Spectral properties of the Google matrix of the World Wide Web and other directed networks

Bertrand Georgeot,[1,2] Olivier Giraud*,[1,2] and Dima L. Shepelyansky[1,2]
[1] *Laboratoire de Physique Théorique (IRSAMC), Université de Toulouse, UPS, F-31062 Toulouse, France*
[2] *LPT (IRSAMC), CNRS, F-31062 Toulouse, France*

(Dated: February 17, 2010)

We study numerically the spectrum and eigenstate properties of the Google matrix of various examples of directed networks such as vocabulary networks of dictionaries and university World Wide Web networks. The spectra have gapless structure in the vicinity of the maximal eigenvalue for Google damping parameter α equal to unity. The vocabulary networks have relatively homogeneous spectral density, while university networks have pronounced spectral structures which change from one university to another, reflecting specific properties of the networks. We also determine specific properties of eigenstates of the Google matrix, including the PageRank. The fidelity of the PageRank is proposed as a new characterization of its stability.

PACS numbers: 89.20.Hh, 89.75.Hc, 05.40.Fb, 72.15.Rn

I INTRODUCTION

The rapid growth of the World Wide Web (WWW) brings the challenge of information retrieval from this enormous database which at present contains about 10^{11} webpages. An efficient algorithm for classification of webpages was proposed in [1], and is now known as the PageRank Algorithm (PRA). This PRA formed the basis of the Google search engine, which is used by the majority of Internet users in everyday life. The PRA allows to determine efficiently a vector ranking the nodes of a network by order of importance. This PageRank vector is obtained as an eigenvector of the Google matrix $\mathbf{G}$ built on the basis of the directed links between WWW nodes (see e.g. [2]):

$$\mathbf{G} = \alpha \mathbf{S} + (1-\alpha)\mathbf{E}/N \ . \tag{1}$$

Here $\mathbf{S}$ is the matrix constructed from the adjacency matrix A_{ij} of the directed links of the network of size N, with $A_{ij} = 1$ if there is a link from node j to node i, and $A_{ij} = 0$ otherwise. Namely, $S_{ij} = A_{ij}/\sum_k A_{kj}$ if $\sum_k A_{kj} > 0$, and $S_{ij} = 1/N$ if all elements in the column j of $\mathbf{A}$ are zero. The last term in Eq.(1) with uniform matrix $E_{ij} = 1$ describes the probability $1-\alpha$ of a random surfer propagating along the network to jump randomly to any other node. The matrix $\mathbf{G}$ belongs to the class of Perron-Frobenius operators. For $0 < \alpha < 1$ it has a unique maximal eigenvalue at $\lambda = 1$, separated from the others by a gap of size at least $1-\alpha$ (see e.g. [2]). The eigenvector associated to this maximal eigenvalue is the PageRank vector, which can be viewed as the steady-state distribution for the random surfer. Usual WWW networks correspond to very sparse matrix $\mathbf{A}$ and repeated applications of $\mathbf{G}$ on a random vector converges quickly to the PageRank vector, after $50-100$ iterations for $\alpha = 0.85$ which is the most commonly used value [2]. The PageRank vector is real nonnegative and can be ordered by decreasing values p_j, giving the relative importance of the node j. It is known that when α varies, all eigenvalues evolve as $\alpha\lambda_i$ where λ_i are the eigenvalues for $\alpha = 1$ and $i = 2, ...N$, while the largest eigenvalue $\lambda_1 = 1$, associated with the PageRank, remains unchanged [2].

The properties of the PageRank vector for WWW have been extensively studied by the computer science community and many important properties have been established [3–7]. For example, it was shown that p_j decreases approximately in an algebraic way $p_j \sim 1/j^\beta$ with the exponent $\beta \approx 0.9$ [3]. It is also known that typically for the Google matrix of WWW at $\alpha = 1$ there are many eigenvalues very close or equal to $\lambda = 1$, and that even at finite $\alpha < 1$ there are degeneracies of eigenvalues with $\lambda = \alpha$ (see e.g. [8]).

In spite of the important progress obtained during these investigations of PageRank vectors, the spectrum of the Google matrix $\mathbf{G}$ was rarely studied as a whole. Nevertheless, it is clear that the structure of the network is directly linked to this spectrum. Eigenvectors other than the PageRank describe the relaxation processes toward the steady-state, and also characterize various communities or subsets of the network. Even if models of directed networks of small-world type [9] have been analyzed, constructed and investigated, the spectral properties of matrices corresponding to such networks were not so much studied. Generally for a directed network the matrix $\mathbf{G}$ is nonsymmetric and thus the spectrum of eigenvalues is complex. Recently the spectral study of the Google matrix for the Albert-Barabasi (AB) model [10] and randomized university WWW networks was performed in [11]. For the AB model the distribution of links is typical of scale-free networks [9]. The distribu-

*present address: Laboratoire de Physique Théorique et Modèles Statistiques, UMR 8626 du CNRS, Université Paris-Sud, Orsay, France.

tion of links for the university network is approximately the same and is not affected by the randomization procedure. Indeed, the randomization procedure corresponds to the one proposed in [12] and is performed by taking pairs of links and inverting the initial vertices, keeping unchanged the number of ingoing and outgoing links for each vertex. It was established that the spectra of the AB model and the randomized university networks were quite similar. Both have a large gap between the largest eigenvalue $\lambda_1 = 1$ and the next one with $|\lambda_2| \approx 0.5$ at $\alpha = 1$. This is in contrast with the known property of WWW where λ_2 is usually very close or equal to unity [2, 8]. Thus it appears that the AB model and the randomized scale-free networks have a very different spectral structure compared to real WWW networks. Therefore it is important to study the spectral properties of examples of real networks (without randomization).

In this paper, we thus study the spectra of Google matrices for the WWW networks of several universities and show that indeed they display very different properties compared to random scale-free networks considered in [11]. We also explore the spectra of a completely different type of real network, built from the vocabulary links in various dictionaries. In addition, we analyze the properties of eigenvectors of the Google matrix for these networks. A special attention is paid to the PageRank vector and in particular we characterize its sensitivity to α through a new quantity, the PageRank fidelity.

The paper is organized as follows. In Section II we give the description of the university and vocabulary networks whose Google matrices we consider. The properties of spectra and eigenstates are investigated in Section III. The fidelity of PageRank and its other properties are analyzed in Section IV. Section V explores various models of random networks for which the spectrum can be closer to the one of real networks. The conclusion is given in Section VI.

II DESCRIPTION OF NETWORKS OF UNIVERSITY WWW AND DICTIONARIES

In order to study the spectra and eigenvectors of Google matrices of real networks, we numerically explored several systems.

Our first example consists in the WWW networks of UK universities, taken from the database [13]. The vertices are the HTML pages of the university websites in 2002. The links correspond to hyperlinks in the pages directing to another webpage. To reduce the size of the matrices in order to perform exact diagonalization, only webpages with at least one outlink were considered. There are still dangling nodes, despite of this selection, since some sites have outlinks only to sites with no outlink. We checked on several examples that the general properties of the spectra were not affected by this reduction in size. We present data on the spectra from five universities:

- University of Wales at Cardiff (www.uwic.ac.uk), with 2778 sites and 29281 links.
- Birmingham City University (www.uce.ac.uk); 10631 sites and 82574 links.
- Keele University (Staffordshire) (www.keele.ac.uk); 11437 sites and 67761 links.
- Nottingham Trent University (www.ntu.ac.uk); 12660 sites and 85826 links.
- Liverpool John Moores University (www.livjm.ac.uk); 13578 sites and 111648 links.

A much larger sample of university networks from the same database was actually used, including universities from the US, Australia and New Zealand, in order to insure that the results presented were representative.

As opposed to the full spectrum of the Google matrix, the PageRank can be computed and studied for much larger matrix sizes. In the studies of Section IV, we therefore included additional data from the university networks of Oxford in 2006 (www.oxford.ac.uk) with 173733 sites and 2917014 links taken from [13], and the network of Notre Dame University from the US taken from the database [14] with 325729 sites and 1497135 links (without removing any node).

In addition, we also investigated several vocabulary networks constructed from dictionaries; the network data were taken from [15].

- Roget dictionary (1022 vertices and 5075 links) [16]. The 1022 vertices correspond to the categories in the 1879 edition of Roget's Thesaurus of English Words and Phrases. There is a link from category X to category Y if Roget gave a reference to Y among the words and phrases of X, or if the two categories are related by their positions in the book.
- ODLIS dictionary (Online Dictionary of Library and Information Science), version December 2000 [17] (2909 vertices and 18419 links).
  A link (X,Y) from term X to term Y is created if the term Y is used in the definition of term X.
- FOLDOC dictionary (Free On-Line Dictionary of Computing) [18] (13356 vertices and 120238 links) A link (X,Y) from term X to term Y is created if the term Y is used in the definition of term X.

Distribution of ingoing and outgoing links for these university WWW networks is similar to those of much larger WWW networks discussed in [3, 7, 9]. An example is shown in the Appendix for the network of Liverpool John Moores University, together with data from AB models discussed in [11] (see Fig. 12).

III PROPERTIES OF SPECTRUM AND EIGENSTATES

To study the spectrum of the networks described in the previous section, we construct the Google matrix $\mathbf{G}$ associated to them at $\alpha = 1$. After that the spectrum λ_i and right eigenstates ψ_i of $\mathbf{G}$ (satisfying the relation $\mathbf{G}\psi_i = \lambda_i \psi_i$) are computed by direct diagonalization using standard LAPACK routines. Since $\mathbf{G}$ is generally a nonsymmetric matrix for our networks, the eigenvalues λ_i are distributed in the complex plane, inside the unit disk $|\lambda_i| \leq 1$.

The spectrum for our eight networks is shown in Fig.1. An important property of these spectra is the presence of eigenvalues very close to $\lambda = 1$ and moreover we find that $\lambda = 1$ eigenvalue has significant degeneracy. It is known that such an exact degeneracy is typical for WWW networks (see e.g. [7, 8]). In addition to this exact degeneracy, there are quasidegenerate eigenvalues very close to $\lambda = 1$. It is important to note that these features are absent in the spectra of random networks studied in [11] based on the AB model and on the randomization of WWW university networks, where the spectrum is characterized by a large gap between the first eigenvalue $\lambda_1 = 1$ and the second one with $|\lambda_2| \approx 0.5$. For example, the spectrum shown in Fig.1 panel H corresponds to the same university whose randomized spectrum was displayed in Fig.1 (bottom panel) in [11]. Clearly the structure of the spectrum becomes very different after randomization of links. Another property of the spectra displayed in Fig.1 that we want to stress is the presence of clearly pronounced structures which are different from one network to another. The structure is less pronounced in the case of the three spectra obtained from dictionary networks. In this case, the spectrum is flattened, being closer to the real axis. In contrast, for the WWW university networks, the spectrum is spread out over the unit disk. However, there is still a significant fraction of eigenvalues close to the real axis. We understand this feature by the existence of a significant number of symmetric ingoing and outgoing links (48 % in the case of the Liverpool John Moores University network), which is larger compared to the case of randomized university networks considered in [11].

To characterize the spectrum, we introduce the relaxation rate γ defined by the relation $|\lambda| = \exp(-\gamma/2)$. For characterization of eigenvectors $\psi_i(j)$, we use the PArticipation Ratio (PAR) defined by $\xi = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$. This quantity gives the effective number of vertices of the network contributing to a given eigenstate ψ_i; it is often used in solid-state systems with disorder to characterize localization properties (see e.g. [19]). The dependence of the density of states $W(\gamma)$ in γ, which gives the number of eigenstates in the interval $[\gamma, \gamma + d\gamma]$, is shown in Figs.2,3,4,5 (top panels). The normalization is chosen such that $\int_0^\infty W(\gamma) d\gamma = 1$, cor-

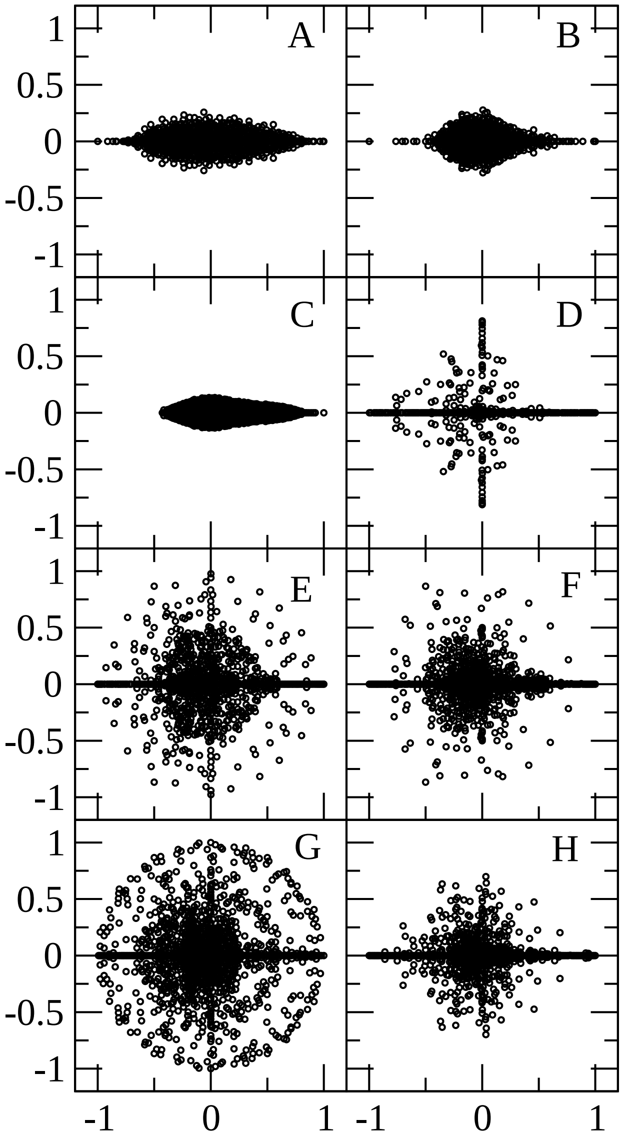

FIG. 1: Distribution of eigenvalues λ_i of Google matrices in the complex plane at $\alpha = 1$ for dictionary networks: Roget (A, N=1022), ODLIS (B, N=2909) and FOLDOC (C, N=13356); university WWW networks: University of Wales (Cardiff) (D, N=2778), Birmingham City University (E, N=10631), Keele University (Staffordshire) (F, N=11437), Nottingham Trent University (G, N=12660), Liverpool John Moores University (H, N=13578)(data for universities are for 2002).

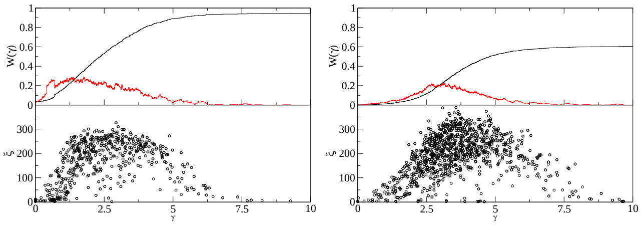

FIG. 2: (Color online) Left: Roget dictionary, $\alpha = 1$. Top panel: normalized density of states W (black) obtained as a derivative of a smoothed version of the integrated density (smoothed over a small interval $\Delta\gamma$ varying with matrix size), integrated density is shown in red (grey) . Bottom panel: PAR of eigenvectors as a function of γ; degeneracy of $\lambda = 1$ is 18 (note that the value $W(0)$ corresponds to eigenvalues with $|\lambda| = 1$). Right: ODLIS dictionary, same as left; degeneracy of $\lambda = 1$ is 4.

responding to the total number of eigenvalues N (equal to the matrix size). We also show the integrated version of this quantity in the same panels. In the same Figs we show the PAR ξ of the eigenstates as a function of γ (bottom panels).

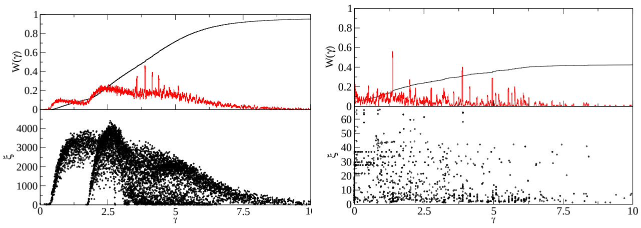

FIG. 3: (Color online) Left: FOLDOC dictionary, same as Fig. 2; degeneracy of $\lambda = 1$ is 1; Right: University of Wales (Cardiff), same as left; degeneracy of $\lambda = 1$ is 69.

It is clear that for the dictionary networks the density of states W depends on γ in a relatively smooth way, with a broad maximum at $\gamma \approx 1 - 2$. The distribution of PAR has also a maximum at approximately the same values. The case of the dictionary FOLDOC is a bit special, showing a bimodal distribution which is also clearly seen in the dependence of ξ on γ. This comes from the fact that the distribution of eigenvalues in Fig. 1 (panel C) is highly asymmetric with respect to the imaginary axis. The latter case has also no degeneracy at $\lambda = 1$. In these three networks the density of states decreases for γ approaching 0. We note that the integrated version of the density of states reaches a plateau for $\gamma \geq 6 - 7$. This saturation value is less than 1, meaning that a certain nonzero fraction of eigenvalues are extremely close to $\lambda = 0$.

For the WWW university networks, the density of states is much more inhomogeneous in γ. Even if a broad maximum is visible, there are sharp peaks at certain values of γ. The sharpest peaks correspond to exact degeneracies at certain complex values of λ. The degeneracies are especially visible at the real values $\lambda = 1/2, \lambda = 1/3$ and other $1/n$ with integer values of n. We attribute this phenomenon to the fact that the small number of links gives only a small number of different values for the matrix elements of the matrix $\mathbf{G}$. For the university networks, the degeneracy at $\lambda = 1$ is much larger than in the case of dictionaries. The integrated densities of states show visible vertical jumps which correspond to the degeneracies; their growth saturates at $\gamma \approx 7$ showing that about $30 - 50\%$ of the eigenvalues are located in the vicinity of $\lambda = 0$.

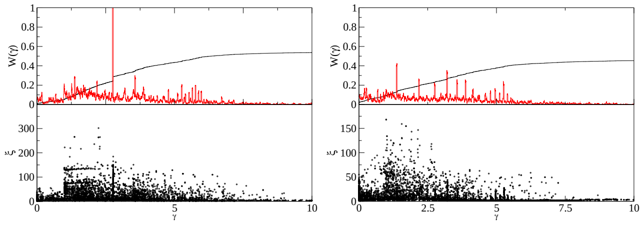

FIG. 4: (Color online) Left: Birmingham City University, same as Fig. 2; degeneracy of $\lambda = 1$ is 71; Right: Keele University (Staffordshire), same as left; degeneracy of $\lambda = 1$ is 205.

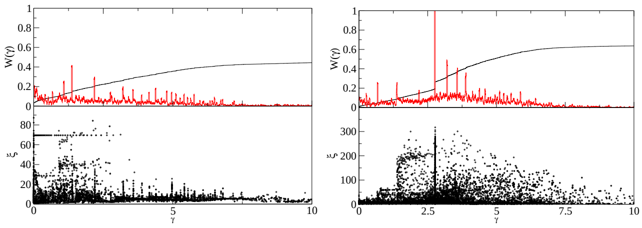

FIG. 5: (Color online) Left: Nottingham Trent University, same as Fig. 2; degeneracy of $\lambda = 1$ is 229. Right: Liverpool John Moores University, same as left; degeneracy of $\lambda = 1$ is 109; other degeneracy peaks correspond to $\lambda = 1/2$ (16), $\lambda = 1/3$ (8); $\lambda = 1/4$ (947), $\lambda = 1/5$ (97), being located at $\gamma = -2\ln\lambda$; other degeneracies are also present, e.g. $\lambda = 1/\sqrt{2}$ (41).

The PAR distribution for the university networks fluctuates strongly, even if a broad maximum is visible. Typical values have $\xi \approx 100$, which is small compared to the matrix size $N \sim 10^4$. This indicates that the majority of eigenstates are localized on certain zones of the network. This does not exclude that certain eigenstates with a larger ξ will be delocalized on a large fraction of the network in the limit of very large N.

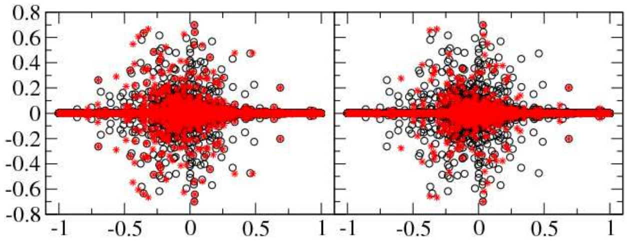

FIG. 6: Cloud of eigenvalues for Liverpool John Moores University, $\alpha = 1$. Circles: full matrix $N = 13578$. Stars: truncated matrix of size 8192 (left) and 4096 (right).

The exact $\mathbf{G}$ matrix diagonalization requires significant computer memory and is practically restricted to matrix size N of about $N < 30000$. However, real networks such as WWW networks can be much larger. It is therefore important to find numerical approaches in order to obtain the spectrum of large networks using approximate methods. A natural possibility is to order the sites through the PageRank method and to consider the spectrum of the (properly renormalized) truncated matrix restricted to the sites with PageRank larger than a certain value. In this way, the truncation takes into account the most important sites of the network. The effect of such a truncation is shown in Fig. 6 for the largest network of our sample. The numerical data show that the global features of the spectrum are preserved by moderate truncation, but individual eigenvalues deviate from their exact values when more than 50% of sites are truncated. Probably future developments of this approach are needed in order to be able to truncate a larger fraction of sites.

IV FIDELITY OF PAGERANK AND ITS OTHER PROPERTIES

In the previous section we studied the properties of the full spectrum and all eigenstates of the $\mathbf{G}$ matrix for several real networks. The PageRank is especially important since it allows to obtain an efficient classification of the sites of the network [1, 2]. Since the networks usually have small number of links, it is possible to obtain the PageRank by vector iteration for enormously large size of networks as described in [1, 2].

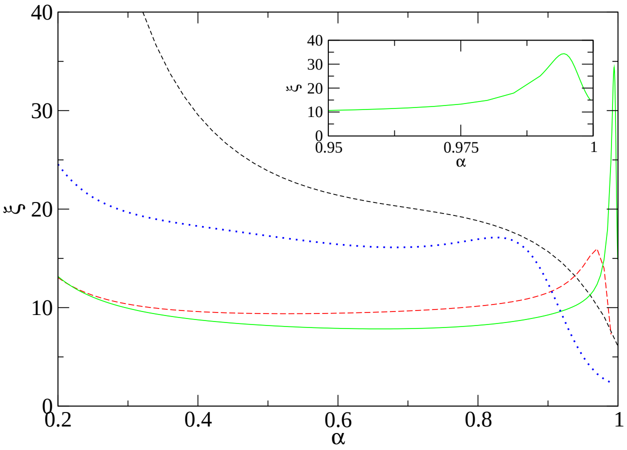

FIG. 7: (Color online) PAR ξ of PageRank as a function of α for University of Wales (Cardiff) (black/dashed), Notre-Dame (blue, dotted), Liverpool John Moores University (red/long dashed) and Oxford (green/solid) Universities (curves from top to bottom at $\alpha = 0.6$). Network sizes vary from $N = 2778$ to $N = 325729$. Inset is a zoom for data from Oxford University close to $\alpha = 1$.

Due to this significance of the PageRank, it is important to characterize its properties. In addition, it is important to know how sensitive the PageRank is with respect to the Google parameter (damping parameter) α in Eq. (1). The localization property of the PageRank can be quantified through the PAR ξ defined above. The dependence of ξ on α is shown in Fig.7 for four University WWW networks, including two from Fig.1 (panels D and H) and two of much larger sizes (Notre Dame and Oxford). For $\alpha \to 0$ the PAR goes to the matrix size since the $\mathbf{G}$ matrix is dominated by the second part of Eq. 1. However, in the interval $0.4 < \alpha < 0.9$ the dependence on α is rather weak, indicating stability of the PageRank. For $0.9 < \alpha < 1$ the PAR value has a local maximum where its value can be increased by a factor $2 - 3$. We attribute this effect to the existence of an exact degeneracy of the eigenvalue $\lambda = 1$ at $\alpha = 1$, discussed in the previous section. In spite of this interesting behavior of ξ in the vicinity of $\alpha = 1$, the value of ξ, which gives the effective number of populated sites, remains much smaller than the network size. In other models considered in [11, 20], a delocalization of the PageRank was observed for some α values, so that ξ was growing with system size N. For the WWW university networks considered here, delocalization is clearly absent (network sizes in Fig. 7 vary over two order of magnitudes). This is in agreement with the value of the exponent $\beta \approx 0.9$ for the PageRank decay, which was found for large samples of WWW data in [3, 7]. Indeed, for that value of β, PAR should be independent of system size.

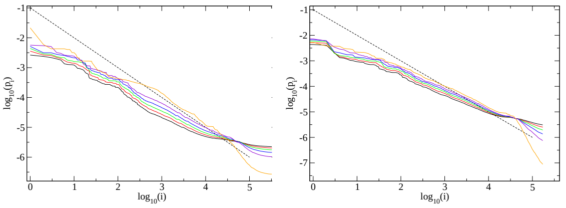

FIG. 8: (Color online) Some PageRank vectors p_j for Notre-Dame university (left panel) and Oxford (right panel). From top to bottom at $\log_{10}(i) = 5$: α=0.49 (black), 0.59 (red), 0.69 (green), 0.79 (blue) , 0.89 (violet) and 0.99 (orange). Dashed line indicates the slope -1.

Our data for PageRank distribution also show its stability as a whole for variation of α in the interval $0.4 < \alpha < 0.9$, as it is shown in Fig. 8.

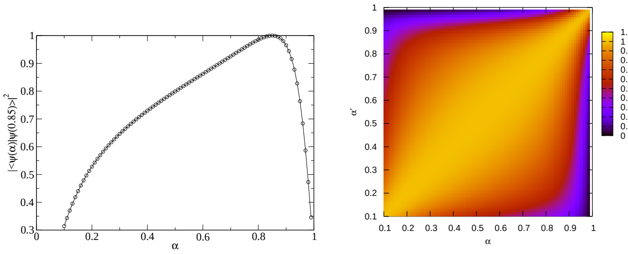

FIG. 9: (Color online) PageRank fidelity $f(\alpha, \alpha')$ for Notre-Dame university ($N = 325729$); left panel: $f(\alpha, \alpha' = 0.85) = |\langle\psi(\alpha)|\psi(0.85)\rangle|^2$ (see Eq. (2)); right panel: color density plot of $f(\alpha, \alpha')$.

The sensitivity of the PageRank with respect to α can be more precisely characterized through the *PageRank fidelity* defined as

$$f(\alpha, \alpha') = |\sum_j \psi_1(j, \alpha)\psi_1(j, \alpha')|^2 , \qquad (2)$$

where $\psi_1(j, \alpha)$ is the eigenstate at $\lambda = 1$ of the Google matrix $\mathbf{G}$ with parameter α in Eq. (1); here the sum over j runs over the network sites (without PageRank reordering). We remind that the eigenvector $\psi_1(j, \alpha)$ is normalized by $\sum_j \psi_1(j, \alpha)^2 = 1$. Fidelity is often used in the context of quantum chaos and quantum computing to characterize the sensitivity of wavefunctions with respect to a perturbation [21, 22]. The variation of this quantity with α and α' is shown in Fig. 9. The fidelity reaches its maximum value $f = 1$ for $\alpha = \alpha'$. According to Fig. 9 (right panel), the stability plateau where fidelity remains close to 1, indicating stability of PageRank, is broadest for $\alpha = 0.5$. This is in agreement with previous results presented in [23], where the same optimal value of α was found based on different arguments.

V SPECTRUM OF MODEL SYSTEMS

The results obtained in [11] compared to those presented in the previous section show that while the spectrum of the network has a large gap between $\lambda = 1$ and the other eigenvalues, still certain properties of the PageRank can be similar in both cases (e.g. the exponent β). In fact the studies performed in the computer science community were often based on simplified models, which can nevertheless give the value of β close to the one of real networks. For example, the model studied by Avrachenkov and Lebedev [5] allows to obtain analytical expressions for β with a value close to the one obtained for WWW. It is interesting to see what are the spectral properties of this model. In Fig. 10 we show the spectrum for this model for $\alpha = 0.85$. Our data show that this model has an enormous gap, thus being very different from spectra of real networks shown in Fig. 1.

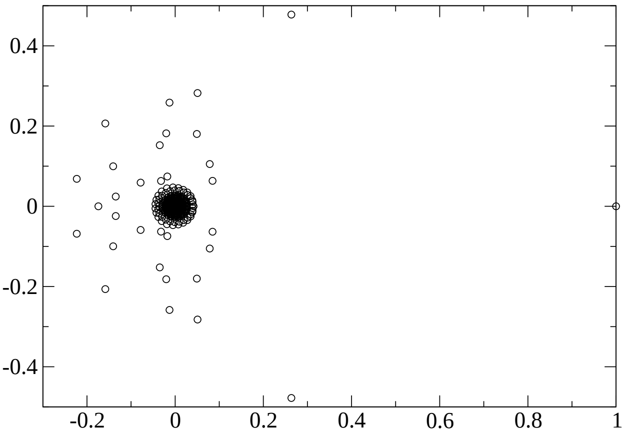

FIG. 10: Spectrum of eigenvalues λ in the complex plane for the Avrachenkov-Lebedev model of [5], with $N = 2^{11}$ (network size), $\alpha = 0.85$, $m = 5$ outgoing links per node. Multiplicity of links is taken into account in the construction of $\mathbf{G}$.

The above results, together with those of [11], show that many commonly used network models are characterized by a large gap between $\lambda = 1$ and the second eigenvalue, in contrast with real networks. In order to build a network model where this gap is absent, we introduce here what we call the color model. It is an extension of the AB model, that allows to obtain results for the spectral distribution that are closer to real networks. We divide the nodes into n sets ("colors"), allowing n to grow with network size. Each node is labeled by an integer between 0 and $n-1$. At each step, links and nodes are added as in the AB model but also with probability η the new node is introduced with a new color. The only links authorized between nodes are links within each set. Such a structure implies that the second eigenvalue of matrix G is real and exactly equal to α [24]. The colors correspond to communities in the network.

In order to have a more realistic model, we allow for the rule for links to be broken with some probability ε. That is, at each time step an link between two nodes is chosen at random according to the rules of the AB model. Then if it agrees with the color rule above it is used; if it does not then with probability $1-\varepsilon$ it is just omitted, and with probability ε it is nevertheless added.

The spectrum of this color model is shown in Fig. 11 for $\alpha = 0.85$. The second eigenvalue is now exactly at $\lambda = 0.85$, demonstrating the absence of a gap. There is also a set of eigenvalues which is located on the real line, but the majority of states remains inside a circle $|\lambda| < 0.3$ as in the AB model.

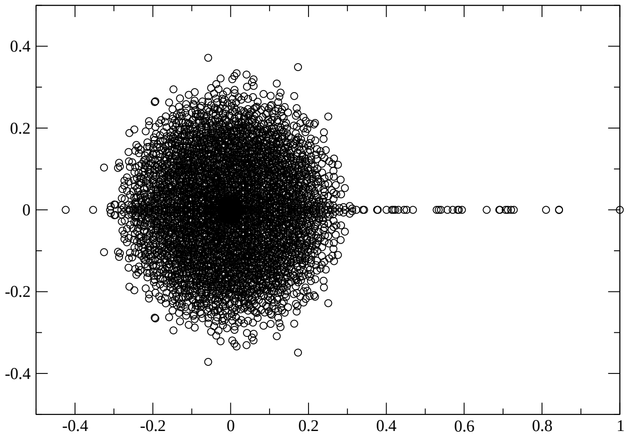

FIG. 11: Spectrum of eigenvalues λ in the complex plane for the color model, $N = 2^{13}$, $p = 0.2$, $q = 0.1$, $\alpha = 0.85$. Nodes are divided into n color sets labeled from $i = 0$ to $n-1$; nodes and links are created according to AB model; only authorized links are links within a color set i. This rule is broken with probability $\epsilon = 10^{-3}$. We start with 3 color sets; with probability η a new color is introduced (we take $\eta = 10^{-2}$). In the example displayed, when the number of nodes reaches N, $n = 83$ colors.

Thus the color model allows to eliminate the gap, but still the distribution of eigenvalues λ in the complex plane remains different from the spectra of real networks shown in Fig. 1: the structures prominent in real networks are not visible, and eigenvalues in the gap are concentrated only on the real axis or close to it.

VI CONCLUSION

In this work we performed numerical analysis of the spectra and eigenstates of the Google matrix $\mathbf{G}$ for several real networks. The spectra of the analyzed networks have no gap between first and second eigenvalues, in contrast with commonly used scale-free network models (e.g. AB model). The spectra of university WWW networks are characterized by complex structures which are different from one university to another. At the same time, PageRank of these university networks look rather similar. In contrast, the Google matrices of vocabulary networks of dictionaries have spectra with much less structure.

These studies show that usual models of random scale-free networks miss many important features of real networks. In particular, they are characterized by a large spectral gap, which is generally absent in real networks. We attribute the physical origin of this gap to the known property of small-world and scale-free networks that only logarithmic time (in system size) is needed to go from any node to any other node (see e.g. [9]). Due to that, the relaxation process in such networks is fast and the gap, being inversely proportional to this time, is accordingly very large. In contrast, the presence of weakly coupled communities in real networks makes the relaxation time very large, at least for certain configurations. Therefore, it is desirable to construct new random scale-free models which could capture in a better way the actual properties of real networks. The color model presented here is a first step in this direction. We note that Ulam networks built from dynamical maps can capture certain properties of real networks in a relatively good manner [20, 25]. In these latter networks, it is possible to have a delocalization of the PageRank when α or map parameters vary; we didn't observe such feature here.

Indeed, our data show that the PageRank remains localized for all values of $\alpha > 0.3$. We also showed that the use of the fidelity as a new quantity to characterize the stability of PageRank enables to identify a stability plateau located around $\alpha = 0.5$.

We think that future investigation of the spectral properties of the Google matrix will open new access to identification of important communities and their properties which can be hidden in the tail of the PageRank and hardly accessible to classification by the PageRank algorithm. Furthermore, the degeneracies at various values of λ and the characteristic patterns directly visible in the spectra of the Google matrix should allow to identify other hidden properties of real networks.

We thank Leonardo Ermann and Klaus Frahm for discussions, and CALcul en MIdi-Pyrénées (CALMIP) for the use of their supercomputers.

[1] S. Brin and L. Page, Computer Networks and ISDN Systems **33**, 107 (1998).

[2] A. M. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press (Princeton, 2006); D. Austin, AMS Feature Columns (2008) available at http://www.ams.org/featurecolumn/archive/pagerank.html

[3] D. Donato, L. Laura, S. Leonardi and S. Millozzi, Eur. Phys. J. B **38**, 239 (2004); G. Pandurangan, P. Raghavan and E. Upfal, Internet Math. **3**, 1 (2005).

[4] P. Boldi, M. Santini, and S. Vigna, in *Proceedings of the 14th international conference on World Wide Web*, A. Ellis and T. Hagino (Eds.), ACM Press, New York p.557 (2005); S. Vigna, **ibid.** p.976.

[5] K. Avrachenkov and D. Lebedev, Internet Math. **3**, 207 (2006).

[6] K. Avrachenkov, N. Litvak, and K.S. Pham, in *Algorithms and Models for the Web-Graph: 5th International Workshop, WAW 2007 San Diego, CA, Proceedings*, A. Bonato and F.R.K. Chung (Eds.), Springer-Verlag, Berlin, Lecture Notes Computer Sci. **4863**, 16 (2007)

[7] K. Avrachenkov, D. Donato and N. Litvak (Eds.), *Algorithms and Models for the Web-Graph: 6th International Workshop, WAW 2009 Barcelona, Proceedings*, Springer-Verlag, Berlin, Lecture Notes Computer Sci. **5427**, Springer, Berlin (2009).

[8] S.Serra-Capizzano, SIAM J. Matrix Anal. Appl. **27**, 305 (2005).

[9] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks*, Oxford University Press (Oxford, 2003).

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

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

[12] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[13] Academic Web Link Database Project http://cybermetrics.wlv.ac.uk/database/

[14] Albert-László Barabási webpage at the University of Notre Dame http://www.nd.edu/˜networks/resources.htm

[15] Vladimir Batagelj and Andrej Mrvar (2006): Pajek datasets. URL: http://vlado.fmf.uni-lj.si/pub/networks/data/.

[16] Peter Mark Roget: Roget's Thesaurus of English Words and Phrases, http://www.gutenberg.org/etext/22

[17] Joan M. Reitz (2002): ODLIS: Online Dictionary of Library and Information Science, http://vax.wcsu.edu/library/odlis.html

[18] Denis Howe, Editor: FOLDOC (2002): Free on-line dictionary of computing, http://foldoc.org/

[19] F. Evers and A. D. Mirlin, Rev. Mod. Phys. **80**, 1355 (2008).

[20] D. L. Shepelyansky and O. V. Zhirov, preprint arXiv:0905.4162 (2009).

[21] T. Gorin, T. Prosen, T. H. Seligman and M. Znidaric, Physics Reports **435**, 33 (2006).

[22] K. M. Frahm, R. Fleckinger and D. L. Shepelyansky, Eur. Phys. J. D **29**, 139 (2004).

[23] K. Avrachenkov, N. Litvak and K. S. Pham, Internet Math. **5**, 47 (2009).

[24] T. Haveliwala and S. Kamvar, *The Second Eigenvalue of the Google Matrix*, Stanford University Technical Report nr 582 (2003).

[25] L. Ermann and D. L. Shepelyansky, preprint arXiv:0911.3823 (2009).

APPENDIX

Here we show the distributions of links for the AB model discussed in [11] and for the university WWW network (panel H of Fig. 1).

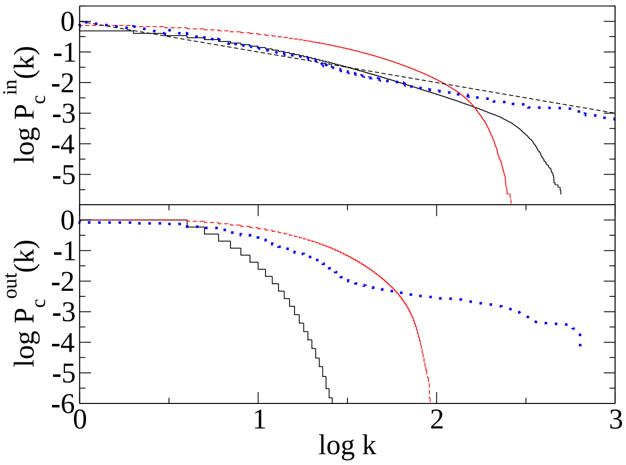

FIG. 12: Cumulative distribution of ingoing links $P_c^{in}(k)$ (top panel) and of outgoing links $P_c^{out}(k)$ (bottom panel) for the AB model with vector size $N = 2^{14}$, for $q = 0.1$ (black/solid) and $q = 0.7$ (red/dashed), data are averaged over 80 realizations of AB model, and for the network of Liverpool John Moores University with $N = 13578$, (panel H in Fig. 1) (blue/dotted). Average number of in- or outgoing links is $< k >= 6.43$ for $q = 0.1$, $< k >= 14.98$ for $q = 0.7$, $< k >= 8.2227$ for LJMU. Dashed straight line indicates the slope -1. Logarithms are decimal.