PageRank equation and localization in the WWW

N. Perra$^{1,2,3(a)}$, V. Zlatić$^{4,5}$, A. Chessa$^{1-2}$, C. Conti$^{6,7}$, D. Donato$^8$ and G. Caldarelli$^{2,4,7}$

$^1$Department of Physics, SLACS-CNR University of Cagliari - S.P. Monserrato-Sestu Km 0.700, 09042 Cagliari, Italy, EU
$^2$Linkalab, Complex Systems Computational Laboratory - 09100 Cagliari Italy, EU
$^3$Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University Bloomington, IN, USA
$^4$Centre SMC CNR-INFM, Department of Physics, University “Sapienza” - P.le Moro 5, 00185 Rome, Italy, EU
$^5$Theoretical Physics Division, Rudjer Boskovic Institute - P.O. Box 180, HR-10002 Zagreb, Croatia
$^6$Centre SOFT INFM-CNR, Department of Physics, University “Sapienza” - P.le Moro 5, 00185 Rome, Italy, EU
$^7$Institute for Complex Systems ISC-CNR - Via dei Taurini 19, 00185 Rome, Italy, EU
$^8$Yahoo! Research - Avinguda Diagonal 177, 8th floor, 08018 Barcelona, Spain, EU

received 17 August 2009; accepted in final form 23 October 2009
published online 20 November 2009

PACS 89.75.Fb – Structures and organization in complex systems
PACS 05.10.-a – Computational methods in statistical physics and nonlinear dynamics
PACS 71.23.-k – Electronic structure of disordered solids

Abstract – We show that the PageRank in a network can be represented as the solution of a differential equation discretized over a directed graph. By exploiting a formal relationship with the time-independent Schrödinger equation it is possible to interpret hub formation and related phenomena as a wave-like localization process in the presence of disorder and trapping potentials. The result opens new perspectives in the physics of networks with interdisciplinary connections and opens the way to the employment of various mathematical techniques to the analysis of self-organization in structured systems. Applications are envisaged in the World-Wide Web, traffic, social and biological networks.

Copyright © EPLA, 2009

Graphs represent one of the most recent approaches to describe and model different natural and technological systems [1,2]. In their simplest definitions, graphs are composed by a set of vertices (sites), connected by edges (links). Thanks to this simple characterization most real networks can be described within the mathematical framework of Graph Theory [3]. Most of the activity in this area has been triggered by the experimental evidence that when considering real systems as the Internet [4], WWW [5], and social systems [6,7] statistical properties of most of the graph quantities are power-law (or fat-tailed) distributed. For example, the degree $k$ (i.e. the total number of edges per vertex) is such that its probability distribution $P(k)$ scales, in most real cases, as $P(k) \propto k^{-\gamma}$, with $2 < \gamma < 3$. This result ensures a convergent mean value for $k$, while its variance diverges for $\gamma < 3$. This scale-free topology has direct consequences on the stability of the whole system (i.e. the number of failures it can undergo before breakdown [8]) and on the dynamics of related processes (e.g., epidemics on a social network diffuse differently than on a lattice [9]). A huge amount of research was devoted to the analysis of real networks topology [10], addressing issues like communities, distance distributions etc. Here, we focus on a topological quantity particularly important for the applications to information technology, that of the PageRank $P_i$ of the vertex $i$ [11]. In particular, we propose a new interpretation for this quantity that can in principle change the way in which this quantity is computed and analyzed in its dynamical evolution.

PageRank is defined as the stationary distribution of a discrete-time, finite-state Markov chain given by a random walk on the graph [12]. In other words the PageRank of a vertex can be defined as the time spent by a random surfer on it. This idea can be formalized by introducing a matrix representation of the graph given by the adjacency matrix $A$. The entries $a_{ij}$ of such matrix are 1 if there is a directed link from $i$ to $j$ and 0 otherwise. The outdegree matrix $K^O$ is defined such that the only not vanishing elements are on the diagonal and are given by the outdegree of the vertices, i.e. $k_i = \sum_j a_{ij}$. In such a way a transfer matrix,
Fig. 1: A very simple directed graph. Vertices C and D have one outgoing link (\(k^O = 1\)) and no incoming links (\(k^I = 0\)), vertex A has three incoming links (\(k^I = 3\)) and one outgoing link (\(k^O = 1\)), vertex B has both one incoming link and one outgoing link (\(k^O = k^I = 1\)). Any walker arriving in A will get trapped in the region A,B.

\(T = (K^O)^{-1}A^T\), can be defined for the process of random walk in the graph. This would allow us to simply compute the PageRank as the principal eigenvector of the equation \(P = TP\) or (by expanding the above vectorial equation) as

\[ P_i = \sum_{j \rightarrow i} \frac{P_j}{k^O_j}, \]

where P is the vector composed by the values of PageRank in all the pages of the system. The above equation is not generally solvable, except in certain particular cases. This can be noticed by considering that in an oriented graph a random walker can be trapped in a confined region. In other words, the matrix is not stochastic (see fig. 1 where \(k^I\) is the in-degree, i.e., the number of edges pointing to a given vertex). In order to overcome this difficulty, one can heuristically add a small perturbation. This is such that the network surfer can escape from a sink (e.g., vertex A in fig. 1) by jumping into a random vertex with a probability \(1/N\), where \(N\) is the total number of pages. In such a way the PageRank equation takes on the form

\[ P_i = \alpha \sum_{j \rightarrow i} \frac{P_j}{k^O_j} + (1 - \alpha) \frac{1}{N}, \]

where the sum runs on all the vertices \(j\) pointing to vertex \(i\). The parameter \(\alpha\) tunes this “teleportation” probability and allows us to solve the above systems of equations. Alternatively the PageRank vector P can be seen as the solution of the equation

\[ P = \alpha ((K^O)^{-1}A^T) + (1 - \alpha)E \]

where \(E\) is a matrix whose elements are \(1/N\), \(\alpha\) (damping factor) weights the two contributions and the correct stochastic matrix of the process is given by \(\alpha ((K^O)^{-1}A^T) + (1 - \alpha)E\).

Here we write eq. (2) in a different way in order to get some more physical insight. The first algebraic passage is to introduce, in the strongly connected component (SCC) of the graph, the variable \(\psi = P_i/k_i^O\) and then to add and subtract the term \(\alpha \psi_i k_i^I/k_i^O\) to eq. (2). In this way we obtain

\[ \psi_i = \frac{\alpha}{k^O_i} \left( \sum_{j \rightarrow i} \psi_j - k_i^I \psi_i \right) + \frac{k_i^I}{k_i^O} \psi_i + \frac{1 - \alpha}{\alpha} \frac{1}{N}, \]

and after some rearrangements

\[- \sum_{j \rightarrow i} \psi_j - k_i^I \psi_i + \left[ \frac{k_i^O - \alpha k_i^I}{\alpha} \psi_i \right] = \frac{1 - \alpha}{\alpha} \frac{1}{N}. \]

We can now spot in (5) the two distinct terms in square brackets. As discussed below, the first one can be put in relation with a discretized graph operator, while the second one can be considered as a potential function \(V\) such that eq. (5) can be rewritten as

\[- \nabla^2_D + V \] \[ \psi = F. \]

Here, \(F\) is a vector with constant entries \(F_i = (1 - \alpha)/\alpha N\), the components of \(V\) are given by \((k_i^O - \alpha k_i^I)/\alpha\) and the operator \(\nabla^2_D\) can be defined as the discretized Laplacian on an oriented graph. Indeed, the discretized Laplacian operator (acting on a test vector \(\phi_i\)) on a regular lattice is given by

\[ (\nabla^2) \phi_i \equiv \sum_{\langle j,i \rangle} \psi_{j} - k_{i} \phi_{i}, \]

where the sum runs on all the \(k\) neighbors \(j\) of site \(i\). The Laplacian of \(\phi\) in a point is given by the sum of the values on all the neighbors minus \(k\) times (where \(k\) is the degree or connectivity of the lattice) the value in the site considered.

The operator \(\nabla^2_D\) in eq. (6) is the directed counterpart of the discretized Laplacian (hence its subscript “\(D\)”). In the case of the WWW and, more in general, for any directed graph, the edges can be travelled only in one direction. This means that the lattice is not “reciprocal”, where with reciprocity we intend the property in a graph to reach vertex B starting from A, if A can be reached from B (see also ref. [13]). In particular, if for every edge in a graph from \(i\) to \(j\) there is also the symmetric edge \(j\) to \(i\), the reciprocity is 100%. More generally the reciprocity is given by the fraction of symmetric edges in the graph.

To our knowledge the properties of the Laplacian on these directed lattices have never been considered before. The operator \(\nabla^2\) and its directed counterpart \(\nabla^2_D\) will, in general, be very different; however we expect this to be dependent on the reciprocity. We tested the solution of the Laplace equation \(\nabla^2\psi = 0\) against the correspondent directed Laplace equation \(\nabla^2_D\psi = 0\) in a series of lattices (both regular as the simple cube and some realizations of Barabási-Albert network) starting from a completely reciprocal case and deleting randomly some of the connections. Generally, provided that the proportion of reciprocal links is above the percolation threshold for

48002-p2
PageRank equation and localization in the WWW

Fig. 2: (Color online) Plot of $\psi$, obeying the equation $\nabla^2 \psi = 0$ (complete reciprocal lattice and complete reciprocal Barabási-Albert graph) and $\nabla^2 D \psi = 0$ (all the other curves) in a 2-d simple cubic lattice and in a Barabási-Albert network. In the simple cubic case the upper and lower layer are kept at the fixed value of $\psi = 0$ and $\psi = 1$, respectively. The average value increases from 0 to 1, even when the reciprocity in the links between pairs of nodes is broken up to 30% of the completely reciprocal case. In the inset the plot of the same quantity for a Barabási-Albert model where the same boundary conditions of $\psi = 1$ and $\psi = 0$ are applied to the leaves and to the core of the structure, respectively. Here four shells are considered starting from a node in the core and the behavior is the same even for a 20% removal of reciprocal links.

Fig. 3: (Color online) 3D plot of potential $V$ and the corresponding PageRank measured along concentric shells around the vertex with the highest value of PageRank.

Fig. 4: (Color online) Shell average potential (red, left scale, dashed line) and shell PageRank $P$ (blue, right scale, continuous line) as obtained in the presence of an hub, i.e. a site with large in-degree, small out-degree and therefore very low potential. These quantities are computed over concentric shells of neighbors. The shell average is obtained by averaging over nodes on the same shell.

the lattice considered [14], the statistical behavior of the directed Laplacian is the same of the non-directed one (see fig. 2). In the particular case in which the network is reciprocal ($V = 0$) no trapping states are present and we can therefore consider $\alpha = 1$ [15]. Equation (6) becomes then a Laplace equation whose solution, $\psi$, is given by a constant function. We find that in this case the PageRank is proportional to the degree of the page [16]. This limit case in which a constant distribution of $V$ gives a trivial distribution of $\psi$ suggest that the term with $V$ in (6) can be considered similar to that of a potential, and we can expect that the PageRank will be localized around the minima of such a potential. To test this hypothesis we considered a snapshot of the real WWW, collected by Dipartimento di Scienze dell’Informazione (DSI), Università degli Studi di Milano and consisting of the 786049 pages of the .eu domain connected by 18120539 edges [17]. In fig. 3 we show a 3D representation of the potential and the corresponding value of PageRank visualized as follows: given one reference node (the node with the highest PageRank value in this case) that is put in the center of a conventional ($x$, $y$)-plane, we arrange in circular shells the first, second, third ... $n$-th neighbors. For each cell the neighbors are settled as equispaced points in a circle and the corresponding Potential and PageRank are given in the $z$-axis. The resulting 3D plots are given in fig. 3, and the highest PageRank value node is such that the potential display a pronounced minimum, and correspondingly the PageRank is localized. In fig. 4 we show the PageRank after averaging over shell nodes as linear plot.

Within this framework, the higher scores of the PageRank will be localized in the potential wells, and correspondingly nodes associated to peaks of the potential ("repulsive" regions) will display a low PageRank, as shown in fig. 5. This clarifies the role of the potential, in the real WWW there is no complete reciprocity and the difference between outdegree and indegree of a page plays the role of a topological disorder. It is important to stress that, since the Web has not a simple topology, the fact that a page is a minimum or a maximum of the potential is only evident when plotting the values at the nearest neighbors in the network. By exploiting topological information in the potential $V$, one is able to gain information about the PageRank spatial distribution. In particular by using the value of the potential as a rule of thumb to determine
starting from eq. (6):

$$\psi = (\mathbf{I} - \mathbf{V}^{-1} \nabla^2_D)^{-1} \mathbf{V}^{-1} \mathbf{F}.$$  

Equation (6) can then be interpreted in various ways, from a Poisson equation in a disordered medium to an inhomogeneous Helmholtz equation [18]. We decided to exploit the similarity with a time independent Schrödinger equation (with the addition of a constant term on the r.h.s.) because, in our opinion, this clarifies in a particularly clear way the role of the potential $\mathbf{V}$. In this perspective, the r.h.s. of (6) plays the role of a stochastic source as often encountered, e.g., in the c-number representation of quantum field-theories [19]. This interpretation suggests a relatively simple way to compute the whole distribution of the PageRank. In principle once the matrices of the Laplacian operator and the potential operator are known, the $\psi$ (and henceforth the set of PageRank values) could be computed by inverting these operators. This simple operation is unfeasible when the size of the matrix is of the order of tens of billions of pages as in the WWW. Here we adopt a different approach based on a matrix expansion that can be also extended to study the time evolution. The idea is to rewrite the above equation by using the common Taylor expansion, while

$$\psi = (\mathbf{I} - \mathbf{V}^{-1} \nabla^2_D)^{-1} \mathbf{V}^{-1} \mathbf{F}.$$  

We now expand the expression in brackets by writing

$$(\mathbf{I} - \mathbf{V}^{-1} \nabla^2_D)^{-1} = \sum_{n=0}^{\infty} (\mathbf{V}^{-1} \nabla^2_D)^n,$$

provided all the eigenvalues $\lambda_n$ of $(\mathbf{V}^{-1} \nabla^2_D)$ have $|\lambda_n| < 1$. This allows to invert only the diagonal matrix $\mathbf{V}$ (that can be done easily by taking the inverse of the elements on the diagonal). The expression above can be rewritten as

$$\psi = (\mathbf{k}^O - \alpha \mathbf{A}^T)^{-1} \mathbf{F}' = (\mathbf{I} - \alpha \mathbf{B})^{-1} (\mathbf{k}^O)^{-1} \mathbf{F}',$$  

where $\mathbf{F}' = \alpha \mathbf{F}$, $\mathbf{k}^O$ is a matrix whose elements are all zero apart on the diagonal where they are given by the outdegree of vertices, $\mathbf{A}^T$ is the transpose of the adjacent matrix and $\mathbf{B} = (\mathbf{k}^O)^{-1} \mathbf{A}^T$.

Equation (10) closely resembles the original equation for PageRank, with the important caveat that we are now working with a wave function $\psi$. In this case, the expansion

$$(\mathbf{I} - \alpha \mathbf{B})^{-1} = \sum_{n=0}^{\infty} (\alpha \mathbf{B})^n$$

converges and we can calculate with the desired precision $\psi$ and so the associated PageRank. The results of this matrix expansion are in good agreement with the solution obtained by traditional methods. One can increase as desired the order of the expansion with a computational cost that increases only linearly with the order.

In conclusion, we have shown that computing the PageRank can be recast in the form of several well known problems in physics, from the charge-distribution in an inhomogeneous medium to the wave-localization phenomena in quantum physics, at the cost of replacing the continuous space by a discretized and oriented graph. While being a fascinating mathematical problem, this not only furnishes a variety of insights in interpreting network theory (as for examples locating “repulsive” regions in a Web and relating hub formation to Anderson-like localizations) but also suggest several further deepenings. Indeed, one can consider the application of the standard perturbation theory to determine the topological effect of new nodes in a given network, or use nonlinear generalization of the proposed PageRank equation to explain condensation phenomena in the Web. These avenues of investigation are currently being developed by the authors, and will potentially find applications in a variety of cases, from the WWW to biological networks.

REFERENCES

[1] Albert R. and Barabási A.-L., Rev. Mod. Phys., 74 (2002) 47.
[2] Newman M. E. J., SIAM Rev., 45 (2003) 167.
[3] Bollobas B., Graph Theory an Introductory Course (Springer-Verlag, New-York, Heidelberg, Berlin) 1979.
[4] Vázquez A., Pastor-Satorras R. and Vespignani A., Phys. Rev. E, 65 (2002) 066130.
[5] Donato D., Leonardi S., Millozzi S. and Tsaparas P., J. Phys. A, 41 (2008) 224017.
[6] Capocci A., Servévé V. D. P., Colaiori F., Buriol L. S., Donato D., Leonardi S. and Caldarelli G., Phys. Rev. E, 74 (2006) 036116.
[7] Catanzaro M., Caldarelli G. and Pietronero L., Phys. Rev. E, 70 (2004) 037101.
[8] Albert R., Jeong H. and Barabási A. L., *Nature*, **406** (2000) 378.

[9] Pastor-Satorras R. and Vespignani A., *Phys. Rev. E*, **65** (2002) 036104.

[10] Bianconi G., Caldarelli G. and Capocci A., *Phys. Rev. E*, **71** (2005) 066116.

[11] Brin S. and Page L., *Comput. Netw.*, **30** (1998) 107.

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

[13] Garlaschelli D. and Loffredo M. I., *Phys. Rev. Lett.*, **93** (2004) 188701.

[14] Stauffer D. and Aharony A., *Introduction to Percolation Theory* (Taylor and Francis) 1994.

[15] Boldi P., Santini M. and Vigna S., *Proceedings of the Fourteenth International World Wide Web Conference (WWW 2005)* (Chiba, Japan, ACM Press) 2005, p. 557–566.

[16] Fortunato S. and Flammini A., *Int. J. Bifurcat. Chaos*, **17** (2007) 2343.

[17] Boldi P. and Vigna S., *Proceedings of the Thirteenth International World Wide Web Conference (WWW 2004)* (Manhattan, USA, ACM Press) 2004, p. 595–602.

[18] Sommerfeld A., *Partial Differential Equations in Physics* (Academic Press, New York) 1949.

[19] Zinn-Justin J., *Quantum Field Theory and Critical Phenomena* (Oxford University Press, London) 2002.