

JORDAN CANONICAL FORM OF THE GOOGLE MATRIX: A POTENTIAL CONTRIBUTION TO THE PAGERANK COMPUTATION*

STEFANO SERRA-CAPIZZANO†

Abstract. We consider the web hyperlink matrix used by Google for computing the PageRank whose form is given by $A(c) = [cP + (1 - c)E]^T$, where P is a row stochastic matrix, E is a row stochastic rank one matrix, and $c \in [0, 1]$. We determine the analytic expression of the Jordan form of $A(c)$ and, in particular, a rational formula for the PageRank in terms of c . The use of extrapolation procedures is very promising for the efficient computation of the PageRank when c is close or equal to 1.

Key words. Google matrix, canonical Jordan form, extrapolation formulae

AMS subject classifications. 65F10, 65F15, 65Y20

DOI. 10.1137/S0895479804441407

1. Introduction. We look at the web as a huge directed graph whose nodes are all the web pages and whose edges are constituted by all the links between pages. In the following $\deg(i)$ denotes the cardinality of the pages which are reached by a direct link from page i . The basic Google matrix P is defined as $P_{i,j} = 1/\deg(i)$ if $\deg(i) > 0$ and there exists a link in the web from page i to a certain page $j \neq i$; for the rows i for which $\deg(i) = 0$ we assume $P_{i,j} = 1/n$, where n is the size of the matrix, i.e., the cardinality of all the web pages. This definition is a model for the behavior of a generic web user: if the user is visiting page i , then with probability $1/\deg(i)$ he will move to one of the pages j linked by i and if i has no links, then the user will make just a random choice with uniform distribution $1/n$. The basic PageRank is an n -sized vector which gives a measure of the importance of every page in the web: a simple reasoning shows that the basic PageRank is the left eigenvector of P associated to the dominating eigenvalue 1 (see, e.g., [9, 6]). Since the matrix P is nonnegative and has row sum equal to 1 it is clear that the right eigenvector related to 1 is \mathbf{e} (the vector of all 1's) and that all the other eigenvalues are in modulus at most equal to 1. The structure of P is such that we have no guarantee for its aperiodicity and for its irreducibility: therefore the gap between 1 and the modulus of the second largest eigenvalue can be zero. This means that the computation of the PageRank by the application of the standard power method (see, e.g., [3]) to the matrix $A = P^T$ (or one of its variations for our specific problem) is not convergent or is very slowly convergent. A solution is found by considering a change in the model: given a value $c \in [0, 1]$, from the basic Google matrix P we define the parametric Google matrix $P(c)$ as $cP + (1 - c)E$ with $E = \mathbf{e}\mathbf{v}^T$ and $v_i > 0$, $\|\mathbf{v}\|_1 = 1$. This change corresponds to the following user behavior: if the user is visiting page i , then the next move will be with probability c according to the rule described by the basic Google matrix P and with probability $1 - c$ according to the rule described by \mathbf{v} . Generally a value

*Received by the editors February 20, 2004; accepted for publication (in revised form) by M. Chu February 21, 2005; published electronically September 15, 2005. This work was partially supported by MIUR grant 2004015437.

<http://www.siam.org/journals/simax/27-2/44140.html>

†Dipartimento di Fisica e Matematica, Università dell'Insubria, Sede di Como, Via Valleggio 11, 22100 Como, Italy (stefano.serrac@uninsubria.it, serra@mail.dm.unipi.it).

of c as 0.85 is considered in the literature (see, e.g., [6]). For $c < 1$, the good news is that the $\text{PageRank}(c)$, i.e., the left dominating eigenvector, can be computed in a fast way since $P(c)$ (which is now with row sum 1, nonnegative, irreducible, and aperiodic) has a second eigenvalue whose modulus is dominated by c [4]: therefore the convergence to $\text{PageRank}(c)$ is such that the error at step k decays as c^k . Of course the computation becomes slow if c is chosen close to 1 and there is no guarantee of convergence if $c = 1$. In this paper, given the Jordan canonical form of the basic Google matrix P , we describe in an analytical way the Jordan canonical form of the Google matrix $P(c)$ and, in particular, we obtain that

$$\text{PageRank}(c) = \text{PageRank} + R(c)$$

with $R(c)$ the rational vector function of c . Since $\text{PageRank}(c)$ can be computed efficiently when c is far away from 1, the use of vector extrapolation methods [1] should allow us to compute in a fast way $\text{PageRank}(c)$ when c is close or equal to 1 (see [2] for more details).

2. Closed form analysis of $\text{PageRank}(c)$. The analysis is given in two steps: first we give the Jordan canonical form and the rational expression of $\text{PageRank}(c)$ under the assumption that P is diagonalizable; then we consider the general case.

THEOREM 2.1. *Let P be a row stochastic matrix of size n , let $c \in (0, 1)$, and let $E = \mathbf{e}\mathbf{v}^T$ be a row stochastic rank one matrix of size n with \mathbf{e} the vector of all 1's and with \mathbf{v} an n -sized vector representing a probability distribution, i.e., $v_i > 0$ and $\|\mathbf{v}\|_1 = 1$. Consider the matrix $P(c) = cP + (1 - c)E$ and assume that P is diagonalizable. If $P = X\text{diag}(1, \lambda_2, \dots, \lambda_n)X^{-1}$ with $X = [\mathbf{e}|\mathbf{x}_2|\dots|\mathbf{x}_n]$, $[X^{-1}]^T = [\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_n]$, then*

$$P(c) = Z\text{diag}(1, c\lambda_2, \dots, c\lambda_n)Z^{-1}, \quad Z = XR^{-1}.$$

Moreover, the following facts hold true:

- $1 \geq |\lambda_2| \geq \dots \geq |\lambda_n|$ and $\lambda_2 = 1$ if P is reducible and its graph has at least two irreducible closed sets.
- We have

$$R = I_n + \mathbf{e}_1\mathbf{w}^T, \quad \mathbf{w}^T = (0, w_2, \dots, w_n), \\ w_j = (1 - c)\mathbf{v}^T\mathbf{x}_j/(1 - c\lambda_j), \quad j = 2, \dots, n.$$

Proof. From the assumptions we have $P = X\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)X^{-1}$, but $P\mathbf{e} = \mathbf{e}$ (P has row sum equal to 1) and P is nonnegative; therefore $X = [\mathbf{x}_1|\mathbf{x}_2|\dots|\mathbf{x}_n]$ with $\mathbf{x}_1 = \mathbf{e}$, $\lambda_1 = 1$, and $|\lambda_j| \leq 1 = \|P\|_\infty$; moreover, $\lambda_2 = 1$ if P is reducible and its graph has at least two irreducible closed sets by standard Markov theory (see, e.g., [4]). Consequently, $[\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_n]^T X = X^{-1}X = I_n$ and, in particular, $[\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_n]^T \mathbf{e} = X^{-1}\mathbf{e} = \mathbf{e}_1$ (the first vector of the canonical basis). We have

$$P(c) = cP + (1 - c)\mathbf{e}\mathbf{v}^T = X\text{diag}(c, c\lambda_2, \dots, c\lambda_n)X^{-1} + (1 - c)\mathbf{e}\mathbf{v}^T,$$

and hence

$$X^{-1}P(c)X = \text{diag}(c, c\lambda_2, \dots, c\lambda_n) + (1 - c)X^{-1}\mathbf{e}\mathbf{v}^T X \\ = \text{diag}(c, c\lambda_2, \dots, c\lambda_n) + (1 - c)\mathbf{e}_1\mathbf{v}^T X \\ = \text{diag}(c, c\lambda_2, \dots, c\lambda_n) + (1 - c)\mathbf{e}_1[\mathbf{v}^T\mathbf{e}, \mathbf{v}^T\mathbf{x}_2, \dots, \mathbf{v}^T\mathbf{x}_n].$$

But $\mathbf{v}^T \mathbf{e} = 1$ since \mathbf{v} is a probability vector by the hypothesis. In conclusion we have

$$(2.1) \quad X^{-1}P(c)X = \begin{bmatrix} 1 & (1-c)\mathbf{v}^T \mathbf{x}_2 & \cdots & (1-c)\mathbf{v}^T \mathbf{x}_{n-1} & (1-c)\mathbf{v}^T \mathbf{x}_n \\ & c\lambda_2 & & & \\ & & \ddots & & \\ & & & c\lambda_{n-1} & \\ & & & & c\lambda_n \end{bmatrix}.$$

The last step is to diagonalize the previous matrix: calling R the matrix

$$(2.2) \quad \begin{bmatrix} 1 & \frac{(1-c)\mathbf{v}^T \mathbf{x}_2}{(1-c\lambda_2)} & \cdots & \frac{(1-c)\mathbf{v}^T \mathbf{x}_{n-1}}{(1-c\lambda_{n-1})} & \frac{(1-c)\mathbf{v}^T \mathbf{x}_n}{(1-c\lambda_n)} \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

and taking into account (2.1), a direct computation shows that

$$R [X^{-1}P(c)X] = \text{diag}(1, c\lambda_2, \dots, c\lambda_n)R,$$

i.e.,

$$X^{-1}P(c)X = R^{-1} \text{diag}(1, c\lambda_2, \dots, c\lambda_n)R,$$

and finally $P(c) = Z \text{diag}(1, c\lambda_2, \dots, c\lambda_n) Z^{-1}$, $Z = XR^{-1}$. \square

COROLLARY 2.2. *With the notation of Theorem 2.1, the PageRank(c) vector is given by*

$$(2.3) \quad [\text{PageRank}(c)]^T = \mathbf{y}_1^T + (1-c) \sum_{j=2}^n \mathbf{v}^T \mathbf{x}_j \mathbf{y}_j^T / (1-c\lambda_j),$$

where \mathbf{y}_1^T is the basic PageRank vector (i.e., when $c = 1$).

Proof. For $c = 1$ there is nothing to prove since $P(c) = P$ and therefore $\text{PageRank}(c) = \text{PageRank}$. We take $c < 1$. By Theorem 2.1 we have that $\text{PageRank}(c)$ is the transpose of the first row of the matrix $Z^{-1} = RX^{-1}$.

Since $X^{-1} = [\mathbf{y}_1 | \mathbf{y}_2 | \cdots | \mathbf{y}_n]^T$, the claimed thesis follows from the structure of R in (2.2). \square

We now take into account the case where P is general (and therefore possibly not diagonalizable): the conclusions are formally identical except for the rational expression $R(c)$ which is a bit more involved. We first observe that if $\lambda_2 = 1$, then, as proved in [4], the graph of P has at least two irreducible closed sets: therefore the geometric multiplicity of the eigenvalue 1 also must be at least 2. In summary in the general case we have $P = XJX^{-1}$, where

$$(2.4) \quad J = \begin{bmatrix} 1 & & & & \\ & \lambda_2 & * & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & * \\ & & & & \lambda_n \end{bmatrix}$$

with $*$ denoting a value that can be 0 or 1.

THEOREM 2.3. *Let P be a row stochastic matrix of size n , let $c \in (0, 1)$, and let $E = \mathbf{e}\mathbf{v}^T$ be a row stochastic rank one matrix of size n with \mathbf{e} the vector of all 1's and with \mathbf{v} an n -sized vector representing a probability distribution, i.e., $v_i > 0$ and $\|\mathbf{v}\|_1 = 1$. Consider the matrix $P(c) = cP + (1 - c)E$ and let $P = XJ(1)X^{-1}$, $X = [\mathbf{e}|\mathbf{x}_2|\cdots|\mathbf{x}_n]$, $[X^{-1}]^T = [\mathbf{y}_1|\mathbf{y}_2|\cdots|\mathbf{y}_n]$,*

$$J(c) = \begin{bmatrix} 1 & & & & \\ & c\lambda_2 & c \cdot * & & \\ & & \ddots & \ddots & \\ & & & c\lambda_{n-1} & c \cdot * \\ & & & & c\lambda_n \end{bmatrix},$$

and

$$(2.5) \quad J(c) = D \begin{bmatrix} 1 & & & & \\ & c\lambda_2 & * & & \\ & & \ddots & \ddots & \\ & & & c\lambda_{n-1} & * \\ & & & & c\lambda_n \end{bmatrix} D^{-1}, \quad D = \text{diag}(1, c, \dots, c^{n-1}),$$

with $*$ denoting a value that can be 0 or 1. Then we have

$$P(c) = ZJ(c)Z^{-1}, \quad Z = XR^{-1},$$

and, in addition, the following facts hold true:

- $1 \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ and $\lambda_2 = 1$ if P is reducible and its graph has at least two irreducible closed sets.
- We have

$$(2.6) \quad R = I_n + \mathbf{e}_1\mathbf{w}^T, \quad \mathbf{w}^T = (0, w_2, \dots, w_n),$$

$$w_2 = (1 - c)\mathbf{v}^T\mathbf{x}_2 / (1 - c\lambda_2),$$

$$(2.7) \quad w_j = [(1 - c)\mathbf{v}^T\mathbf{x}_j + [J(c)]_{j-1,j}w_{j-1}] / (1 - c\lambda_j), \quad j = 3, \dots, n.$$

Proof. From the assumptions we have $P = X\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)X^{-1}$, but $P\mathbf{e} = \mathbf{e}$ (P has row sum equal to 1) and P is nonnegative: therefore $X = [\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_n]$ with $\mathbf{x}_1 = \mathbf{e}$, $\lambda_1 = 1$, and $|\lambda_j| \leq 1 = \|P\|_\infty$; moreover, $\lambda_2 = 1$ if the graph of P has at least two irreducible closed sets by standard Markov theory (see, e.g., [4]). Consequently, $[\mathbf{y}_1|\mathbf{y}_2|\cdots|\mathbf{y}_n]^T X = X^{-1}X = I_n$ and, in particular, $[\mathbf{y}_1|\mathbf{y}_2|\cdots|\mathbf{y}_n]^T \mathbf{e} = X^{-1}\mathbf{e} = \mathbf{e}_1$ (the first vector of the canonical basis). From the relation

$$P(c) = cP + (1 - c)\mathbf{e}\mathbf{v}^T = X\text{diag}(c, c\lambda_2, \dots, c\lambda_n)X^{-1} + (1 - c)\mathbf{e}\mathbf{v}^T,$$

we deduce

$$\begin{aligned} X^{-1}P(c)X &= cJ(1) + (1 - c)X^{-1}\mathbf{e}\mathbf{v}^T X \\ &= cJ(1) + (1 - c)\mathbf{e}_1\mathbf{v}^T X \\ &= cJ(1) + (1 - c)\mathbf{e}_1[\mathbf{v}^T\mathbf{e}, \mathbf{v}^T\mathbf{x}_2, \dots, \mathbf{v}^T\mathbf{x}_n]. \end{aligned}$$

But $\mathbf{v}^T \mathbf{e} = 1$ since \mathbf{v} is a probability vector by the hypothesis. In summary we infer

$$(2.8) \quad X^{-1}P(c)X = \begin{bmatrix} 1 & (1-c)\mathbf{v}^T \mathbf{x}_2 & \cdots & (1-c)\mathbf{v}^T \mathbf{x}_{n-1} & (1-c)\mathbf{v}^T \mathbf{x}_n \\ & c\lambda_2 & c \cdot * & & \\ & & \ddots & & \\ & & & c\lambda_{n-1} & c \cdot * \\ & & & & c\lambda_n \end{bmatrix}.$$

The last step is to diagonalize the previous matrix: setting R the matrix

$$(2.9) \quad \begin{bmatrix} 1 & w_2 & \cdots & w_{n-1} & w_n \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix},$$

with w_j as in (2.6)–(2.7) and using (2.8), a direct computation shows that

$$R[X^{-1}P(c)X] = J(c)R,$$

i.e.,

$$X^{-1}P(c)X = R^{-1}J(c)R,$$

and therefore $P(c) = ZJ(c)Z^{-1}$, $Z = XR^{-1}$. As a final remark we observe that the identity in (2.5) can be proved by direct inspection. \square

COROLLARY 2.4. *With the notation of Theorem 2.1, the PageRank(c) vector is given by*

$$(2.10) \quad [\text{PageRank}(c)]^T = \mathbf{y}_1^T + \sum_{j=2}^n w_j \mathbf{y}_j^T,$$

where \mathbf{y}_1^T is the basic PageRank vector (i.e., when $c = 1$) and the quantities w_j are expressed as in (2.6)–(2.7).

Proof. By Theorem 2.3, the decomposition

$$P(c) = ZJ(c)Z^{-1}, \quad Z = XR^{-1},$$

with $J(c)$ as in (2.5), is a Jordan decomposition where

$$\text{diag}(1, c^{-1}, \dots, c^{1-n})Z^{-1}$$

is the left eigenvector matrix. Therefore $[\text{PageRank}(c)]^T$ is the first row of the matrix

$$\text{diag}(1, c^{-1}, \dots, c^{1-n})Z^{-1} = \text{diag}(1, c^{-1}, \dots, c^{1-n})RX^{-1}.$$

Since $X^{-1} = [\mathbf{y}_1 | \mathbf{y}_2 | \cdots | \mathbf{y}_n]^T$, the claimed thesis follows from the structure of R in (2.9) and from the fact that the first row of $\text{diag}(1, c^{-1}, \dots, c^{1-n})$ is \mathbf{e}_1^T . \square

- all the remaining eigenvalues of \hat{A} (and hence of A) have modulus weakly dominated by 1 since each matrix $P_{i,i}$ for $i = 1, \dots, r$ is either a null matrix or is strictly substochastic and irreducible;
- the canonical basis of the dominating eigenspace is given by $\{\mathbf{Z}[i] \geq 0, i = 1, \dots, t = m - r\}$ with

$$\hat{A}\mathbf{Z}[i] = \hat{A} \begin{bmatrix} 0 \\ \mathbf{z}[i] \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ P_{r+i,r+i}\mathbf{z}[i] \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{z}[i] \\ 0 \end{bmatrix} = \mathbf{Z}[i]$$

and $\sum_{j=1}^n (\mathbf{Z}[i])_j = 1, (\mathbf{Z}[i])_j \geq 0$, for all $j = 1, \dots, n$, for all $i = 1, \dots, t$. In conclusion we are able to characterize all the nonnegative normalized dominating eigenvectors of the Google matrix as

$$\mathbf{v}(\lambda_1, \dots, \lambda_t) = \sum_{i=1}^t \lambda_i \mathbf{Z}[i], \quad \sum_{i=1}^t \lambda_i = 1, \quad \lambda_i \geq 0.$$

Now if we put the above relations into Corollary 2.4, we deduce that (2.10) can be read as

$$[\text{PageRank}(c)]^T = \mathbf{y}_1^T + (\mathbf{v}^T \mathbf{x}_2) \mathbf{y}_2^T + \dots + (\mathbf{v}^T \mathbf{x}_t) \mathbf{y}_t^T + \sum_{j=t+1}^n w_j \mathbf{y}_j^T,$$

with $w_j = w_j(c)$ and $\lim_{c \rightarrow 1} w_j(c) = 0$. Therefore the unique vector $\text{PageRank}(1) = \mathbf{y}_1 + (\mathbf{v}^T \mathbf{x}_2) \mathbf{y}_2 + \dots + (\mathbf{v}^T \mathbf{x}_t) \mathbf{y}_t$ that we compute is one of the vectors $\mathbf{v}(\lambda_1, \dots, \lambda_t) = \sum_{i=1}^t \lambda_i \mathbf{Z}[i], \sum_{i=1}^t \lambda_i = 1, \lambda_i \geq 0$. The question is, which $\lambda_1, \dots, \lambda_t$? By comparison with Corollary 2.4, the answer is

$$\lambda_j = \sum_{i=1}^t (\mathbf{v}^T \mathbf{x}_i) \alpha_{i,j},$$

where $\alpha = (\alpha_{i,j})_{i,j=1}^t$ is the transformation matrix from the basis $\{\mathbf{Z}[i] \geq 0, i = 1, \dots, t\}$ to the basis $\{\mathbf{y}_i, i = 1, \dots, t\}$, i.e., $\mathbf{y}_i = \sum_{j=1}^t \alpha_{i,j} \mathbf{Z}[j], i = 1, \dots, t$. It is interesting to observe that the vector that we compute, as limit of $\text{PageRank}(c)$, depends on \mathbf{v} , and this is welcome since according to the model, it is correct that the personalization vector \mathbf{v} decides which PageRank vector is chosen.

Acknowledgments. We give warm thanks to the referees for very pertinent and useful remarks and to Michele Benzi, Claude Brezinski, and Michela Redivo Zaglia for interesting discussions.

REFERENCES

[1] C. BREZINSKI AND M. REDIVO ZAGLIA, *Extrapolation Methods. Theory and Practice*, Stud. Comput. Math., North-Holland, Amsterdam, 1991.
 [2] C. BREZINSKI, M. REDIVO ZAGLIA, AND S. SERRA-CAPIZZANO, *Extrapolation methods for Page-Rank computations*, Les Comptes Rendus de l'Academie de Sciences de Paris Ser. 1, 340 (2005), pp. 393–397.
 [3] G.H. GOLUB AND C.F. VAN LOAN, *Matrix Computations*, Johns Hopkins Ser. Math. Sci. 3, The Johns Hopkins University Press, Baltimore, MD, 1983.
 [4] H. HAVELIWALA AND S.D. KAMVAR, *The Second Eigenvalue of the Google Matrix*, Technical report, Stanford University, Stanford, CA, 2003.

- [5] S.D. KAMVAR AND H. HAVELIWALA, *The Condition Number of the PageRank Problem*, Technical report, Stanford University, Stanford, CA, 2003.
- [6] S.D. KAMVAR, H. HAVELIWALA, C.D. MANNING, AND G.H. GOLUB, *Extrapolation methods for accelerating PageRank computations*, in Proceedings of the 12th International WWW Conference, Budapest, Hungary, 2003.
- [7] S.D. KAMVAR, H. HAVELIWALA, C.D. MANNING, AND G.H. GOLUB, *Exploiting the block structure of the Web for computing PageRank*, SCCM TR.03-9, Stanford University, Stanford, CA, 2003.
- [8] C.D. MEYER, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
- [9] L. PAGE, S. BRIN, R. MOTWANI, AND T. WINOGRAD, *The PageRank citation ranking: Bringing order to the web*, Stanford Digital Libraries Working Paper, 1998.