, we now apply the high precision Arnoldi method with 256, 512 and 756 bits and $n_A = 2000$ to the original CNPR. The results for the core space eigenvalues are shown in Fig. 7 where we compare the spectrum of the highest precision of 756 bits with lower precision spectra of 52, 256 and 512 bits. As in Fig. 6 for the triangular CNPR, for CNPR we also observe that the radius and angular extension of the cloud or circle of incorrect Jordan block eigenvalues decreases with increasing precision. Despite the lower number of $n_A = 2000$ as compared to $n_A = 8000$ of Fig. 3 the number of accurate eigenvalues with 756 bit precision is certainly considerably higher.

The higher precision Arnoldi method certainly improves the quality of the smaller eigenvalues, e.g. for $|\lambda| < 0.3 - 0.4$, but it also implies a strange shortcoming as far as the degeneracies of certain particular eigenvalues are concerned. This can be seen in Fig. 8 which shows the core space eigenvalues $|\lambda_j|$ versus the level number j for various values of the Arnoldi dimension and the precision. In these curves we observe flat plateaux at certain values $|\lambda_j| = 1/\sqrt{n}$ with $n = 2, 3, 4, 5, \ldots$ corresponding to degenerate eigenvalues which turn out to be real but with positive or negative values: $\lambda_j = \pm 1/\sqrt{n}$. For fixed standard double-precision arithmetic with 52 binary digits the degeneracies increase with increasing Arnoldi dimension and seem to saturate for $n_A \geq 4000$. However at the given value of $n_A = 2000$ the degeneracies *decrease* with increasing precision of the Arnoldi method. Apparently the higher precision Arnoldi method is less able to determine the correct degeneracy of a degenerate eigenvalue.

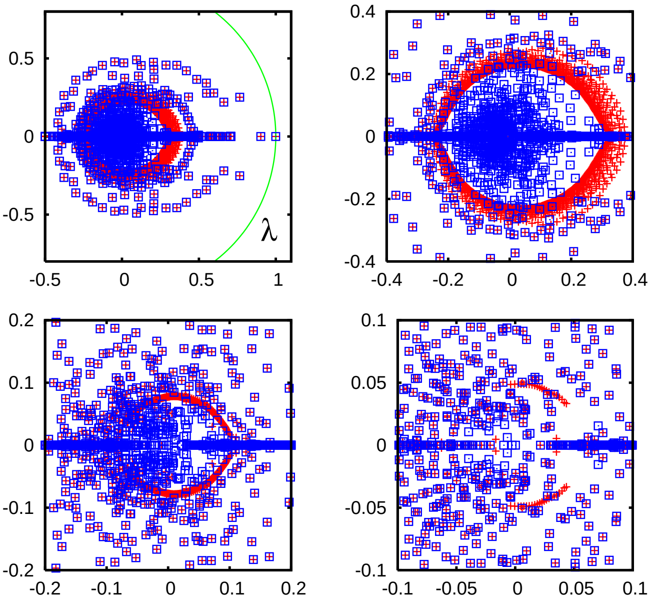

FIG. 7: (Color online) Comparison of the core space eigenvalue spectrum of S of CNPR, obtained by the high precision Arnoldi method using 768 binary digits (blue squares, all panels), with lower precision data of the Arnoldi method (red crosses). In both top panels the red crosses correspond to double-precision with 52 binary digits (extended range in left top panel and zoomed range in right top panel). In the bottom left (right) panel red crosses correspond to the numerical precision of 256 (512) binary digits. In these two cases only a zoomed range is shown. The eigenvalues outside the zoomed ranges coincide for both data sets up to graphical precision. In all cases the Arnoldi dimension is $n_A = 2000$. High precision calculations are done with the library GMP [30].

This point can be understood as follows. In theory, assuming perfect precision, the simple version of Arnoldi method used here (in contrast to more complicated block Arnoldi methods) can only determine one eigenvector for a degenerate eigenvalue. The reason is that for a degenerate eigenvalue we have a particular linear combination of the eigenvectors for this eigenvalue which contribute in any initial vector (in other words "one particular" eigenvector for this eigenvalue) and during the Arnoldi iteration this particular eigenvector will be perfectly conserved and the generated Krylov space will only contain this and no other eigenvector for this eigenvalue. However, due to round-off errors we obtain at each step new random contributions from other eigenvectors of the same eigenvalue and it is only due to these round-off errors that we can see the flat plateaux in Fig. 8. Obviously, increasing the precision reduces this round-off error effect and the flat plateaux are indeed considerably smaller for higher precisions.

The question arises about the origin of the degenerate eigenvalues in the core space spectrum. In other examples, such as the WWW for certain university networks [10], the degeneracies, especially of the leading eigenvalue 1, could be treated by separating and diagonalizing the exact subspaces and the remaining core space spectrum contained much less or nearly no degenerate eigenvalues. However, here for the CNPR we have "only" 27 subspaces with maximal dimension of 6 containing 71 nodes in total. The eigenvalues due to these subspaces are 1, -1, -0.5, 0 with degeneracies 27, 18, 4, 22 (see blue dots in the upper panels of Fig. 3). These exact subspaces exist only due to the modest number of future citation links. Even when we take care that in all cases the Arnoldi method is applied to the core space without these 71 subspace nodes, there are still remain a lot of degenerate eigenvalues in the core space spectrum.

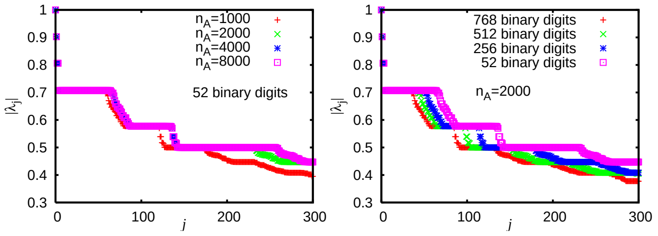

FIG. 8: (Color online) Modulus $|\lambda_j|$ of the core space eigenvalues of S of CNPR, obtained by the Arnoldi method, shown versus level number j. *Left panel:* data for standard double-precision with 52 binary digits with different Arnoldi dimensions $1000 \leq n_A \leq 8000$. *Right panel:* data for Arnoldi dimension $n_A = 2000$ with different numerical precisions between 52 and 768 binary digits.

In order to understand the mechanism of these degenerate core space eigenvalues we extend the argumentation of the last subsection for triangular CNPR to the case of nearly triangular networks. Consider again the matrix S given by Eq. (4) but now S_0 is not nilpotent. There are two groups of eigenvectors ψ of S with eigenvalue λ. The first group is characterized by the orthogonality $d^T\psi = 0$ of the eigenvector ψ with respect to the dangling vector d and the second group is characterized by the non-orthogonality $d^T\psi \neq 0$. In the following, we describe efficient methods to determine all eigenvalues of the first group and a considerable number of eigenvalues of the second group. We note that for the case of a purely triangular network the first group contains only eigenvectors for the eigenvalue 0 and the second group contains the eigenvectors for the l non-vanishing eigenvalues as discussed in the last subsection. In principle there are also complications due to generalized eigenvectors (associated to non-trivial Jordan blocks) but they appear mainly for zero eigenvalue and we for the moment do not discuss these complications.

First we note that the subspace eigenvectors of S belong to the first group because the nodes of the subspaces of S cannot contain dangling nodes which are by construction of S are linked to any other node and therefore belong to the core space. Since any subspace eigenvector ψ has non-vanishing values only for subspace nodes being different from dangling nodes we have obviously $d^T\psi = 0$. We also note that an eigenvector of S of the first group with $d^T\psi = 0$ is due to (4) also an eigenvector of S_0 with the same eigenvalue.

For the remaining eigenvectors in the first group one might try to diagonalize the matrix S_0 and check for each eigenvector of S_0 if the identity $d^T\psi = 0$ holds in which case we would obtain an eigenvector of S of the first group but generically, and apart from the subspace eigenvectors, there is no reason that eigenvectors of S_0 with isolated non-degenerate eigenvalues obey this identity. However, if we have an eigenvalue of S_0 with a degeneracy $m \geq 2$ we may construct by suitable linear combinations $m-1$ linearly independent eigenvectors of S_0 which also obey $d^T\psi = 0$ and therefore this eigenvalue with degeneracy m of S_0 is also an eigenvalue with degeneracy $m-1$ of S. In order to determine the degenerate eigenvalues of S_0 it is useful to determine the subspaces of S_0 which (in contrast to the subspaces of S) may contain dangling nodes. Actually, each dangling node is a trivial subspace of dimension 1 with a network matrix of size 1×1 and being zero. Explicitly we have implemented the following procedure: first we determine the subspaces of S (with 71 nodes in total) and remove these nodes from the network. Then we determine all subspaces of S_0 whose dimension is below 10. Each time such a subspace is found its nodes are immediately removed from the network. When we have tested in a first run all nodes as potential subspace nodes the procedure is repeated until no new subspaces of maximal dimension 10 are found since removal of former subspaces may have created new subspaces. Then the limit size of 10 is doubled to 20, 40, 80 etc. to ensure that we do not miss large subspaces. However, for the CNPR it turns out that the limit size of 10 allows to find all subspaces. In our procedure a subsequently found subspace may potentially have links to a former subspace leading to a block-triangular (and not block-diagonal structure as it was done in ref. [10]). This method to determine "relative" subspaces of a network already reduced by former subspaces is more convenient for the CNPR which is nearly triangular and it allows also to determine correctly all subspace eigenvalues by diagonalizing each relative subspace network. The removal of subspace nodes of S and S_0 reduces the network size from $N = 463348$ to 404959. In the next step we remove in the same way the subspaces of the transpose S_0^T of S_0 (since the eigenvalues of S_0^T and S_0 are the same) which reduces the network size furthermore to 90965. In total this procedure provides a block triangular structure of S_0 as:

$$S_0 = \begin{pmatrix} S_1 & * & \cdots & & & \cdots & * \\ 0 & S_2 & * & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ & & 0 & B & * & & \\ \vdots & & & 0 & T_1 & * & \vdots \\ \vdots & & & & 0 & T_2 & * \\ 0 & \cdots & & & \cdots & \ddots & \ddots \end{pmatrix} \qquad (13)$$

where S_1, S_2, ... represent the diagonal subblocks associated to the subspaces of S and S_0 while T_1, T_2, ... represent the diagonal subblocks associated to the subspaces of S_0^T and B is the "bulk" part for the remaining network of 90965 nodes. The stars represent potential non-vanishing entries whose values do not influence the eigenvalues of S_0. The subspace blocks S_1, S_2, ... and T_1, T_2, ... which are individually of maximal dimension 10 can be directly diagonalized and it turns that out of 372382 eigenvalues in these blocks only about 4000 eigenvalues (counting degeneracies) or 950 eigenvalues (non-counting degeneracies) are different from zero. Most of these eigenvalues are not degenerate and are therefore not eigenvalues of S but there are still quite many degenerate eigenvalues at $\lambda = \pm 1/\sqrt{n}$ with $n \geq 2$ taking small integer values and who are also eigenvalues of S with a degeneracy reduced by one.

Concerning the bulk block B we can write it in the form $B = B_0 + f_1\, e_1^T$ where f_1 is the first column vector of B and $e_1^T = (1,\, 0,\, \ldots,\, 0)$. The matrix B_0 is obtained from B by replacing its first column to zero. We can apply the above argumentation between S and S_0 in the same way to B and B_0, i.e. the degenerate eigenvalues of B_0 with degeneracy m are also eigenvalues of B with degeneracy $m-1$ (with eigenvectors obeying $e_1^T\psi = 0$) and therefore eigenvalues of S with degeneracy $m-2$. The matrix B_0 is decomposed in a similar way as in (13) with subspace blocks, which can be diagonalized numerically, and a new bulk block $\tilde{B}$ of dimension 63559 and which may be treated in the same way by taking out its first column. This procedure provides a recursive scheme which after 9 iterations stops with a final bulk block of zero size. At each iteration we keep only subspace eigenvalues with degeneracies $m \geq 2$ and which are joined with reduced degeneracies $m-1$ to the subspace spectrum of the previous iteration. For this joined spectrum we keep again only eigenvalues with degeneracies $m \geq 2$ which are joined with the subspace spectrum of the next higher level etc.

In this way we have determined all eigenvalues of S_0 with a degeneracy $m \geq 2$ which belong to the eigenvalues of S of the first group. Including the direct subspace of S there are 4999 non-vanishing eigenvalues (counting degeneracies) or 442 non-vanishing eigenvalues (non-counting degeneracies). The degeneracy of the zero

λ	degeneracy
1	27
-1	18
$\pm 1/\sqrt{2}$	27
$\pm 1/\sqrt{3}$	20
$1/2$	58
$-1/2$	52
$\pm 1/\sqrt{5}$	20
$\pm 1/\sqrt{6}$	52
$\pm 1/\sqrt{7}$	6
$\pm 1/\sqrt{8}$	44
$1/3$	47
$-1/3$	39
$\pm 1/\sqrt{10}$	33
$\pm 1/\sqrt{11}$	1
$\pm 1/\sqrt{12}$	85
$\pm 1/\sqrt{14}$	15
$\pm 1/\sqrt{15}$	46
$1/4$	52
$-1/4$	42
$\pm 1/\sqrt{18}$	29
$\pm 1/\sqrt{20}$	60
$\pm 1/\sqrt{21}$	30
$\pm 1/\sqrt{22}$	3
$\pm 1/\sqrt{24}$	69
$1/5$	20
$-1/5$	11

TABLE I: Degeneracies of the eigenvalues with largest modulus for the whole CNPR whose eigenvectors ψ belong to the first group and obey the orthogonality $d^T\psi = 0$ with the dangling vector d.

eigenvalue (or the dimension of the generalized kernel) is found by this procedure to be 455789 but this would only be correct assuming that there are no general eigenvectors of higher order (representation vectors of non-trivial Jordan blocks) which is clearly not the case. The Jordan subspace structure of the zero eigenvalue complicates the argumentation. Here at each iteration step the degeneracy has to be reduced from m to $m-D$ where $D > 1$ is the dimension of the maximal Jordan block since each generalized eigenvector at a given order has to be treated as an independent vector when constructing vectors obeying the orthogonality with respect to the dangling vector d. Therefore the degeneracy of the zero eigenvalue cannot be determined exactly but we may estimate its degeneracy of about ~ 455000 out of 463348 nodes in total. This implies that the number of non-vanishing eigenvalues is about $\sim 8000 - 9000$ which is considerably larger than the value of 352 for the triangular CNPR but still much smaller than the total network size.

In Table I we provide the degeneracies for some of the eigenvalues $\pm 1/\sqrt{n}$ for integer n in the range $1 \le n \le 25$. The degeneracies for $+1/\sqrt{n}$ and $-1/\sqrt{n}$ are identical for non-square numbers n (with non-integer $\sqrt{n}$) and different for square numbers (with integer $\sqrt{n}$). Apparently for non-square numbers the eigenvalues are only generated from effective 2×2 blocks:

$$\begin{pmatrix} 0 & 1/n_1 \\ 1/n_2 & 0 \end{pmatrix} \quad \Rightarrow \quad \lambda = \pm\frac{1}{\sqrt{n_1\, n_2}} \qquad (14)$$

with positive integers n_1 and n_2 such that $n = n_1\, n_2$ while for square numbers $n = m^2$ they may be generated by such blocks or by simple 1×1 blocks containing $1/m$ such that the degeneracy for $+1/\sqrt{n} = +1/m$ is larger than the degeneracy for $-1/\sqrt{n} = -1/m$. Furthermore, statistically the degeneracy is smaller for prime numbers n or numbers with less factorization possibilities and larger for numbers with more factorization possibilities. The Arnoldi method (with 52 bits for double-precision arithmetic and $n_A = 8000$) provides according to the sizes of the plateaux visible in Fig. 8 the overall approximate degeneracies ~ 60 for $|\lambda| = 1/\sqrt{2}$ (i.e. $\pm 1/\sqrt{2}$ counted together), ~ 50 for $|\lambda| = 1/\sqrt{3}$ and ~ 115 for $|\lambda| = 1/2$. These values are coherent with (but slightly larger than) the values 54, 40 and 110 taken from Table I. Actually, as we will see below, the slight differences between the degeneracies obtained from Fig. 8 and from Table I are indeed relevant and correspond to some eigenvalues of the second group which are close but not identical to $\pm 1/\sqrt{2}$, $\pm 1/\sqrt{3}$ or $\pm 1/2$ and do not contribute in Table I.

We now consider the eigenvalues λ of S for the eigenvectors of the second group with non-orthogonality $d^T\psi \neq 1$ or $d^T\psi = 1$ after proper renormalization of ψ. Now ψ cannot be an eigenvector of S_0 and λ is not an eigenvalue of S_0. As in the last subsection the eigenvalue equation $S\psi = \lambda\psi$, the condition $d^T\psi = 1$ and (4) imply that the eigenvalue λ of S is a zero of the rational function

$$\mathcal{R}(\lambda) = 1 - d^T \frac{1\!\!1}{\lambda 1\!\!1 - S_0} e/N = 1 - \sum_{j,q} \frac{C_{jq}}{(\lambda - \rho_j)^q} \qquad (15)$$

where we have formally expanded the vector e/N in eigenvectors of S_0 and with ρ_j being the eigenvalues of S_0 and q is the *order* of the eigenvector of ρ_j used in this expansion, i.e. $q = 1$ for simple eigenvectors and $q > 1$ for generalized eigenvectors of higher order due to Jordan blocks. Note that even the largest possible value of q for a given eigenvalue may be (much) smaller than its multiplicity m. Furthermore the case of simple repeating eigenvalues (with simple eigenvectors) with higher multiplicity $m > 1$ leads only to several identical terms $\sim (\lambda - \rho_j)^{-1}$ for any eigenvector of this eigenvalue thus all contributing to the coefficients C_{jq} and whose precise values we do not need to know in the following. For us the important point is that the second identity in (15) establishes that $\mathcal{R}(\lambda)$ is indeed a rational function whose denominator and numerator polynomials have the same degree and whose poles are (some of) the eigenvalues of S_0.

We mention that one can also show by a simple determinant calculation (similar to a calculation shown in [19] for triangular networks with nilpotent S_0) that:

$$P_S(\lambda) = P_{S_0}(\lambda)\,\mathcal{R}(\lambda) \tag{16}$$

where $P_S(\lambda)$ [or $P_{S_0}(\lambda)$] is the characteristic polynomial of S (S_0). Therefore those zeros of $\mathcal{R}(\lambda)$ which are not zeros of $P_{S_0}(\lambda)$ (i.e. not eigenvalues of S_0) are indeed zeros of $P_S(\lambda)$ (i.e. eigenvalues of S) since there are not poles of $\mathcal{R}(\lambda)$. Furthermore, generically the *simple* zeros $P_{S_0}(\lambda)$ also appear as poles in $\mathcal{R}(\lambda)$ and are therefore not eigenvalues of S. However, for a zero of $P_{S_0}(\lambda)$ (eigenvalue of S_0) with *higher multiplicity* $m > 1$ (and unless m is equal to the maximal Jordan block order q associated to this eigenvalue of S_0) the corresponding pole in $\mathcal{R}(\lambda)$ only reduces the multiplicity to $m-1$ (or $m-q$ in case of higher order generalized eigenvectors) and we have also a zero of $P_S(\lambda)$ (eigenvalue of S). Some of the eigenvalues of S_0, whose eigenvectors ψ are orthogonal to the dangling vector ($d^T\psi = 0$) and do not contribute in the expansion in (15), are not poles of $\mathcal{R}(\lambda)$ and therefore also eigenvalues of S. This concerns essentially the direct subspace eigenvalues of S which are also direct subspace eigenvalues of S_0 as already mentioned above. In total the identity (16) confirms exactly the above picture that there are two groups of eigenvalues and with the special role of direct subspace eigenvalues belonging to the first group.

Our aim is to determine numerically the zeros of the rational function $\mathcal{R}(\lambda)$. In order to evaluate this function we expand the first identity in (15) in a matrix geometric series and we obtain

$$\mathcal{R}(\lambda) = 1 - \sum_{j=0}^{\infty} c_j \lambda^{-1-j} \tag{17}$$

with the coefficients c_j defined in (5) and provided that this series converges. In the last subsection, where we discussed the case of a nilpotent matrix S_0 with $S_0^l = 0$, the series was finite and for this particular case we had $\mathcal{R}(\lambda) = \lambda^{-l}\mathcal{P}_r(\lambda)$ where $\mathcal{P}_r(\lambda)$ was the reduced polynomial defined in (8) and whose zeros provided the l nonvanishing eigenvalues of S for nilpotent S_0.

However, for the CNPR the series are infinite since all c_j are different from zero. One may first try a crude approximation and simply replace the series by a finite sum for $j < l$ and using some rather large cutoff value for l and determine the zeros in the same way as for the nilpotent case (high precision calculation of the zeros of the reduced polynomial of degree l). It turns that in this way we obtain correctly the largest core space eigenvalue of S as $\lambda_1 = 0.999751822283878$ which is also obtained by (any variant of) the Arnoldi method. However, the other zeros obtained by this approximation lie all on a circle of radius ≈ 0.9 in the complex plane and do not obviously represent any valid eigenvalues. Increasing the cutoff value l does not help either and it increases only the density of zeros on this circle. To understand this behavior we note that in the limit $j \to \infty$ the coefficients c_j behave as $c_j \propto \rho_1^j$ where $\rho_1 = 0.902448280519224$ is the largest eigenvalue of the matrix S_0 with an eigenvector nonorthogonal to d. Note that the matrix S_0 has also some degenerate eigenvalues at $+1$ and -1 but these eigenvalues are obtained from the direct subspace eigenvectors of S (which are also direct subspace eigenvectors of S_0) and which are orthogonal to the dangling vector d and do not contribute in the rational function (15). It turns actually out that the eigenvalue ρ_1 is also the largest *subspace space* eigenvalue of S_0 (after having removed the direct subspace nodes of S). By analyzing explicitly the small-dimensional subspace related to this eigenvalue one can show that ρ_1 is given as the largest solution of the polynomial equation $x^3 - \frac{2}{3}x - \frac{2}{15} = 0$ and can therefore be expressed as $\rho_1 = 2\,\mathrm{Re}\,[(9 + i\sqrt{119})^{1/3}]/(135)^{1/3}$. The asymptotic behavior $c_j \propto \rho_1^j$ is also confirmed by the direct numerical evaluation of c_j. Therefore the series (17) converges only for $|\lambda| > \rho_1$ and a simple (even very large) cutoff in the sum implies that only eigenvalues $|\lambda_j| > \rho_1$ can be determined as a zero of the finite sum. The only eigenvalue respecting this condition is the largest core space eigenvalue λ_1 given above.

One may try to improve this by a "better" approximation which consists of evaluating the sum exactly up to some value l and than to replace the remaining sum as a geometric series with the approximation: $c_j \approx c_l \rho_1^{j-l}$ for $j \geq l$ and with ρ_1 determined as the ratio $\rho_1 = c_l/c_{l-1}$ (which provides a sufficient approximation) or taken as its exact (high precision) value. This improved approximation results in $\mathcal{R}(\lambda) \approx \lambda^{-l}(\lambda - \rho_1)^{-1}\mathcal{P}(\lambda)$ with a polynomial $\mathcal{P}(\lambda)$ whose zeros provide in total four correct eigenvalues. Apart from λ_1 it also gives $\lambda_2 = 0.902445536212661$ (note that this eigenvalue of S is very close but different to the eigenvalue ρ_1 of S_0) and $\lambda_{3,4} = 0.765857950563684 \pm i\,0.251337495625571$ such that $|\lambda_{3,4}| = 0.806045245100386$. All these four core space eigenvalues coincide very well with the first four eigenvalues obtained from the Arnoldi method. However, the other zeros of the Polynomial $\mathcal{P}(\lambda)$ lie again on a circle, now with a reduced radius ≈ 0.7, and do not coincide with eigenvalues of S. This can be understood by the fact that the coefficients c_j obey for $j \to \infty$ the more precise asymptotic expression $c_j \approx C_1\rho_1^j + C_2\rho_2^j + C_2\rho_3^j + \ldots$ with the next eigenvalues $\rho_2 = 1/\sqrt{2} \approx 0.707$ and $\rho_3 = -\rho_2$. Here the first term $C_1\rho_1^j$ is dealt with analytically by the replacement of the geometric series but the other terms create a new convergence problem. Therefore the improved approximation allows only to determine the four core space eigenvalues with $|\lambda_j| > |\rho_{2,3}| = 1/\sqrt{2}$. To obtain more valid eigenvalues it seems to be necessary to sum up by geometric series many of the next terms, not only the next two terms due to ρ_2 and ρ_3, but also the following terms of smaller eigenvalues ρ_j of S_0. In other words the exact pole structure of the rational function $\mathcal{R}(\lambda)$ has be kept as best as possible.

Therefore due to the rational structure of the function $\mathcal{R}(\lambda)$ with many eigenvalues ρ_j of S_0 that determine its precise pole structure we suggest the following numerical approach using high precision arithmetic. For a given number p of binary digits, e.g. $p = 1024$, we determine the coefficients c_j for $j < l$ where the cutoff value

$$l \approx \frac{\ln(1-\rho_1) - p\ln(2)}{\ln(\rho_1)} \approx 6.753\,p + \text{const.} \tag{18}$$

is sufficiently large to evaluate the sum (17) accurately in the given precision of p binary digits (error below 2^{-p}) for *all complex values λ on the unit circle*, i.e. $|\lambda| = 1$, where the series converges well. Furthermore we choose a number n_R of "eigenvalues" we want to calculate, e.g. $n_R = 300$, and evaluate the rational function $\mathcal{R}(z)$ at $n_S = 2n_R + 1$ support points $z_j = \exp(2\pi i\, j/n_S)$ ($j = 0, \ldots, n_S - 1$) uniformly distributed on the unit circle using the series (17). Then we calculate the rational function $R_I(z)$ which interpolates $\mathcal{R}(z)$ at the n_S support points z_j, $R_I(z_j) = \mathcal{R}(z_j)$, using Thiele's interpolation formula. Then the numerator and denominator polynomials of $R_I(z)$ are both of degree n_R. Thiele's interpolation formula expresses $R_I(z)$ in terms of a continued fraction expansion using inverse differences. This method is quite standard and well described in the literature of numerical mathematics, see for example [31]. After having evaluated a table of n_S inverse differences (with $n_S^2/2$ operations) one can evaluate arbitrary values of $R_I(z)$ using the continued fraction expansion (with n_S operations). It is not very difficult to derive from the continued fraction expansion a recursive scheme to evaluate the values of the numerator and denominator polynomials separately as well as their derivatives. Using this scheme we determine the n_R complex zeros of the numerator polynomial using the (high precision variant of the) Newton-Maehly method. These zeros correspond to the zeros of the rational functional $\mathcal{R}(z)$ and are taken as approximate eigenvalues of the matrix S of the second group. The main idea of this approach is to evaluate these zeros from the analytical continuation of $\mathcal{R}(z)$ using values for $|z| = 1$ to determine its zeros well inside the unit circle.

We also consider a second variant of the method where the number of support points $n_S = 2n_R + 2$ is even (instead of $n_S = 2n_R + 1$ being odd as for the first variant). In this case the numerator polynomial is of degree $n_R + 1$ (instead of n_R) while the denominator polynomial is of degree n_R and we choose to interpolate the inverse of the rational function $1/\mathcal{R}(z)$ (instead of $\mathcal{R}(z)$ itself) by $R_I(z)$ such that the zeros of $\mathcal{R}(z)$ are given by the n_R zeros of the denominator (instead of the numerator) polynomial of $R_I(z)$.

The number n_R must not be too small in order to well approximate the second identity in (15) by the fit function. On the other hand for a given precision of p binary digits the number of n_R must not be too large as well because the coefficients c_j, which may be written as the expansion $c_j = \sum_\nu C_\nu\, \rho_\nu^j$, do not contain enough information to resolve its structure for the smaller eigenvalues ρ_j of S_0. Therefore for too large values of n_R (for a given precision), we obtain additional artificial zeros of the numerator polynomial (or of the denominator polynomial for the second variant) of $R_I(z)$, mostly close to the unit circle, somehow as additional nodes around the support points.

It turns out that for the proper combination of p and n_R values the method provides highly accurate eigenvalues and works astonishingly well. In particular for values of n_R below a certain threshold (depending on the precision p) both variants of the method with odd or even number of support points provide numerically identical zeros (with final results rounded to 52 binary digits) which indeed coincide very accurately (for most of them) with the eigenvalues of S we want to determine.

For example, as can be seen in Fig. 9, for $p = 1024$ we obtain $n_R = 300$ eigenvalues for which the big majority coincides numerically (error $\sim 10^{-14}$) with the eigenvalues obtained from the high precision Arnoldi method for 768 binary digits and furthermore both variants of the rational interpolation method provide identical spectra.

However for $n_R = 340$ some of the zeros do not coincide with eigenvalues of S and most of these deviating zeros lie close to the unit circle. We can even somehow distinguish between "good" zeros (associated to eigenvalues of S) being identical for both variants of the method and "bad" artificial zeros which are completely different for both variants (see Fig. 9). We note that for the case of too large n_R values the artificial zeros are extremely sensitive to numerical round-off errors (in the high precision variables) and that they change strongly, when slightly modifying the support points (e.g. a random modification $\sim 10^{-18}$ or simply changing their order in the interpolation scheme) or when changing the precise numerical algorithm (e.g. between direct sum or Horner scheme for the evaluation of the series of the rational function). Furthermore, they do not respect the symmetry that the zeros should come in pairs of complex conjugate numbers in case of complex zeros. This is because Thiele's rational interpolation scheme breaks the symmetry due to complex conjugation once round-off errors become relevant.

However, we have carefully verified that for the proper values of n_R not being too large (e.g. $n_R = 300$ for $p = 1024$) the obtained zeros are numerically identical (with 52 binary digits in the final result) with respect to small changes of the support points (or their order) or with respect to different numerical algorithms and that they respect perfectly the symmetry due to complex conjugation.

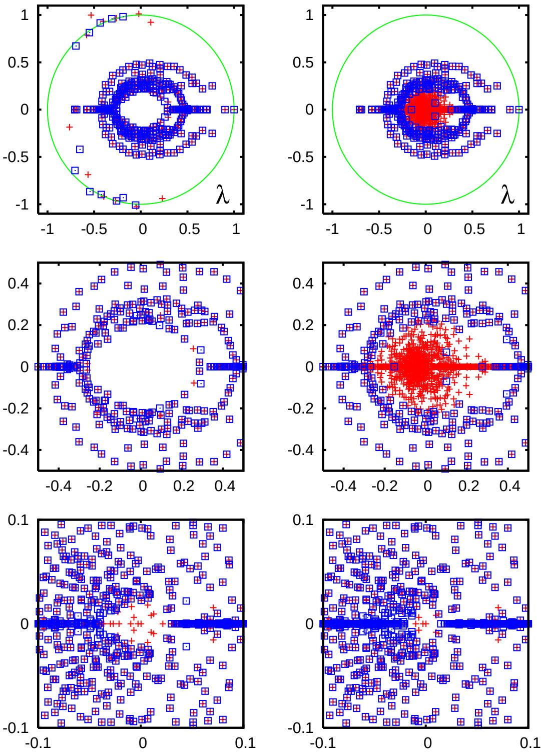

FIG. 9: (Color online) *Top Panels: Left:* Comparison of $n_R = 340$ core space eigenvalues of S for CNPR obtained by two variants of the rational interpolation method (see text) with the numerical precision of $p = 1024$ binary digits, 681 support points (first variant, red crosses) or 682 support points (second variant, blue squares). *Right:* Comparison of the core space eigenvalues of CNPR obtained by the high precision Arnoldi method with $n_A = 2000$ and $p = 768$ binary digits (red crosses, same data as blue squares in Fig. 7) with the eigenvalues obtained by (both variants of) the rational interpolation method with the numerical precision of $p = 1024$ binary digits and $n_R = 300$ eigenvalues (blue squares). Here both variants with 601 or 602 support points provide identical spectra (differences below 10^{-14}). *Middle panels:* Same as top panels with a zoomed range: $-0.5 \leq \mathrm{Re}(\lambda), \mathrm{Im}(\lambda) \leq 0.5$. *Bottom Panels: Left:* Comparison of the core space spectra obtained by the high precision Arnoldi method (red crosses, $n_A = 2000$ and $p = 768$) and by the rational interpolation method with $p = 12288$, $n_R = 2000$ eigenvalues (blue squares). *Right:* Same as left panel with $p = 16384$, $n_R = 2500$ for the rational interpolation method. Both panels are shown in a zoomed range: $-0.1 \leq \mathrm{Re}(\lambda), \mathrm{Im}(\lambda) \leq 0.1$. Eigenvalues outside the shown range coincide up to graphical precision and both variants of the rational interpolation method provide numerically identical spectra.

This method, despite the necessity of high precision calculations, is not very expensive, especially for the memory usage, compared, for example, with the high precision Arnoldi method. Furthermore, its efficiency for the computation time can be improved by the trick of summing up the largest terms in the series (17) as a geometrical series which allows to reduce the cutoff value of l by a good factor 3, i.e. replacing $\rho_1 \approx 0.902$ by $\rho_2 = 1/\sqrt{2} \approx 0.707$ in the estimate (18) of l which gives $l \approx 2\,p + \mathrm{const.}$ We have increased the number of binary digits up to $p = 16384$ and we find that for $p = 1024, 2048, 4096, 6144, 8192, 12288, 16384$ we may use $n_R = 300, 500, 900, 1200, 1500, 2000, 2500$ and still avoid the appearance of artificial zeros. In Fig. 9 we also compare the result of the highest precisions $p = 12288$ (and $p = 16384$) using $n_R = 2000$ ($n_R = 2500$) with the high precision Arnoldi method with $n_A = 2000$ and $p = 768$ and these spectra coincide well apart from a minor number of smallest eigenvalues. In general, the complex isolated eigenvalues converge very well (with increasing values of p and n_R) while the strongly clustered eigenvalues on the real axis have more difficulties to converge. Comparing the results between $n_R = 2000$ and $n_R = 2500$ we see that the complex eigenvalues coincide on graphical precision for $|\lambda| \geq 0.04$ and the real eigenvalues for $|\lambda| \geq 0.1$. The Arnoldi method has even more difficulties on the real axis (convergence roughly for $|\lambda| \geq 0.15$) since it has implicitly to take care of the highly degenerate eigenvalues of the first group and for which it has difficulties to correctly find the degeneracies (see also Fig. 8).

Fig. 10 shows as summary the highest precision spectra of S with core space eigenvalues obtained by the Arnoldi method or the rational interpolation method (both at best parameter choices) and also taking into account the direct subspace eigenvalues of S and the above determined eigenvalues of the first group (degenerate subspace eigenvalues of S_0).

We remind that the rational interpolation method allows only to determine the eigenvalues of S of the second group, i.e. the eigenvalues which are not eigenvalues of S_0 and whose eigenvectors obey $d^T\psi \neq 0$. The eigenvalues of the first group (with $d^T\psi = 0$) have to be determined separately by the above described scheme of degenerate subspace eigenvalues of S_0. In particular the eigenvalues given in Table I and belonging to the first group are not zeros of the rational function $\mathcal{R}(z)$ (they are actually poles of this function) but it turns out that there are some zeros of $\mathcal{R}(z)$ which are very close but not identical to some of the values in Table I. For example the rational interpolation method provides the following zeros: $1/2 + 3.13401098 \times 10^{-5}$, $1/2 + 1.3279300 \times 10^{-7}$, $1/\sqrt{2} - 1.1597 \times 10^{-10}$ or $1/\sqrt{2} - 6.419004 \times 10^{-8}$ which are indeed accurate in the given precision since they are stable for all values of $p \geq 1024$ and the corresponding maximal value of n_R and we have stopped the Newton iteration when the error of a zero was clearly below 10^{-18}. These zeros are also found with the same precision in the data of the high precision Arnoldi method for the three different values of 256, 512 or 768 binary digits. However, based only on results of the Arnoldi method it is not really clear if the small corrections to $1/2$ or $1/\sqrt{2}$ are real and exact or numerically artificial since the Arnoldi method has indeed problems with degenerate and clustered eigenvalues [17]. Therefore the rational interpolation method provides an independent and strong confirmation of the accuracy of these type of eigenvalues. We attribute their existence to a quasi-subspace structure, similarly as discussed in [10], with a matrix subblock as in (14) but which is still very weakly coupled (by many indirect network links) to the core space.

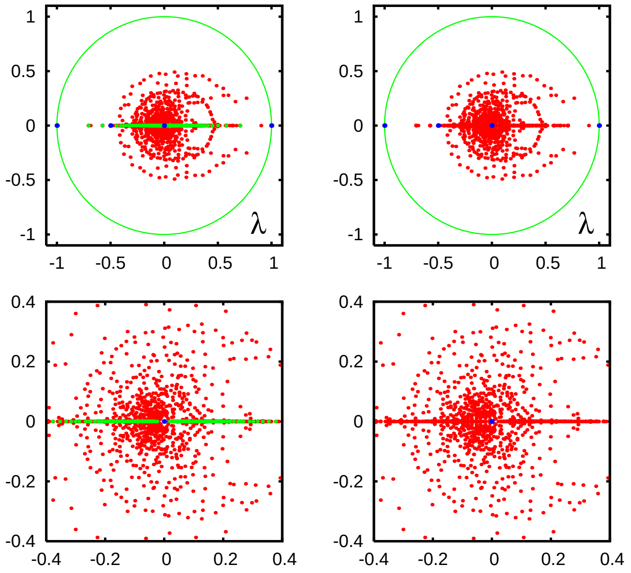

FIG. 10: (Color online) The most accurate spectrum of eigenvalues of S for CNPR. *Top panels: Left:* red dots represent the core space eigenvalues obtained by the rational interpolation method with the numerical precision of $p = 16384$ binary digits, $n_R = 2500$ eigenvalues. Green dots show the degenerate subspace eigenvalues of the matrix S_0 which are also eigenvalues of S with a degeneracy reduced by one (eigenvalues of the first group, see text). Blue dots show the direct subspace eigenvalues of S (same as blue dots in left upper panel in Fig. 3). *Right:* red dots represent the core space eigenvalues obtained by the high precision Arnoldi method with $n_A = 2000$ and the numerical precision of $p = 768$ binary digits and blue dots show the direct subspace eigenvalues of S. Note that the Arnoldi method determines implicitly also the degenerate subspace eigenvalues of S_0 which are therefore not shown in another color. *Bottom panels*: Same as top panels with a zoomed range: $-0.4 \leq \mathrm{Re}(\lambda), \mathrm{Im}(\lambda) \leq 0.4$.

III. FRACTAL WEYL LAW FOR CNPR

The concept of the fractal Weyl law [32, 33],[34] states that the number of states N_λ in a ring of complex eigenvalues with $\lambda_c \leq |\lambda| \leq 1$ scales in a polynomial way with the growth of matrix size:

$$N_\lambda = aN^b \,. \tag{19}$$

where the exponent b is related to the fractal dimension of underlying invariant set $d_f = 2b$. The fractal Weyl law was first discussed for the problems of quantum chaotic scattering in the semiclassical limit [32, 33],[34]. Later it was shown that this law also works for the Ulam matrix approximant of the Perron-Frobenius operators of dissipative chaotic systems with strange attractors [6, 7]. In [11] it was established that the time growing Linux Kernel network is also characterized by the fractal Weyl law with the fractal dimension $d_f \approx 1.3$.

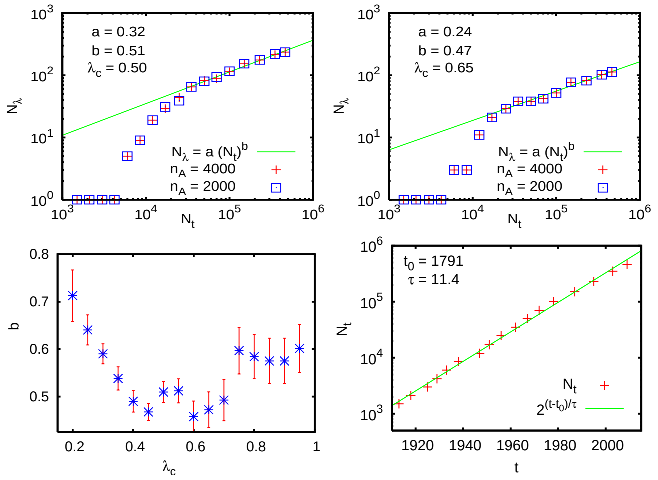

FIG. 11: (Color online) Data for the whole CNPR at different moments of time. *Top panels:* the left (right) panel shows the number N_λ of eigenvalues with $\lambda_c \leq \lambda \leq 1$ for $\lambda_c = 0.50$ ($\lambda_c = 0.65$) versus the effective network size N_t where the nodes with publication times after a cut time t are removed from the network. The green line shows the Weyl law $N_\lambda = a\,(N_t)^b$ with parameters $a = 0.32 \pm 0.08$ ($a = 0.24 \pm 0.11$) and $b = 0.51 \pm 0.02$ ($b = 0.47 \pm 0.04$) obtained from a fit in the range $3 \times 10^4 \leq N_t < 5 \times 10^5$. The number N_λ includes both exactly determined invariant subspace eigenvalues and core space eigenvalues obtained from the Arnoldi method with double-precision (52 binary digits) for $n_A = 4000$ (red crosses) and $n_A = 2000$ (blue squares). *Bottom panels: Left:* exponent b with error bars obtained from the fit $N_\lambda = a\,(N_t)^b$ in the range $3 \times 10^4 \leq N_t < 5 \times 10^5$ versus cut value λ_c. *Right:* effective network size N_t versus cut time t (in years). The green line shows the exponential fit $2^{(t-t_0)/\tau}$ with $t_0 = 1791 \pm 3$ and $\tau = 11.4 \pm 0.2$ representing the number of years after which the size of the network (number of papers published in all Physical Review journals) is effectively doubled.

The fact that $b < 1$ implies that the majority of eigenvalues drop to zero. We see that this property also appears for the CNPR if we test here the validity of the fractal Weyl law by considering a time reduced CNPR of size N_t including the N_t papers published until the time t (measured in years) for different times t in order to obtain a scaling behavior of N_λ as a function of N_t. The data presented in Fig. 11 shows that the network size grows approximately exponentially as $N_t = 2^{(t-t_0)/\tau}$ with the fit parameters $t_0 = 1791$, $\tau = 11.4$. The time interval considered in Fig. 11 is $1913 \leq t \leq 2009$ since the first data point corresponds to $t = 1913$ with $N_t = 1500$ papers published between 1893 and 1913. The results for N_λ show that its growth is well described by the relation $N_\lambda = a\,(N_t)^b$ for the range when the number of articles becomes sufficiently large $3 \times 10^4 \leq N_t < 5 \times 10^5$. This range is not very large and probably due to that there is a certain dependence of the exponent b on the range parameter λ_c. However, we have $0.47 < b < 0.6$ for all $\lambda_c \geq 0.4$ that is definitely smaller than unity and thus the fractal Weyl law is well applicable to the CNPR. The value of b increases up to 0.7 for the data points with $\lambda_c < 0.4$ but this is due to the fact here N_λ also includes some numerically incorrect eigenvalues related to the numerical instability of the Arnoldi method at standard double-precision (52 binary digits) as discussed in the beginning of the previous section.

We think that the most appropriate choice for the description of the data is obtained at $\lambda_c = 0.4$ which from one side excludes small, partly numerically incorrect, values of λ and on the other side gives sufficiently large values of N_λ. Here we have $b = 0.49 \pm 02$ corresponding to the fractal dimension $d = 0.98 \pm 0.04$. Furthermore, for $0.4 \leq \lambda_c \leq 0.7$ we have a rather constant value $b \approx 0.5$ with $d_f \approx 1.0$. Of course, it would be interesting to extend this analysis to a larger size N of CNPR but for that we still should wait about 10 years until the network size will be doubled comparing to the size studied here.

IV. PROPERTIES OF EIGENVECTORS

The results for the eigenvalue spectra of CNPR presented in the previous sections show that most of the visible eigenvalues on the real axis (except for the largest one) in Figs. 9 and 10 are due to the effect of future citations. They appear either directly due to 2×2 subblocks of the type (14) with a cycle where two papers mutually cite each other giving the degenerate eigenvalues of the first group, or indirectly by eigenvalues of the second group which are also numerous on the real axis. On the other hand, as can be seen in Fig. 6, for the triangular CNPR, where all future citations are removed, there is only the leading eigenvalue $\lambda = 1$ and a small number of negative eigenvalues with $-0.27 < \lambda < 0$ on the real axis. All other eigenvalues are complex and a considerable number of the largest ones are relatively close to corresponding complex eigenvalues for the whole CNPR with future citations.

The appearance of future citations is quite specific and is not a typical situation for citation networks. Therefore we consider the eigenvectors of complex eigenvalues for the triangular CNPR which indeed represent the typical physical situation without future citations. There is no problem to evaluate these eigenvectors by the Arnoldi method, either with double-precision, provided the eigenvalue of the eigenvector is situated in the region of numerically accurate eigenvalues, or with the high precision variant of the Arnoldi method. However, for the triangular CNPR we have, according to the semi-analytical theory presented above, the explicit formula:

$$\psi \propto (\lambda \mathbb{1} - S_0)^{-1}\, e/N = \sum_{j=0}^{l-1} \lambda^{-(1+j)}\, S_0^j e/N \qquad (20)$$

where the normalization is given by $\sum_i |\psi(i)| = 1$. This expression is quite convenient and we verified that it provides the same eigenvectors (up to numerical errors) as the Arnoldi method.

In Fig. 12 we show two eigenvectors of S: one ψ_0 for the leading eigenvalue $\lambda_0 = 1$ and another ψ_{39} for a complex eigenvalue at $|\lambda_{39}| < 1$. The eigenvector of λ_0 gives the PageRank probability for the triangular CNPR (at $\alpha = 1$). We also consider the eigenvector for the complex eigenvalue $\lambda_{39} = -0.3738799 + i\,0.2623941$ (eigenvalues are ordered by their absolute values starting from $\lambda_0 = 1$). In this figure the modulus of $|\psi_j(N_t)|$ is shown versus the time index N_t as introduced in Fig. 11. We also indicate the positions of five famous papers: BCS 1957 [35] at $K = 6$, Anderson 1958 [36] $K = 63$, Benettin et al. 1976 [37] $K = 441$, Thouless 1977 [38] $K = 256$ and Abrahams et al. 1979 [39] $K = 74$. In the first eigenvector for $\lambda_0 = 1$ all of these papers have quite dominating positions, especially BCS 1957 and Abrahams et al. 1979 which are the most important ones if compared to papers of comparable publication date. Only considerably older papers have higher positions in this vector.

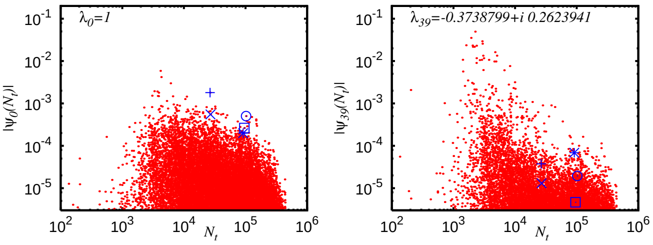

FIG. 12: (Color online) Two eigenvectors of the matrix S for the triangular CNPR. Both panels show the modulus of the eigenvector components $|\psi_j(N_t)|$ versus the time index N_t (as used in Fig. 11) with nodes/articles ordered by the publication time (small red dots). The blue points represent five particular articles: BCS 1957 (+), Anderson 1958 (×), Benettin et al. 1976 (∗), Thouless 1977 (⊡) and Abrahams et al. 1979 (⊙). The left (right) panel corresponds to the real (complex) eigenvalue $\lambda_0 = 1$ ($\lambda_{39} = -0.3738799 + i\,0.2623941$).

For the second eigenvector with complex eigenvalue the older papers (with $10^3 < N_t < 10^4$ corresponding to publications times between 1910 and 1940) are strongly enhanced in its importance while the above five famous papers lose their importance. The top 3 positions of largest amplitude $|\psi_{39}(i)|$ correspond to DOI 10.1103/PhysRev.14.409 (1919), 10.1103/PhysRev.8.561 (1916), 10.1103/PhysRev.24.97 (1917). These old articles study the radiating potentials of nitrogen, ionization impact in gases and the abnormal low voltage arc. It is clear that this eigenvector selects a certain community of old articles related to a certain ancient field of interest. This fact is in agreement with the studies of eigenvectors of Wikipedia network [13] showing that the eigenvectors with $0 < |\lambda| < 1$ select specific communities.

It is interesting to note that the top node of the vector ψ_0 appears in the position $K_{39} = 39$ in local rank index of the vector ψ_{39} (ranking in decreasing order by modulus of $|\psi(i)|$). On the other side the top node of ψ_{39} appears at position $K_0 = 30$ of vector ψ_0. This illustrates how different nodes contribute to different eigenvectors of S.

It is useful to characterize the eigenvectors by their Inverse Participation Ratio (IPR) $\xi_i = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$ which gives an effective number of nodes populated by an eigenvector ψ_i (see e.g. [8, 13]). For the above two vectors we find $\xi_0 = 20.67$ and $\xi_{39} = 10.76$. This means that ξ_{39} is mainly located on approximately 11 nodes. For ξ_0 this number is twice larger in agreement with data of Fig. 12 which show a clearly broader distribution comparing to ξ_{39}.

We also considered a few tens of eigenstates of S of the whole CNPR. They are mainly located on the complex plane around the largest oval curve well visible in the spectrum (see Fig. 10 top right panel). The IPR value of these eigenstates with $|\lambda| \sim 0.4$ varies in the range $4 < \xi < 13$ showing that they are located on some effective quasi-isolated communities of articles. About 10 of them are related to the top article of ψ_{39} shown in Fig. 12 meaning that these ten vectors represent various linear combinations of vectors on practically the same community. In global, we can say that the eigenstates of G are well localized since $\xi \ll N$. A similar situation was seen for the Wikipedia network [13].

Of course, in addition to ξ it is also useful to consider the whole distribution of ψ amplitudes over the nodes. Such a consideration has been done for the Wikipedia network in [13]. For the CNPR we leave such detailed studies for further investigations.

V. CHEIRANK VERSUS PAGERANK FOR CNPR

The dependence of PageRank probability $P(K)$ on PageRank index K is shown in Fig. 13. The results are similar to those of [22]. We note that the PageRank of the triangular CNPR has the same top 9 articles as for the whole CNPR (both at $\alpha = 0.85$ and with a slight interchanged order of positions 7, 8, 9). This confirms that the future citations produce only a small effect on the global ranking.

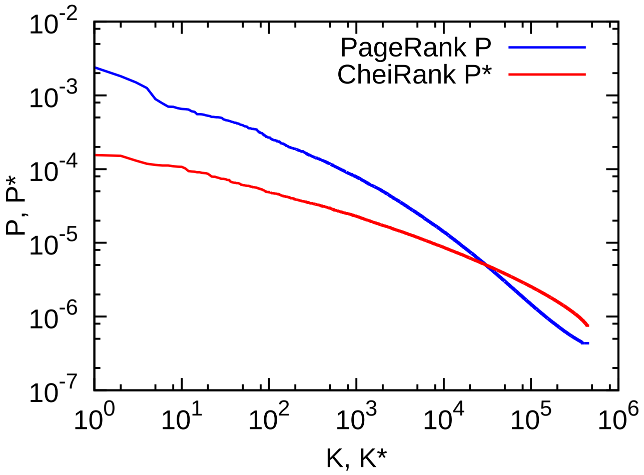

FIG. 13: (Color online) Dependence of probability of PageRank P (CheiRank P^*) on corresponding index K (K^*) for the CNPR at $\alpha = 0.85$.

Following previous studies [24],[25, 26], in addition to the Google matrix G we also construct the matrix G^* following the same definition (1) but for the network with inverted direction of links. The PageRank vector of this matrix G^* is called the CheiRank vector with probability $P^*(K_i^*)$ and CheiRank index K^*. The dependence of $P^*(K_i^*)$ is shown in Fig. 13. We find that the IPR values of P and P^* are $\xi = 59.54$ and 1466.7 respectively. Thus P^* is extended over significantly larger number of nodes comparing to P. A power law fit of the decay $P \propto 1/K^\beta$, $P^* \propto 1/K^{*\beta}$, done for a range $K, K^* \leq 2 \times 10^5$ gives $\beta \approx 0.57$ for P and $\beta \approx 0.4$ for P^*. However, this is only an approximate description since there is a visible curvature (in a double logarithmic representation) in these distributions. The corresponding frequency distributions of ingoing links have exponents $\mu = 2.87$ while the distribution of outgoing links has $\mu \approx 3.7$ for outdegree $k \geq 20$, even if the whole frequency dependence in this case is rather curved and a power law fit is rather approximate in this case. Thus the usual relation $\beta = 1/(\mu - 1)$ [4, 8, 25] approximately works.

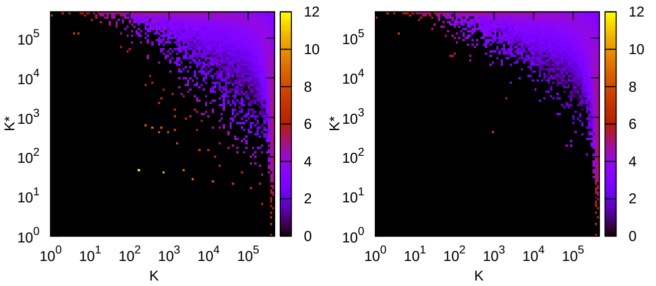

FIG. 14: (Color online) Density distribution $W(K, K^*) = dN_i/dKdK^*$ of Physical Review articles in the PageRank-CheiRank plane (K, K^*). Color bars show the natural logarithm of density, changing from minimal nonzero density (dark) to maximal one (white), zero density is shown by black. Left panel: all articles of CNPR; right panel: CNPR without Rev. Mod. Phys.

The correlation between PageRank and CheiRank vectors can be characterized by the correlator $\kappa = N \sum_{i=1}^{N} P(i)P^*(i) - 1$ [24, 26]. Here we find $\kappa = -0.2789$ for all CNPR, and $\kappa = -0.3187$ for CNPR without Rev. Mod. Phys. This is the most strong negative value of κ among all directed networks studied previously [26]. In a certain sense the situation is somewhat similar to the Linux Kernel network where $\kappa \approx 0$ or slightly negative ($\kappa > -0.1$ [24]). For CNPR, we can say that due to a almost triangular structure of G and G^* there is a very little overlap of top ranking in K and K^* that leads to a negative correlator value, since the components $P(i)P^*(i)$ of the sum for κ are small.

Each article i has two indexes K_i, K_i^* so that it is convenient to see their distribution on 2D PageRank-CheiRank plane. The density distribution $W(K, K^*) = dN_i/dKdK^*$ is shown in Fig. 14. It is obtained from 100×100 cells equidistant in log-scale (see details in [25, 26]). For the CNPR the density is homogeneous along lines $K = -K^* + const$ that corresponds to the absence of correlations between P and P^* [25, 26]. For the CNPR without Rev. Mod. Phys. we have an additional suppression of density at low K^* values. Indeed, Rev. Mod. Phys. contains mainly review articles with a large number of citations that place them on top of CheiRank. At the top 3 positions of K^* of CNPR we have DOI 10.1103/PhysRevA.79.062512, 10.1103/PhysRevA.79.062511, 10.1103/RevModPhys.81.1551 of 2009. These are articles with long citation lists on K shell diagram 4d transition elements; hypersatellites of 3d transition metals; superconducting phases of f electron compounds. For CNPR without Rev. Mod. Phys. the first two articles are the same and the third one has DOI 10.1103/PhysRevB.80.224501 being about model for the coexistence of d wave superconducting and charge density wave order in in high temperature cuprate superconductors. We see that the most recent articles with long citation lists are dominating.

The top PageRank articles are analyzed in detail in [22] and we do not discuss them here.

It is also useful to consider two-dimensional rank 2DRank K_2 defined by counting nodes in order of their appearance on ribs of squares in (K, K^*) plane with the square size growing from $K = 1$ to $K = N$ [25]. It selects highly cited articles with a relatively long citation list. For CNPR, we have top 3 such articles with DOI 10.1103/RevModPhys.54.437 (1982), 10.1103/RevModPhys.65.851 (1993), 10.1103/RevModPhys.58.801 (1986). Their topics are electronic properties of two dimensional systems, pattern formation outside of equilibrium, spin glasses facts and concepts. The 1st one located at $K = 183$, $K^* = 49$ is well visible in the left panel of Fig. 14. For CNPR without Rev. Mod. Phys. we find at $K_2 = 1$ the article with DOI 10.1103/PhysRevD.54.1 (1996) entitled *Review of Particle Physics* with a lot of information on physical constants.

For the ranking of articles about persons in Wikipedia networks [14, 25],[40], PageRank, 2DRank, CheiRank highlights in a different manner various sides of human activity. For the CNPR, these 3 ranks also select different types of articles, however, due a triangular structure of G, G^* and absence of correlations between PageRank and CheiRank vectors the useful side of 2DRank and CheiRank remains less evident.

VI. IMPACTRANK FOR INFLUENCE PROPAGATION

It is interesting to quantify how an influence of a given article propagates through the whole CNPR. To analyze this property we consider the following propagator acting on an initial vector v_0 located on a given article:

$$v_f = \frac{1-\gamma}{1-\gamma G} v_0 \quad , \quad v_f^* = \frac{1-\gamma}{1-\gamma G^*} v_0 \ . \qquad (21)$$

Here G, G^* are the Google matrices defined above, γ is a new impact damping factor being in a range $\gamma \sim 0.5 - 0.9$, v_f in the final vector generated by the propagator (21). This vector is normalized to unity $\sum_i v_f(i) = 1$ and one can easily show that it is equal to the PageRank vector of a modified Google matrix given by

$$\tilde{G} = \gamma \, G + (1-\gamma) \, v_0 \, e^T \qquad (22)$$

where e is the vector with unit elements. This modified Google matrix corresponds to a stochastic process where at a certain time a given probability distribution is propagated with probability γ using the initial Google matrix G and with probability $(1-\gamma)$ the probability distribution is reinitialized with the vector v_0. Then v_f is the stationary vector from this stochastic process. Since the initial Google matrix G has a similar form, $G = \alpha S + (1-\alpha) e \, e^T/N$ with the damping factor α, the modified Google matrix can also be written as:

$$\tilde{G} = \tilde{\alpha} \, S + (1-\tilde{\alpha}) \, v_p \, e^T \quad , \quad \tilde{\alpha} = \gamma\alpha \quad , \qquad (23)$$

with the personalization vector [4]

$$v_p = \frac{\gamma(1-\alpha)e/N + (1-\gamma)v_0}{1-\gamma\alpha} \tag{24}$$

which is also sum normalized: $\sum_i v_p(i) = 1$. Obviously similar relations hold for G^* and v_f^*.

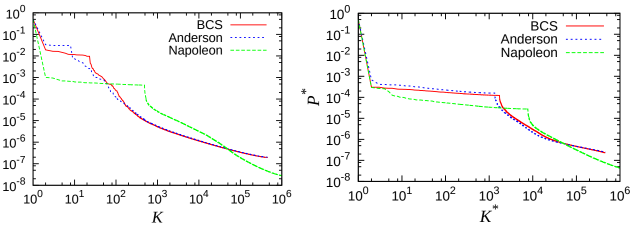

FIG. 15: (Color online) Dependence of impact vector v_f probability P and P^* (left and right panels) on the corresponding ImpactRank index K and K^* for an initial article v_0 as BCS [35] and Anderson [36] in CNPR, and Napoleon in English Wikipedia network from [40]. Here the impact damping factor is $\gamma = 0.5$.

The relation (21) can be viewed as a Green function with damping γ. Since $\gamma < 1$ the expansion in a geometric series is convergent and v_f can be obtained from about 200 terms of the expansion for $\gamma \sim 0.5$. The stability of v_f is verified by changing the number of terms. The obtained vectors v_f, v_f^* can be considered as effective PageRank, CheiRank probabilities P, P^* and all nodes can be ordered in the corresponding rank index K, K^*, which we will call ImpactRank.

The results for 2 initial vectors located on BCS [35] and Anderson [36] articles are shown in Fig. 15. In addition we show the same probability for the Wikipedia article *Napoleon* for the English Wikipedia network analyzed in [40]. The direct analysis of the distributions shows that the original article is located at the top position, the next step like structure corresponds to the articles reached by first outgoing (ingoing) links from v_0 for G (G^*). The next visible step correspond to a second link step.

Top ten articles for these 3 vectors are shown in Tables II, III, IV, V, VI. The analysis of these top articles confirms that they are closely linked with the initial article and thus the ImpactRank gives relatively good ranking results. At the same time, some questions for such ImpactRanking still remain to be clarified. For example, in Table V we find at the third position the well known Rev. Mod. Phys. on Anderson transitions but the paper of Abrahams *et al.* [39] appears only on far positions $K^* \approx 300$. The situation is changed if we consider all CNPR links as bi-directional obtaining a non-directional network. Then the paper [39] appears on the second position directly after initial article [36]. We think that such a problem appears due to triangular structure of CNPR where there is no intersection of forward and backward flows. Indeed, for the case of Napoleon we do not see such difficulties. Thus we hope that such an approach can be applied to other directed networks.

VII. MODELS OF RANDOM PERRON-FROBENIUS MATRICES

In this section we discuss the spectral properties of several random matrix models of Perron-Frobenius operators characterized by non-negative matrix elements and column sums normalized to unity. We call these models Random Perron-Frobenius Matrices (RPFM). To construct these models for a given matrix G of dimension N we draw N^2 independent matrix elements $G_{ij} \geq 0$ from a given distribution $p(G)$ (with $p(G) = 0$ for $G < 0$) with average $\langle G \rangle = 1/N$ and finite variance $\sigma^2 = \langle G^2 \rangle - \langle G \rangle^2$. A matrix obtained in this way obeys the column sum normalization only in average but not exactly for an arbitrary realization. Therefore we renormalize all columns to unity after having drawn the matrix elements. This renormalization provides some (hopefully small) correlations between the different matrix elements.

Neglecting these correlations for sufficiently large N the statistical average of the RPFM is simply given by $\langle G_{ij} \rangle = 1/N$ which is a projector matrix with the eigenvalue $\lambda = 1$ of multiplicity 1 and the corresponding eigenvector being the uniform vector e (with $e_i = 1$ for all i). The other eigenvalue $\lambda = 0$ is highly degenerate of multiplicity $N - 1$ and its eigenspace contains all vectors orthogonal to the uniform vector e. Writing the matrix elements of a RPFM as $G_{ij} = \langle G_{ij} \rangle + \delta G_{ij}$ we may consider the fluctuating part δG_{ij} as a perturbation which only weakly modifies the unperturbed eigenvector e for $\lambda = 1$ but for the eigenvalue $\lambda = 0$ we have to apply degenerate perturbation theory which requires the diagonalization of δG_{ij}. According to the theory of non-symmetric real random Gaussian matrices [5, 41, 42] it is well established that the complex eigenvalue density of such a matrix is uniform on a circle of radius $R = \sqrt{N}\sigma$ with σ^2 being the variance of the matrix elements. One can also expect that this holds for more general, non-Gaussian, distributions with finite variance provided that we exclude extreme long tail distribution where the typical values are much smaller than σ. Therefore we expect that the eigenvalue density of a RPFM is determined by a single parameter being the variance σ^2 of the matrix elements resulting in a uniform density on a circle of radius $R = \sqrt{N}\sigma$ around $\lambda = 0$, in addition to the unit eigenvalue $\lambda = 1$ which is always an exact eigenvalue due to sum normalization of columns.

We now consider different variants of RPFM. The first variant is a full matrix with each element uniformly distributed in the interval $[0, 2/N[$ which gives the variance $\sigma^2 = 1/(3N^2)$ and the spectral radius $R = 1/\sqrt{3N}$. The second variant is a sparse RPFM matrix with Q nonvanishing elements per column and which are uniformly distributed in the interval $[0, 2/Q[$. Then the probability distribution is given by $p(G) = (1 - Q/N)\delta(G) +$

TABLE II: Spreading of impact on "Theory of superconductivity" paper by "J. Bardeen, L. N. Cooper and J. R. Schrieffer (doi:10.1103/PhysRev.108.1175) by Google matrix G with $\alpha = 0.85$ and $\gamma = 0.5$

ImpactRank	DOI	Title of paper
1	10.1103/PhysRev.108.1175	Theory of superconductivity
2	10.1103/PhysRev.78.477	Isotope effect in the superconductivity of mercury
3	10.1103/PhysRev.100.1215	Superconductivity at millimeter wave frequencies
4	10.1103/PhysRev.78.487	Superconductivity of isotopes of mercury
5	10.1103/PhysRev.79.845	Theory of the superconducting state. i. the ground ...
6	10.1103/PhysRev.80.567	Wave functions for superconducting electrons
7	10.1103/PhysRev.79.167	The hyperfine structure of ni^{61}
8	10.1103/PhysRev.97.1724	Theory of the Meissner effect in superconductors
9	10.1103/PhysRev.81.829	Relation between lattice vibration and London ...
10	10.1103/PhysRev.104.844	Transmission of superconducting films ...

TABLE III: Spreading of impact on "Absence of diffusion in certain random lattices" paper by P. W. Anderson (doi:10.1103/PhysRev.109.1492) by Google matrix G. with $\alpha = 0.85$ and $\gamma = 0.5$

ImpactRank	DOI	Title of paper
1	10.1103/PhysRev.109.1492	Absence of diffusion in certain random lattices
2	10.1103/PhysRev.91.1071	Electronic structure of f centers: saturation of ...
3	10.1103/RevModPhys.15.1	Stochastic problems in physics and astronomy
4	10.1103/PhysRev.108.590	Quantum theory of electrical transport phenomena
5	10.1103/PhysRev.48.755	Theory of pressure effects of foreign gases on spectral lines
6	10.1103/PhysRev.105.1388	Multiple scattering by quantum-mechanical systems
7	10.1103/PhysRev.104.584	Spectral diffusion in magnetic resonance
8	10.1103/PhysRev.74.206	A note on perturbation theory
9	10.1103/PhysRev.70.460	Nuclear induction
10	10.1103/PhysRev.90.238	Dipolar broadening of magnetic resonance lines ...

TABLE IV: Spreading of impact on "Theory of superconductivity" paper by "J. Bardeen, L. N. Cooper and J. R. Schrieffer (doi:10.1103/PhysRev.108.1175) by Google matrix G^* with $\alpha = 0.85$ and $\gamma = 0.5$

ImpactRank	DOI	Title of paper
1	10.1103/PhysRev.108.1175	Theory of superconductivity
2	10.1103/PhysRevB.77.104510	Temperature-dependent gap edge in strong-coupling ...
3	10.1103/PhysRevC.79.054328	Exact and approximate ensemble treatments of thermal ...
4	10.1103/PhysRevB.8.4175	Ultrasonic attenuation in superconducting molybdenum
5	10.1103/RevModPhys.62.1027	Properties of boson-exchange superconductors
6	10.1103/PhysRev.188.737	Transmission of far-infrared radiation through thin films ...
7	10.1103/PhysRev.167.361	Superconducting thin film in a magnetic field - theory of ...
8	10.1103/PhysRevB.77.064503	Exact mesoscopic correlation functions of the Richardson ...
9	10.1103/PhysRevB.10.1916	Magnetic field attenuation by thin superconducting lead films
10	10.1103/PhysRevB.79.180501	Exactly solvable pairing model for superconductors with ...

$(Q/N)\,\chi_{[0,2/Q[}(G)$ where $\chi_{[0,2/Q[}(G)$ is the characteristic function on the interval $[0,2/Q[$ (with values being 1 for G in this interval and 0 for G outside this interval). The average is indeed $\langle G\rangle = 1/N$ and the vari-

TABLE V: Spreading of impact on "Absence of diffusion in certain random lattices" paper by P. W. Anderson (doi:10.1103/PhysRev.109.1492) by Google matrix G^*. with $\alpha = 0.85$ and $\gamma = 0.5$

ImpactRank	DOI	Title of paper
1	10.1103/PhysRev.109.1492	Absence of diffusion in certain random lattices
2	10.1103/PhysRevA.80.053606	Effects of interaction on the diffusion of atomic ...
3	10.1103/RevModPhys.80.1355	Anderson transitions
4	10.1103/PhysRevE.79.041105	Localization-delocalization transition in hessian ...
5	10.1103/PhysRevB.79.205120	Statistics of the two-point transmission at ...
6	10.1103/PhysRevB.80.174205	Localization-delocalization transitions ...
7	10.1103/PhysRevB.80.024203	Statistics of renormalized on-site energies and ...
8	10.1103/PhysRevB.79.153104	Flat-band localization in the Anderson-Falicov-Kimball model
9	10.1103/PhysRevB.74.104201	One-dimensional disordered wires with Poschl-Teller potentials
10	10.1103/PhysRevB.71.235112	Critical wave-packet dynamics in the power-law bond ...

TABLE VI: Spreading of impact on the article of "Napoleon" in English Wikipedia by Google matrix G and G^*. with $\alpha = 0.85$ and $\gamma = 0.5$

ImpactRank	Articles (G case)	Articles (G^* case)
1	Napoleon	Napoleon
2	French Revolution	List of orders of battle
3	France	Lists of state leaders by year
4	First French Empire	Names inscribed under the Arc de Triomphe
5	Napoleonic Wars	List of battles involving France
6	French First Republic	Order of battle of the Waterloo Campaign
7	Saint Helena	Napoleonic Wars
8	French Consulate	Wagram order of battle
9	French Directory	Departments of France
10	National Convention	Jena-Auerstedt Campaign Order of Battle

ance is $\sigma^2 = 4/(3NQ)$ (for $N \gg Q$) providing the spectral radius $R = 2/\sqrt{3Q}$. We may also consider a sparse RPFM where we have exactly Q non-vanishing constant elements of value $1/Q$ in each column with random positions resulting in a variance $\sigma^2 = 1/(NQ)$ and $R = 1/\sqrt{Q}$. The theoretical predictions for these three variants of RPFM coincide very well with numerical simulations. In Fig. 16 the complex eigenvalue spectrum for one realization of each of the three cases is shown for $N = 400$ and $Q = 20$ clearly confirming the circular uniform eigenvalue density with the theoretical values of R. We also confirm numerically the scaling behavior of R as a function of N or Q.

Motivated by the Google matrices of DNA sequences [43], where the matrix elements are distributed with a power law, we also considered a power law variant of RPFM with $p(G) = D(1 + aG)^{-b}$ for $0 \le G \le 1$ and with an exponent $2 < b < 3$. The condition $G \le 1$ is required because of the column sum normalization. The parameters D and a are determined by normalization and the average $\langle G\rangle = 1/N$. In the limit $N^{b-2} \gg 1$ we find $a \approx N/(b-2)$ and $D \approx N(b-1)/(b-2)$. For $b > 3$ the variance would scale with $\sim N^{-2}$ resulting in $R \sim 1/\sqrt{N}$ as in the first variant with uniformly distributed matrix elements. However, for $b < 3$ this scaling is different and we find (for $N^{b-2} \gg 1$) :

$$R = C(b)\, N^{1-b/2} \quad , \quad C(b) = (b-2)^{(b-1)/2} \sqrt{\frac{b-1}{3-b}} \quad . \tag{25}$$

Fig. 17 shows the results of numerical diagonalization for one realization with $N = 400$ and $b = 2.5$ such that we expect $R \sim N^{-0.25}$. It turns out that the circular eigenvalue density is rather well confirmed and the "theoretical radius" is indeed given by $R = \sqrt{N}\sigma$ if the variance σ^2 of matrix elements is determined by an average over the N^2 matrix elements of the given matrix. A study for different values of N with $50 \le N \le 2000$ also confirms the dependence $R = C\,N^{-\eta}$ with fit values $C = 0.67 \pm 0.03$ and $\eta = 0.22 \pm 0.01$. The value of $\eta = 0.22$ is close to the theoretical value $1 - b/2 = 0.25$ but the prefactor $C = 0.67$ is smaller than its theoretical value $C(2.5) \approx 1.030$. This is due to the correlations introduced by the additional column sum normalization after drawing the random matrix elements. Furthermore for the power law model with $b < 3$ we should not expect a precise confirmation of the uniform circular density obtained for Gaussian distribution matrix elements. Actually, a more detailed numerical analysis of the density shows that the density for the power law model is not exactly uniform, in particular for values of b close to 2.

The important observation is that a generic RPFM (full, sparse or with power law distributed matrix elements) has a complex eigenvalue density rather close to a uniform circle of a quite small radius (depending on the parameters N, Q or b). The fact, that the realistic networks (e.g. certain university WWW-networks) have Google matrix spectra very different from this [10], shows that in these networks there is indeed a subtle network structure and that already slight random perturbations or variations immediately result in uniform circular eigenvalue spectra. This was already observed in [8, 9], where it was shown that certain modest random changes in the network links already provide such circular eigenvalue spectra.

We also determine the PageRank for the different variants of the RPFM, i.e. the eigenvector for the eigenvalue $\lambda = 1$. It turns out that it is rather uniform that is rather natural since this eigenvector should be close to the uniform vector e which is the "PageRank" for the average matrix $\langle G_{ij}\rangle = 1/N$. This also holds when we use a damping factor $\alpha = 0.85$ for the RPFM.

Following the above discussion about triangular networks (with $G_{ij} = 0$ for $i \ge j$) we also study numerically a triangular RPFM where for $j \ge 2$ and $i < j$ the matrix elements G_{ij} are uniformly distributed in the interval $[0, 2/(j-1)[$ and for $i \ge j$ we have $G_{ij} = 0$. Then the first column is empty, that means it corresponds to a dangling node and it needs to be replaced by $1/N$ entries. For the triangular RPFM the situation changes completely since here the average matrix $\langle G_{ij}\rangle = 1/(j-1)$ (for $i < j$ and $j \ge 2$) has already a non-trivial structure and eigenvalue spectrum. Therefore the argument of degenerate perturbation theory which allowed to apply the results of standard full non-symmetric random matrices does not apply here. In Fig. 16 one clearly sees that for $N = 400$ the spectra for one realization of a triangular RPFM and its average are very similar for the eigenvalues with large modulus but both do not have at all a uniform circular density in contrast to the RPRM models without the triangular constraint discussed above. For the triangular RPFM the PageRank behaves as $P(K) \sim 1/K$ with the ranking index K being close to the natural order of nodes $\{1, 2, 3, \ldots\}$ that reflects the fact that the node 1 has the maximum of $N-1$ incoming links etc.

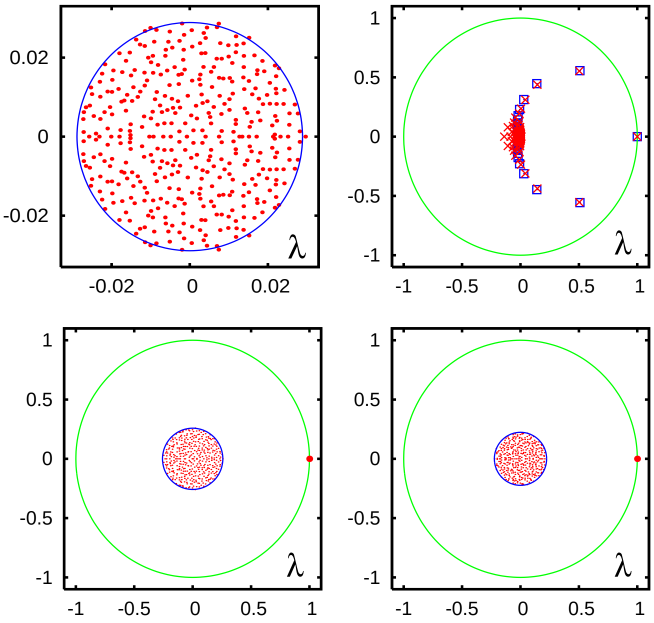

FIG. 16: (Color online) Top left panel shows the spectrum (red dots) of one realization of a full uniform RPFM with dimension $N = 400$ and matrix elements uniformly distributed in the interval $[0, 2/N[$. The blue circle represents the theoretical spectral border with radius $R = 1/\sqrt{3N} \approx 0.02887$. The unit eigenvalue $\lambda = 1$ is not shown due to the zoomed presentation range. Top right panel shows the spectrum of one realization of triangular RPFM (red crosses) with non-vanishing matrix elements uniformly distributed in the interval $[0, 2/(j-1)[$ and a triangular matrix with non-vanishing elements $1/(j-1)$ (blue squares). Here $j = 2, 3, \ldots, N$ is the index-number of non-empty columns and the first column with $j = 1$ corresponds to a dangling node with elements $1/N$ for both triangular cases. Bottom panels show the complex eigenvalue spectrum (red dots) of a sparse RPFM with dimension $N = 400$ and $Q = 20$ non-vanishing elements per column at random positions. The left (right) panel corresponds to the case of uniformly distributed non-vanishing elements in the interval $[0, 2/Q[$ (constant non-vanishing elements being $1/Q$). The blue circle represents the theoretical spectral border with radius $R = 2/\sqrt{3Q} \approx 0.2582$ ($R = 1/\sqrt{Q} \approx 0.2236$). In both bottom panels $\lambda = 1$ is shown by a larger red dot for better visibility. The unit circle is shown by green line (top right and bottom panels).

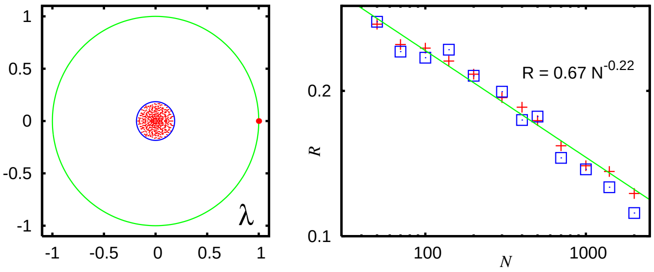

FIG. 17: (Color online) Left panel shows the spectrum (red dots) of one realization of the power law RPFM with dimension $N = 400$ and decay exponent $b = 2.5$ (see text). The unit eigenvalue $\lambda = 1$ is shown by a large red dot, the unit circle is shown by green curve. The blue circle represents the spectral border with theoretical radius $R =\approx 0.1850$ (see text). Right panel shows the dependence of the spectrum border radius on matrix size N for $50 \leq N \leq 2000$. Red crosses represent the radius obtained from theory (see text). Blue squares correspond to the spectrum border radius obtained numerically from a small number of eigenvalues with maximal modulus. The green line shows the fit $R = C\,N^{-\eta}$ of red crosses with $C = 0.67 \pm 0.03$ and $\eta = 0.22 \pm 0.01$.

The study of above models shows that it is not so simple to find a good RPFM model which reproduces a typical spectral structure of real directed networks.

VIII. DISCUSSION

In this study we presented a detailed analysis of the spectrum of the CNPR for the period 1893 – 2009. It happens that the numerical simulations should be done with a high accuracy (up to $p = 16384$ binary digits for the rational interpolation method or $p = 768$ binary digits for the high precision Arnoldi method) to determine correctly the eigenvalues of the Google matrix of CNPR at small eigenvalues λ. Due to the time ordering of citations, the CNPR G matrix is close to the triangular form with a nearly nilpotent matrix structure. We show that special semi-analytical methods allow to determine efficiently the spectrum of such matrices. The eigenstates with large modulus of λ are shown to select specific communities of articles in certain research fields but there is no clear way on how to identify a community one is interested in. The obtained results show that the spectrum of CNPR is characterized by the fractal Weyl law with the fractal dimension $d_f \approx 1$.

The ranking of articles is analyzed with the help of PageRank and CheiRank vectors corresponding to forward and backward citation flows in time. It is shown that the correlations between these two vectors are small and even negative that is similar to the case of Linux Kernel networks [26] and significantly different from networks of universities and Wikipedia. The 2DRanking on the PagRank-CheiRank plane allows to select articles which efficiently redistribute information flow on the CNPR.

To characterize the local impact propagation for a given article we introduce the concept of ImpactRank which efficiently determines its domain of influence.

Finally we perform the analysis of several models of RPFM showing that such full random matrices are very far from the realistic cases of directed networks. Random sparse matrices with a limited number Q of links per nodes seem to be closer to typical Google matrices concerning the matrix structure. However, still such random models give a rather uniform eigenvalue density with a spectral radius $\sim 1/\sqrt{Q}$ and also a flat PageRank distribution. Furthermore they do not capture the existence of quasi-isolated communities which generates quasi-degenerate spectrum at $\lambda = 1$. Further development of RPFM models is required to reproduce the spectral properties of real modern directed networks.

IX. ACKNOWLEDGMENTS

We thank the American Physical Society for letting us use their citation database for Physical Review [15]. This research 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.

[1] S. Brin and L. Page, Computer Networks and ISDN Systems **30**, 107 (1998).

[2] A.A. Markov, *Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga*, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, **15**, 135 (1906) (in Russian) [English trans.: *Extension of the limit theorems of probability theory to a sum of variables connected in a chain* reprinted in Appendix B of Howard RA *Dynamic Probabilistic Systems*, volume 1: *Markov models*, Dover Publ. (2007)].

[3] M. Brin and G. Stuck, *Introduction to Dynamical Systems*, Cambridge University Press, Cambridge, England, 2002.

[4] A. M. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press (Princeton, 2006).

[5] M.L.Mehta, *Random matrices*, Elsevier-Academic Press, Amsterdam (2004).

[6] D.L.Shepelyansky and O.V.Zhirov, Phys. Rev. E **81**, 036213 (2010).

[7] L.Ermann and D.L.Shepelyansky, Eur. Phys. J. B **75**, 299 (2010).

[8] O.Giraud, B.Georgeot and D.L.Shepelyansky, Phys. Rev. E **80**, 026107 (2009).

[9] B.Georgeot, O.Giraud and D.L.Shepelyansky, Phys. Rev. E **81**, 056109 (2010).

[10] K.M.Frahm, B.Georgeot and D.L.Shepelyansky, J. Phys, A: Math. Theor. **44**, 465101 (2011)

[11] L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, Eur. Phys. J. B **79**, 115 (2011).

[12] K.M.Frahm and D.L.Shepelyansky, Eur. Phys. J. B **85**, 355 (2012).

[13] L.Ermann, K.M.Frahm and D.L.Shepelyansky, Eur. Phys. J. B **86**, 193 (2013).

[14] Y.-H.Eom, K.M.Frahm, A.Benczur and D.L.Shepelyansky, preprint arXiv:1304.6601 [physics.soc-ph] (2013).

[15] Web page of Physical Review http://publish.aps.org/

[16] R. Albert and A.-L. Barabási, Phys. Rev. Lett. **85**, 5234 (2000).

[17] G. W. Stewart, *Matrix Algorithms Volume II: Eigensystems*, SIAM, 2001.

[18] K.M. Frahm and D.L. Shepelyansky, Eur. Phys. J. B **76**, 57 (2010).

[19] K.M.Frahm, A.D.Chepelianskii and D.L.Shepelyansky, J. Phys. A: Math. Theor. **45**, 405101 (2012).

[20] S. Redner, Phys. Today **58(6)**, 49 (2005)

[21] P. Chen, H. Xie, S. Maslov and S.Redner, J. Infometrics **1**, 8 (2007)

[22] F. Radicchi, S. Fortunato, B. Markines and A. Vespignani, Phys. Rev. E **80**, 056103 (2009).

[23] J.D. West, T.C. Bergstrom and C.T. Bergstrom, Coll. Res. Libr. **71**, 236 (2010); http://www.eigenfactor.org/

[24] A.D. Chepelianskii, *Towards physical laws for software architecture*, preprint arXiv:1003.5455[cs.Se] (2010)

[25] A.O.Zhirov, O.V.Zhirov and D.L.Shepelyansky, Eur. Phys. J. B **77**, 523 (2010)

[26] L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, J. Phys. A: Math. Theor. **45**, 275101 (2012)

[27] This number depends on the exact time ordering which is used and which is not unique because many papers are published at the same time and the order between them is not specified. We have chosen a time ordering where between these papers, degenerate in publication time, the initial node order of the raw data is kept.

[28] Note that some of the non-vanishing components of the iteration vector $S_0^i e$ may become very small, e.g. $\sim 10^{-100}$. In this context we count such components still as occupied despite their small size and N_i is the number of nodes which can be reached from some arbitrary other node after i iterations with the matrix S_0.

[29] In [19] a set of vectors without this prefactor was used but this provided a representation matrix which is numerically unstable for a direct diagonalization. The prefactor c_{j-1}^{-1} ensures that the representation matrix is numerically (rather) stable and of course both matrices are mathematically related by a similarity transformation and have identical eigenvalues.

[30] T. Granlund and the GMP development team, *GNU MP: The GNU Multiple Precision Arithmetic Library*, Version 5.0.5 (2012), http://gmplib.org/.

[31] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer (2002).

[32] J. Sjöstrand, Duke Math. J. **60**, 1 (1990).

[33] J. Sjöstrand and M. Zworski, Duke Math. J. **137**, 381 (2007).

[34] S. Nonnenmacher and M. Zworski, Commun. Math. Phys. **269**, 311 (2007).

[35] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).

[36] P.W. Anderson, Phys. Rev. **109**, 1492 (1958)

[37] G. Benettin, L. Galgani, and J.-M. Strelcyn, Phys. Rev. A **14**, 2338 (1976).

[38] D.J. Thouless, Phys. Rev. Lett. **39**, 1167 (1977).

[39] E. Abrahams, P.W. Anderson, D.C. Licciardello, and T.V. Ramakrishnan, Phys. Rev. Lett. **42**, 673 (1979).

[40] Y.-H. Eom and D.L. Sepelyansky, PLoS ONE **8(10)**, e74554 (2013).

[41] J. Ginibre, J. Math. Phys. Sci. **6**, 440 (1965).

[42] H.-J. Sommers, A. Crisanti, H. Sompolinsky and Y. Stein, Phys. Rev. Lett. **60**, 1895 (1988); N. Lehmann and H.-J. Sommers, Phys. Rev. Lett. **67**, 941 (1991).

[43] V. Kandiah and D. L. Shepelyansky, PLoS One **8(5)**, e61519 (2013).