A singular perturbation approach for choosing PageRank damping factor

Konstantin Avrachenkov, Nelly Litvak, and Kim Son Pham

[1] INRIA Sophia Antipolis, 2004, Route des Lucioles, 06902, France;
k.avrachenkov@sophia.inria.fr
[2] University of Twente, Dept. of Applied Mathematics, P.O. Box 217,
7500AE Enschede, The Netherlands; n.litvak@ewi.utwente.nl
[3] St.Petersburg State University,35, University Prospect, 198504, Peterhof, Russia;
sonsecure@yahoo.com.sg

Abstract. We study the PageRank mass of principal components in a bow-tie Web Graph, as a function of the damping factor c. It is known that the Web graph can be divided into three principal components: SCC, IN and OUT. The Giant Strongly Connected Component (SCC) contains a large group of pages all having a hyper-link path to each other. The pages in the IN (OUT) component have a path to (from) the SCC, but not back. Using a singular perturbation approach, we show that the PageRank share of IN and SCC components remains high even for very large values of the damping factor, in spite of the fact that it drops to zero when c tends to one. However, a detailed study of the OUT component reveals the presence of "dead-ends" (small groups of pages linking only to each other) that receive an unfairly high ranking when c is close to one. We argue that this problem can be mitigated by choosing c as small as 1/2.

1 Introduction

The link-based ranking schemes such as PageRank [1], HITS [2], and SALSA [3] have been successfully used in search engines to provide adequate importance measures for Web pages. In the present work we restrict ourselves to the analysis of the PageRank criterion and use the following definition of PageRank from [4]. Denote by n the total number of pages on the Web and define the $n \times n$ hyper-link matrix W as follows:

$$w_{ij} = \begin{cases} 1/d_i, & \text{if page } i \text{ links to } j, \\ 1/n, & \text{if page } i \text{ is dangling,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

for $i, j = 1, \ldots, n$, where d_i is the number of outgoing links from page i. A page is called *dangling* if it does not have outgoing links. The PageRank is defined as a stationary distribution of a Markov chain whose state space is the set of all Web pages, and the transition matrix is

$$G = cW + (1-c)(1/n)\mathbf{1}\mathbf{1}^T. \tag{2}$$

Here and throughout the paper we use the symbol $\mathbf{1}$ for a column vector of ones having by default an appropriate dimension. In (2), $\mathbf{1}\mathbf{1}^T$ is a matrix whose all entries are equal to one, and $c \in (0,1)$ is the parameter known as a *damping factor*. Let π be the PageRank vector. Then by definition, $\pi G = \pi$, and $||\pi|| = \pi\mathbf{1} = 1$, where we write $||\mathbf{x}||$ for the L_1-norm of vector $\mathbf{x}$.

The damping factor c is a crucial parameter in the PageRank definition. It regulates the level of the uniform noise introduced to the system. Based on the publicly available information Google originally used $c = 0.85$, which appears to be a reasonable compromise between the true reflection of the Web structure and numerical efficiency (see [5] for more details). However, it was mentioned in [6] that the value of c too close to one results into distorted ranking of important pages. This phenomenon was also independently observed in [7]. Moreover, with smaller c, the PageRank is more robust, that is, one can bound the influence of outgoing links of a page (or a small group of pages) on the PageRank of other groups [8] and on its own PageRank [7].

In this paper we explore the idea of relating the choice of c to specific properties of the Web structure. In papers [9,10] the authors have shown that the Web graph can be divided into three principal components. The Giant Strongly Connected Component (SCC) contains a large group of pages all having a hyperlink path to each other. The pages in the IN (OUT) component have a path to (from) the SCC, but not back. Furthermore, the SCC component is larger than the second largest strongly connected component by several orders of magnitude.

In Section 3 we consider a Markov walk governed by the hyperlink matrix W and explicitly describe the limiting behavior of the PageRank vector as $c \to 1$ with the help of the singular perturbation theory [11,12,13,14]. We experimentally study the OUT component in more detail to discover a so-called Pure OUT component (the OUT component without dangling nodes and their predecessors) and show that Pure OUT contains a number of small sub-SCC's, or dead-ends, that absorb the total PageRank mass when $c = 1$. In Section 4 we analyze the shape of the PageRank of IN+SCC as a function of c. The dangling nodes turn out to play an unexpectedly important role in the qualitative behavior of this function. Our analytical and experimental results suggest that the PageRank mass of IN+SCC is sustained on a high level for quite large values of c, in spite of the fact that it drops to zero as $c \to 1$. Furthermore, the PageRank mass of IN+SCC has a unique maximum. Then, in Section 5 we show that the total PageRank mass of Pure OUT component increases with c. We argue that $c = 0.85$ results in an inadequately high ranking for Pure OUT pages and we present an argument based on the singular perturbation theory for choosing c as small as $1/2$. We confirm our theoretical argument by experiments with log files. We would like to mention that the value $c = 1/2$ was also used in [15] to find gems in scientific citations. This choice was justified intuitively by stating that researchers may check references in cited papers but on average they hardly go deeper than two levels. Nowadays, when search engines work really fast, this argument also applies to Web search. Indeed, it is easier for the user to refine a query and receive a more relevant page in fraction of a second than to look for this page by clicking on hyper-links. Therefore, we may assume that a surfer searching for a page, on average, does not go deeper than two clicks.

2 Datasets

We have collected two Web graphs, which we denote by INRIA and FrMathInfo. The Web graph INRIA was taken from the site of INRIA, the French Research Institute of Informatics and Automatics. The seed for the INRIA collection was Web page `www.inria.fr`. It is a typical large Web site with around 300.000 pages and 2 millions hyper-links. We have collected all pages belonging to INRIA. The Web graph FrMathInfo was crawled with the initial seeds of 50 mathematics and informatics laboratories of France, taken from Google Directory. The crawl was executed by Breadth First Search of depth 6. The FrMathInfo Web graph contains around 700.000 pages and 8 millions hyper-links. As the Web seems to have a fractal structure [16], we expect our datasets to be enough representative.

The link structure of the two Web graphs is stored in Oracle database. We could store the adjacency lists in RAM to speed up the computation of PageRank and other quantities of interest. This enables us to make more iterations, which is extremely important when the damping factor c is close to one. Our PageRank computation program consumes about one hour to make 500 iterations for the FrMathInfo dataset and about half an hour for the INRIA dataset for the same number of iterations. Our algorithms for discovering the structures of the Web graph are based on Breadth First Search and Depth First Search methods, which are linear in the sum of number of nodes and links.

3 The structure of the hyper-link transition matrix

Let us refine the bow-tie structure of the Web graph [9,10]. We recall that the transition matrix W induces artificial links to all pages from dangling nodes. Obviously, the graph with many artificial links has a much higher connectivity than the original Web graph. In particular, if the random walk can move from a dangling node to an arbitrary node with the uniform distribution, then the Giant SCC component increases further in size. We refer to this new strongly connected component as the Extended Strongly Connected Component (ESCC). Due to the artificial links from the dangling nodes, the SCC component and IN component are now inter-connected and are parts of the ESCC. Furthermore, if there are dangling nodes in the OUT component, then these nodes together with all their predecessors become a part of the ESCC.

In the mini-example in Figure 1, node 0 represents the IN component, nodes from 1 to 3 form the SCC component, and the rest of the nodes, nodes from 4 to 11, are in the OUT component. Node 5 is a dangling node, thus, artificial links go from the dangling node 5 to all other nodes. After addition of the artificial links, all nodes from 0 to 5 form the ESCC.

The part of the OUT component without dangling nodes and their predecessors forms a block that we refer to as a Pure OUT component. In Figure 1 the

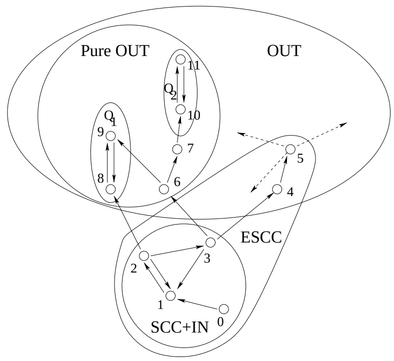

Fig. 1. Example of a graph

#	$INRIA$	$FrMathInfo$
total nodes	318585	764119
nodes in SCC	154142	333175
nodes in IN	0	0
nodes in OUT	164443	430944
nodes in ESCC	300682	760016
nodes in Pure OUT	17903	4103
SCCs in OUT	1148	1382
SCCs in Pure Out	631	379

Fig. 2. Component sizes in INRIA and FrMathInfo datasets

Pure OUT component consists of nodes from 6 to 11. Typically, the Pure OUT component is much smaller than the Extended SCC.

The sizes of all components for our two datasets are given in Figure 2. Here the size of the IN components is zero because in the Web crawl we used the Breadth First Search method and we started from important pages in the Giant SCC. For the purposes of the present research it does not make any difference since we always consider IN and SCC together.

Let us now analyze the structure of the Pure OUT component in more detail. It turns out that inside Pure OUT there are many disjoint strongly connected components. We refer to these sub-SCC's as "dead-ends", since once the random walk induced by transition matrix W enters such a component it will not be able to leave the component. In Figure 1 there are two dead-end components $\{8, 9\}$ and $\{10, 11\}$. We have observed that in our two datasets the majority of dead-ends are of sizes 2 or 3.

Let us now characterize the new refined structure of the Web graph in terms of ergodic structure of the Markov chain induced by the matrix W. First, we note that all states in the dead-ends are *recurrent*, that is, the Markov chain started from any of these states always returns back to it. In contrast, all the states from ESCC are *transient*, that is, with probability 1, the Markov chain induced by W eventually leaves this set of states and never returns back. The stationary probability of all these states is zero. We note that the Pure OUT component also contains transient states that eventually bring the random walk into one of the dead-ends. For simplicity, we add these states to the giant transient ESCC component.

Now, by appropriate renumbering of the states, we can refine the matrix W by subdividing all states into one giant transient block and a number of small recurrent blocks as follows:

$$W = \begin{bmatrix} Q_1 & & 0 & 0 \\ & \ddots & & \\ 0 & & Q_m & 0 \\ R_1 & \cdots & R_m & T \end{bmatrix} \begin{array}{l} \text{dead-end (recurrent)} \\ \cdots \\ \text{dead-end (recurrent)} \\ \text{ESCC+[transient states in Pure OUT] (transient)} \end{array} \qquad (3)$$

Here for $i = 1, \ldots, m$, a block Q_i corresponds to transitions inside the i-th recurrent block, and a block R_i contains transition probabilities from transient states to the i-th recurrent block. Block T corresponds to transitions between the transient states. For instance, in example of the graph from Figure 1, the nodes 8 and 9 correspond to block Q_1, nodes 10 and 11 correspond to block Q_2, and all other nodes belong to block T. Let us denote by $\bar{\pi}_{\text{OUT},i}$ the stationary distribution corresponding to block Q_i.

We would like to emphasis that the recurrent blocks here are really small, constituting altogether about 5% for INRIA and about 0.5% for FrMathInfo. We believe that for larger data sets, this percentage will be even less. By far most important part of the pages is contained in the ESCC, which constitutes the major part of the giant transient block. However, if the random walk is governed by transition matrix W it is absorbed with probability 1 into one of the recurrent blocks.

The use of the Google transition matrix G with $c < 1$ (2) instead of W ensures that all the pages are recurrent states with positive stationary probabilities. However, if $c = 1$, the majority of pages turn into transient states with stationary probability zero. Hence, the random walk governed by the Google transition matrix G is in fact a singularly perturbed Markov chain. Informally, by singular perturbation we mean relatively small changes in elements of the matrix, that lead to altered connectivity and stationary behavior of the chain. Using the results of the singular perturbation theory (see e.g., [11,12,13,14]), in the next proposition we characterize explicitly the limiting PageRank vector as $c \to 1$.

Proposition 1. *Let $\bar{\pi}_{OUT,i}$ be a stationary distribution of the Markov chain governed by Q_i, $i = 1, \ldots, m$. Then, we have*

$$\lim_{c \to 1} \pi(c) = [\pi_{OUT,1} \;\; \cdots \;\; \pi_{OUT,m} \;\; \mathbf{0}] \,,$$

where

$$\pi_{OUT,i} = \left(\frac{\#\text{ nodes in block } Q_i}{n} + \frac{1}{n} \mathbf{1}^T [I - T]^{-1} R_i \mathbf{1} \right) \bar{\pi}_{OUT,i} \qquad (4)$$

for $i = 1, ..., m$, I is the identity matrix, and $\mathbf{0}$ is a row vector of zeros that correspond to stationary probabilities of the states in the transient block.

Proof. First, we note that if we make a change of variables $\varepsilon = 1 - c$ the Google matrix becomes a transition matrix of a singularly perturbed Markov chain as in Lemma 1 (see Appendix) with $A = W$ and $C = \frac{1}{n}\mathbf{1}\mathbf{1}^T - W$. Specifically, $A_i = Q_i$,

$L_i = R_i$, $E = T$ and $\mu_i = \bar{\pi}_{\mathrm{OUT},i}$. Next, define the aggregated generator matrix D as follows:

$$D = \frac{1}{n}\mathbf{1}\mathbf{1}^T B - I = \frac{1}{n}\mathbf{1}[n_1 + \mathbf{1}[I-T]^{-1}R_1\mathbf{1}, \cdots, n_m + \mathbf{1}[I-T]^{-1}R_m\mathbf{1}] - I. \tag{5}$$

Using the definition of C together with identities $\bar{\pi}_{\mathrm{OUT},i}(1/n)\mathbf{1}\mathbf{1}^T = (1/n)\mathbf{1}\mathbf{1}^T$ and $\bar{\pi}_{\mathrm{OUT},i}Q_i = \bar{\pi}_{\mathrm{OUT},i}$, it is easy to verify that the matrix D in (5) has been computed in exactlly the same way as the matrix D in Lemma 1. Furthermore, since the aggregated transition matrix $D + I$ has identical rows, its stationary distribution ν is simply equal to each of these rows. Thus, invoking Lemma 1 we obtain (4).

The second term inside the brackets in formula (4) corresponds to the PageRank mass received by a dead-end from the Extended SCC. If c is close to one, then this contribution can outweight by far the fair share of the PageRank, whereas the PageRank mass of the giant transient block decreases to zero. How large is the neighborhood of one where the ranking is skewed towards the Pure OUT? Is the value $c = 0.85$ already too large? We will address these questions in the remainder of the paper. In the next section we analyze the PageRank mass IN+SCC component, which is an important part of the transient block.

4 PageRank mass of IN+SCC

In Figure 3 we depict the PageRank mass of the giant component IN+SCC, as a function of the damping factor, for FrMathInfo. Here we see a typical behavior

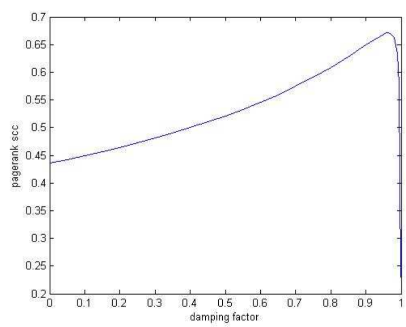

Fig. 3. The PageRank mass of IN+SCC as a function of c.

also observed for several pages in the mini-web from [6]: the PageRank first grows with c and then decreases to zero. In our case, the PageRank mass of IN+SCC drops drastically starting from some value c close to one. Our goal now is to explain this behavior. Clearly, since IN+SCC is a part of the transient block, we do expect that the corresponding PageRank mass drops to zero when c goes to one. Thus, the two phenomena that remain to be justified are: the growth of the PageRank mass when c is not too large, and the abrupt drop to zero after reaching a (unique) extreme point.

The plan of the analysis in this section is as follows. First, we write the expression for the PageRank mass of IN+SCC, $||\pi_{\text{IN+SCC}}||$, as a function of c. Then we consider the derivative of $||\pi_{\text{IN+SCC}}(c)||$ at $c=0$ and prove that, surprisingly, this derivative is always positive in the graph with sufficiently large fraction of dangling nodes. This explains the fact that $||\pi_{\text{IN+SCC}}(c)||$ is initially increasing. Further, we use singular perturbation theory to show that the derivative of $||\pi_{\text{IN+SCC}}(c)||$ at $c=1$ is a large negative number, and that $||\pi_{\text{IN+SCC}}(c)||$ can have only one extreme point in $(0,1)$.

We base our analysis on the model where the Web graph sample is subdivided into three subsets of nodes: IN+SCC, OUT, and the set of dangling nodes DN. We assume that all links to dangling nodes come from IN+SCC. This simplifies the derivation but does not change our conclusions. Then the Web hyper-link matrix W in (1) can be written in the form

$$W = \begin{array}{c} \left[\begin{array}{ccc} Q & 0 & 0 \\ R & P & S \\ \frac{1}{n}\mathbf{1}\mathbf{1}^T & \frac{1}{n}\mathbf{1}\mathbf{1}^T & \frac{1}{n}\mathbf{1}\mathbf{1}^T \end{array}\right] \end{array} \begin{array}{l} \text{OUT} \\ \text{IN+SCC} \\ \text{DN} \end{array},$$

where the block Q corresponds to the hyper-links inside the OUT component, the block R corresponds to the hyper-links from IN+SCC to OUT, the block P corresponds to the hyper-links inside the IN+SCC component, and the block S corresponds to the hyper-links from SCC to dangling nodes. In the above, n is the total number of pages in the Web graph sample, and the blocks $\mathbf{1}\mathbf{1}^T$ are the matrices of ones adjusted to appropriate dimensions.

Let us derive the expression for the PageRank mass of IN+SCC. Dividing the PageRank vector in segments corresponding to the blocks OUT, IN+SCC and DN, namely, $\pi = [\pi_{\text{OUT}}\ \pi_{\text{IN+SCC}}\ \pi_{\text{DN}}]$, we can rewrite the well-known formula (see e.g. [17])

$$\pi = \frac{1-c}{n}\mathbf{1}^T[I-cW]^{-1} \tag{6}$$

as a system of three linear equations:

$$\pi_{\text{OUT}}[I-cQ] - \pi_{\text{IN+SCC}}cR - \frac{c}{n}\pi_{\text{DN}}\mathbf{1}\mathbf{1}^T = \frac{1-c}{n}\mathbf{1}^T, \tag{7}$$

$$\pi_{\text{IN+SCC}}[I-cP] - \frac{c}{n}\pi_{\text{DN}}\mathbf{1}\mathbf{1}^T = \frac{1-c}{n}\mathbf{1}^T, \tag{8}$$

$$-\pi_{\text{IN+SCC}}cS + \pi_{\text{DN}} - \frac{c}{n}\pi_{\text{DN}}\mathbf{1}\mathbf{1}^T = \frac{1-c}{n}\mathbf{1}^T. \tag{9}$$

Now we would like to solve (7–9) for $\pi_{\text{IN+SCC}}$. To this end, we first observe that if $\pi_{\text{IN+SCC}}$ and $\pi_{\text{DN}}\mathbf{1}$ are known then from (7) it is straightforward to obtain π_{OUT}:

$$\pi_{\text{OUT}} = \pi_{\text{IN+SCC}}cR[I-cQ]^{-1} + \left(\frac{1-c}{n} + \pi_{\text{DN}}\mathbf{1}\frac{c}{n}\right)\mathbf{1}^T[I-cQ]^{-1}.$$

Therefore, let us solve equations (8) and (9). Towards this goal, we sum the elements of the vector equation (9), which corresponds to the postmultiplication of equation (9) by vector $\mathbf{1}$.

$$-\pi_{\text{IN+SCC}} cS\mathbf{1} + \pi_{\text{DN}}\mathbf{1} - \frac{c}{n}\pi_{\text{DN}}\mathbf{1}\mathbf{1}^T\mathbf{1} = \frac{1-c}{n}\mathbf{1}^T\mathbf{1}$$

Now, denote by n_{IN}, n_{OUT}, n_{SCC} and n_{DN} the number of pages in the IN component, OUT component, SCC component and the number of dangling nodes. Since $\mathbf{1}^T\mathbf{1} = n_{DN}$, we have

$$\pi_{\text{DN}}\mathbf{1} = \frac{n}{n - cn_{DN}}(\pi_{\text{IN+SCC}} cS\mathbf{1} + \frac{1-c}{n}n_{DN}).$$

Substituting the above expression for $\pi_{\text{DN}}\mathbf{1}$ into (8), we get

$$\pi_{\text{IN+SCC}}\left[I - cP - \frac{c^2}{n - cn_{DN}}S\mathbf{1}\mathbf{1}^T\right] = \frac{c}{n - cn_{DN}}\frac{1-c}{n}n_{DN}\mathbf{1}^T + \frac{1-c}{n}\mathbf{1}^T.$$

Denote by $\alpha = (n_{IN} + n_{SCC})/n$ and $\beta = n_{DN}/n$ the fractions of nodes in IN+SCC and DN, respectively, and let $\mathbf{u}_{\text{IN+SCC}} = (n_{IN} + n_{SCC})^{-1}\mathbf{1}^T$ be a uniform probability row-vector of dimension $n_{IN} + n_{SCC}$. Then from the last equation we directly obtain

$$\pi_{\text{IN+SCC}}(c) = \frac{(1-c)\alpha}{1-c\beta}\mathbf{u}_{\text{IN+SCC}}\left[I - cP - \frac{c^2\alpha}{1-c\beta}S\mathbf{1}\mathbf{u}_{\text{IN+SCC}}\right]^{-1}. \qquad (10)$$

Equation (10) gives the desired expression for the PageRank mass of IN+SCC as a function of c, and we can analyze the behavior of this function by looking at its derivatives. Define

$$k(c) = \frac{(1-c)\alpha}{1-c\beta}, \quad \text{and} \quad U(c) = P + \frac{c\alpha}{1-c\beta}S\mathbf{1}\mathbf{u}_{\text{IN+SCC}}. \qquad (11)$$

Then the derivative of $\pi_{\text{IN+SCC}}(c)$ with respect to c is given by

$$\pi'_{\text{IN+SCC}}(c) = \mathbf{u}_{\text{IN+SCC}}\left\{k'(c)I + k(c)[I - cU(c)]^{-1}(cU(c))'\right\}[I - cU(c)]^{-1}, \quad (12)$$

where from (11) after simple calculations we get $k'(c) = -(1-\beta)\alpha/(1-c\beta)^2$, $(cU(c))' = U(c) + c\alpha(1-c\beta)^{-2}S\mathbf{1}\mathbf{u}_{\text{IN+SCC}}$.

Now we are ready to explain the fact that $||\pi_{\text{IN+SCC}}(c)||$ is increasing when c is small. Consider the point $c = 0$. Using (12), we get

$$\pi'_{\text{IN+SCC}}(0) = -\alpha(1-\beta)\mathbf{u}_{\text{IN+SCC}} + \alpha\mathbf{u}_{\text{IN+SCC}}P. \qquad (13)$$

One can see from the above equation that the PageRank of pages in IN+SCC with many incoming links will increase as c increases from zero, which explains the graphs presented in [6]. Next, for the total mass of the IN+SCC component, from (13) we obtain

$$||\pi'_{\text{IN+SCC}}(0)|| = -\alpha(1-\beta)\mathbf{u}_{\text{IN+SCC}} + \alpha\mathbf{u}_{\text{IN+SCC}}P\mathbf{1} = \alpha(-1+\beta+p_1),$$

where $p_1 = \mathbf{u}_{\text{IN+SCC}} P\mathbf{1}$ is the probability that a random walk on the hyperlink matrix stays in IN+SCC for one step if the initial distribution is uniform over IN+SCC. If $1-\beta < p_1$ then the derivative at 0 is positive. Since dangling nodes typically constitute more than 25% of the graph [18], and p_1 is usually close to one, the condition $1-\beta < p_1$ seems to be comfortably satisfied in Web samples. Thus, the total PageRank of the IN+SCC increases in c when c is small. Note by the way that if $\beta = 0$ then $||\pi_{\text{IN+SCC}}(c)||$ is strictly decreasing in c. Hence, surprisingly, the presence of dangling nodes qualitatively changes the behavior of the IN+SCC PageRank mass.

Now let us consider the point $c = 1$. Again using (12), we get

$$\pi'_{\text{IN+SCC}}(1) = -\frac{\alpha}{1-\beta}\mathbf{u}_{\text{IN+SCC}}[I - P - \frac{\alpha}{1-\beta}S\mathbf{1}\mathbf{u}_{\text{IN+SCC}}]^{-1}. \tag{14}$$

We will show that the derivative above is a negative number with a large absolute value. Note that the matrix in the square braces is close to singular. Denote by $\bar{P}$ the hyper-link matrix of IN+SCC when the outer links are neglected. Then, $\bar{P}$ is an irreducible stochastic matrix. Denote its stationary distribution by $\bar{\pi}_{\text{IN+SCC}}$. Then we can apply Lemma 2 (see Appendix) from the singular perturbation theory to (14) by taking $A = \bar{P}$, $\varepsilon C = \bar{P} - P - \alpha(1-\beta)^{-1}S\mathbf{1}\mathbf{u}_{\text{IN+SCC}}$, and noting that

$$\varepsilon C\mathbf{1} = R\mathbf{1} + (1-\alpha-\beta)(1-\beta)^{-1}S\mathbf{1}.$$

Combining all terms together and using $\bar{\pi}_{\text{IN+SCC}}\mathbf{1} = ||\bar{\pi}_{\text{IN+SCC}}|| = 1$ and $\mathbf{u}_{\text{IN+SCC}}\mathbf{1} = ||\mathbf{u}_{\text{IN+SCC}}|| = 1$, by Lemma 2 we obtain

$$||\pi'_{\text{IN+SCC}}(1)|| \approx -\frac{\alpha}{1-\beta}\frac{1}{\bar{\pi}_{\text{IN+SCC}}R\mathbf{1} + \frac{1-\beta-\alpha}{1-\beta}\bar{\pi}_{\text{IN+SCC}}S\mathbf{1}}.$$

It is expected that the value in the denominator of the second fraction is typically small (indeed, in our dataset INRIA, the value is 0.022), and hence the mass $||\pi_{\text{IN+SCC}}(c)||$ decreases very fast as c approaches one.

Having described the behavior of the PageRank mass $||\pi_{\text{IN+SCC}}(c)||$ at the boundary points $c = 0$ and $c = 1$, now we would like to show that there is at most one extremum on $(0,1)$. It is sufficient to prove that if $||\pi'_{\text{IN+SCC}}(c_0)|| \le 0$ for some $c_0 \in (0,1)$ then $||\pi'_{\text{IN+SCC}}(c)|| \le 0$ for all $c > c_0$. To this end, we apply the Sherman-Morrison formula to (10), which yields

$$\pi_{\text{IN+SCC}}(c) = \tilde{\pi}_{\text{IN+SCC}}(c) + \frac{\frac{c^2\alpha}{1-c\beta}\mathbf{u}_{\text{IN+SCC}}[I-cP]^{-1}S\mathbf{1}}{1 + \frac{c^2\alpha}{1-c\beta}\mathbf{u}_{\text{IN+SCC}}[I-cP]^{-1}S\mathbf{1}}\,\tilde{\pi}_{\text{IN+SCC}}(c), \tag{15}$$

where

$$\tilde{\pi}_{\text{IN+SCC}}(c) = \frac{(1-c)\alpha}{1-c\beta}\mathbf{u}_{\text{IN+SCC}}[I-cP]^{-1} \tag{16}$$

represents the main term in the right-hand side of (15). (The second summand in (15) is about 10% of the total sum for the INRIA dataset for $c = 0.85$.) Now the behavior of $\pi_{\text{IN+SCC}}(c)$ in Figure 3 can be explained by means of the next proposition.

Proposition 2. *The term* $||\tilde{\pi}_{IN+SCC}(c)||$ *given by (16) has exactly one local maximum at some* $c_0 \in [0,1]$*. Moreover,* $||\tilde{\pi}''_{IN+SCC}(c)|| < 0$ *for* $c \in (c_0, 1]$*.*

Proof. Multiplying both sides of (16) by $\mathbf{1}$ and taking the derivatives, after some tedious algebra we obtain

$$||\tilde{\pi}'_{\text{IN+SCC}}(c)|| = -\,a(c) + \frac{\beta}{1-c\beta}\,||\tilde{\pi}_{\text{IN+SCC}}(c)||, \tag{17}$$

where the real-valued function $a(c)$ is given by

$$a(c) = \frac{\alpha}{1-c\beta}\mathbf{u}_{\text{IN+SCC}}[I-cP]^{-1}[I-P][I-cP]^{-1}\mathbf{1}.$$

Differentiating (17) and substituting $\frac{\beta}{1-c\beta}||\tilde{\pi}_{\text{IN+SCC}}(c)||$ from (17) in the resulting expression, we get

$$||\tilde{\pi}''_{\text{IN+SCC}}(c)|| = \left\{-a'(c) + \frac{\beta}{1-c\beta}a(c)\right\} + \frac{2\beta}{1-c\beta}\,||\tilde{\pi}'_{\text{IN+SCC}}(c)||.$$

Note that the term in the curly braces is negative by definition of $a(c)$. Hence, if $||\tilde{\pi}'_{\text{IN+SCC}}(c)|| \le 0$ for some $c \in [0,1]$ then $||\tilde{\pi}''_{\text{IN+SCC}}(c)|| < 0$ for this value of c.

We conclude that $||\tilde{\pi}_{\text{IN+SCC}}(c)||$ is decreasing and concave for $c \in [c_0, 1]$, where $||\tilde{\pi}'_{\text{IN+SCC}}(c_0)|| = 0$. This is exactly the behavior we observe in the experiments. The analysis and experiments suggest that c_0 is definitely larger than 0.85 and actually is quite close to one. Thus, one may want to choose large c in order to maximize the PageRank mass of IN+SCC. However, in the next section we will indicate important drawbacks of this choice.

5 PageRank mass of ESCC

Let us now consider the PageRank mass of the Extended SCC component (ESCC) described in Section 3, as a function of $c \in [0,1]$. Subdividing the PageRank vector in the blocks $\pi = [\pi_{\text{PureOUT}}\ \pi_{\text{ESCC}}]$, from (6) we obtain

$$\pi_{\text{ESCC}}(c) = (1-c)\gamma\mathbf{u}_{\text{ESCC}}[I-cT]^{-1} = (1-c)\gamma\mathbf{u}_{\text{ESCC}}\sum_{k=1}^{\infty} c^k T^k, \tag{18}$$

where T represents the transition probabilities inside the ESCC block, $\gamma = |ESCC|/n$ is the fraction of pages contained in the ESCC, and $\mathbf{u}_{\text{ESCC}}$ is a uniform probability row-vector over ESCC. Clearly, we have $||\pi_{\text{ESCC}}(0)|| = \gamma$ and $||\pi_{\text{ESCC}}(1)|| = 0$. Furthermore, it is easy to see that $||\pi_{\text{ESCC}}(c)||$ is a concave decreasing function, since

$$\frac{d}{dc}||\pi_{\text{ESCC}}(c)|| = -\gamma\mathbf{u}_{\text{ESCC}}[I-cT]^{-2}[I-T]\mathbf{1} < 0$$

and

$$\frac{d^2}{dc^2}||\pi_{\text{ESCC}}(c)|| = -2\gamma\mathbf{u}_{\text{ESCC}}[I-cT]^{-3}T[I-T]\mathbf{1} < 0.$$

The next proposition establishes the upper and lower bounds for $||\pi_{\text{ESCC}}(c)||$.

Proposition 3. *Let λ_1 be the Perron-Frobenius eigenvalue of T, and let $p_1 = \mathbf{u}_{ESCC}T\mathbf{1}$ be the probability that the random walk started from a randomly chosen state in ESCC, stays in ESCC for one step. If $p_1 \le \lambda_1$ and*

$$p_1 \le \frac{\mathbf{u}_{ESCC}T^k\mathbf{1}}{\mathbf{u}_{ESCC}T^{k-1}\mathbf{1}} \le \lambda_1 \quad \textit{for all } k \ge 1, \tag{19}$$

then

$$\frac{\gamma(1-c)}{1-cp_1} < ||\pi_{ESCC}(c)|| < \frac{\gamma(1-c)}{1-c\lambda_1}, \quad c \in (0,1). \tag{20}$$

Proof. From condition (19) it follows by induction that

$$p_1^k \le \mathbf{u}_{\mathrm{ESCC}}T^k\mathbf{1} \le \lambda_1^k, \quad k \ge 1,$$

and thus the statement of the proposition is obtained directly from the series expansion of $\pi_{\mathrm{ESCC}}(c)$ in (18).

The conditions of Proposition 3 have a natural probabilistic interpretation. The value p_1 is the probability that the Markov random walk on the Web sample stays in the block T for one step, starting from the uniform distribution over T. Furthermore, $p_k = \mathbf{u}_{\mathrm{ESCC}}T^k\mathbf{1}/(\mathbf{u}_{\mathrm{ESCC}}T^{k-1}\mathbf{1})$ is the probability that the random walk stays in T for one step provided that it has stayed there for the first $k-1$ steps. It is a well-known fact that, as $k \to \infty$, p_k converges to λ_1, the Perron-Frobenius eigenvalue of T. Let $\hat{\pi}_{\mathrm{ESCC}}$ be the probability-normed left Perron-Frobenius eigenvector of T. Then $\hat{\pi}_{\mathrm{ESCC}}$, also known as a *quasi-stationary* distribution of T, is the limiting probability distribution of the Markov chain given that the random walk never leaves the block T (see e.g. [19]). Since $\hat{\pi}_{\mathrm{ESCC}}T = \lambda_1\hat{\pi}_{\mathrm{ESCC}}$, the condition $p_1 < \lambda_1$ means that the chance to stay in ESCC for one step in the quasi-stationary regime is higher than starting from the uniform distribution $\mathbf{u}_{\mathrm{ESCC}}$. This is quite natural since the quasi-stationary distribution tends to avoid the states, from which the random walk is likely to leave the block T. Furthermore, the condition in (19) says that if the random walk is about to make its k-th steps in T, then it leaves T most easily at step $k = 1$, and it is most hard to leave T after an infinite number of steps. Both conditions of Proposition 3 are satisfied in our experiments on both data sets. Moreover, we noticed that the sequence $(p_k, k \ge 1)$ was increasing from p_1 to λ_1.

With the help of the derived bounds we conclude that $||\pi_{\mathrm{ESCC}}(c)||$ decreases very slowly for small and moderate values of c, and it decreases extremely fast when c becomes close to 1. This typical behavior is clearly seen in Figure 4, where $||\pi_{\mathrm{ESCC}}(c)||$ is plotted with a solid line. The bounds are plotted in Figure 4 with dashed lines. For the INRIA dataset we have $p_1 = 0.97557$, $\lambda_1 = 0.99954$, and for the FrMathInfo dataset we have $p_1 = 0.99659$, $\lambda_1 = 0.99937$.

From the above we conclude that the PageRank mass of ESCC is smaller than γ for any value $c > 0$. On contrary, the PageRank mass of Pure OUT increases in c beyond its "fair share" $\delta = |PureOUT|/n$. With $c = 0.85$, the PageRank mass of the Pure OUT component in the INRIA dataset is equal to

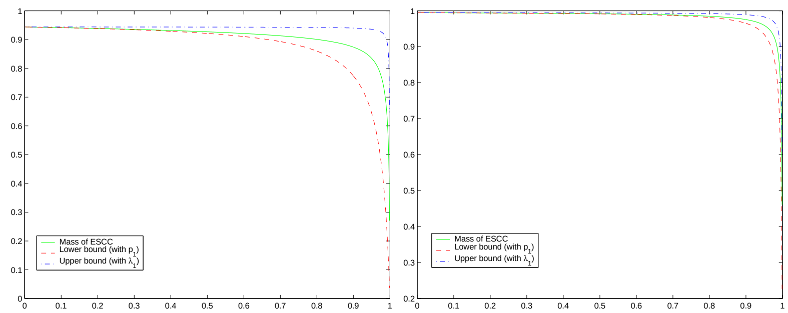

Fig. 4. PageRank mass of ESCC and bounds, INRIA (left) and FrMathInfo (right)

1.95δ. In the FrMathInfo dataset, the unfairness is even more pronounced: the PageRank mass of the Pure OUT component is equal to 3.44δ. This gives users an incentive to create dead-ends: groups of pages that link only to each other. Clearly, this can be mitigated by choosing a smaller damping factor. Below we propose one way to determine an "optimal" value of c.

Since the PageRank mass of ESCC is always smaller than γ we would like to choose the damping factor in such a way that the ESCC receives a "fair" fraction of γ. Formally, we would like to define a number $\rho \in (0,1)$ such that a desirable PageRank mass of ESCC could be written as $\rho\gamma$, and then find the value c^* that satisfies

$$||\pi_{\text{ESCC}}(c^*)|| = \rho\gamma. \tag{21}$$

Then $c \le c^*$ will ensure that $||\pi_{\text{ESCC}}(c)|| \ge \rho\gamma$. Naturally, ρ should somehow reflect the properties of the substochastic block T. For instance, as T becomes closer to stochastic matrix, ρ should also increase. One possibility to do it is to define

$$\rho = \mathbf{v}T\mathbf{1},$$

where $\mathbf{v}$ is a row vector representing some probability distribution on ESCC. Then the damping factor c should satisfy

$$c \le c^*,$$

where c^* is given by

$$||\pi_{\text{ESCC}}(c^*)|| = \gamma\mathbf{v}T\mathbf{1}. \tag{22}$$

In this setting, ρ is a probability to stay in ESCC for one step if initial distribution is $\mathbf{v}$. For given $\mathbf{v}$, this number increases as T becomes closer to stochastic matrix. Now, the problem of choosing ρ comes down to the problem of choosing $\mathbf{v}$. The advantage of this approach is twofold. First, we still have all the flexibility because, depending on $\mathbf{v}$, the value of ρ may vary considerably except it can not become too small if T is really close to stochastic matrix. Second, we can use a probabilistic interpretation of $\mathbf{v}$ to make a reasonable choice. One can think for instance of the following three intuitive choices of $\mathbf{v}$: 1) $\hat{\pi}_{\text{ESCC}}$, the quasi-stationary distribution of T, 2) the uniform vector $\mathbf{u}_{\text{ESCC}}$, and 3) the normalized

PageRank vector $\pi_{\text{ESCC}}(c)/||\pi_{\text{ESCC}}(c)||$. The first choice reflects the proximity of T to a stochastic matrix. The second choice is inspired by definition of PageRank (restart from uniform distribution), and the third choice combines both these features.

If conditions of Proposition 3 are satisfied, then (20) holds, and thus the value of c^* satisfying (22) must be in the interval (c_1, c_2), where

$$(1-c_1)/(1-p_1c_1) = ||\mathbf{v}T||, \quad (1-c_2)/(1-\lambda_1 c_2) = ||\mathbf{v}T||.$$

Numerical results for all three choices of $\mathbf{v}$ are presented in Table 1.

v	c	INRIA	FrMathInfo
$\hat{\pi}_{\text{ESCC}}$	c_1	0.0184	0.1956
	c_2	0.5001	0.5002
	c^*	.02	.16
$\mathbf{u}_{\text{ESCC}}$	c_1	0.5062	0.5009
	c_2	0.9820	0.8051
	c^*	.604	.535
$\pi_{\text{ESCC}}/\vert\vert\pi_{\text{ESCC}}\vert\vert$	$1/(1+\lambda_1)$	0.5001	0.5002
	$1/(1+p_1)$	0.5062	0.5009

Table 1. Values of c^* with bounds.

If $\mathbf{v} = \hat{\pi}_{\text{ESCC}}$ then we have $||\mathbf{v}T|| = \lambda_1$, which implies $c_1 = (1-\lambda_1)/(1-\lambda_1 p_1)$ and $c_2 = 1/(\lambda_1+1)$. In this case, the upper bound c_2 is only slightly larger than 1/2 and c^* is close to zero in our data sets (see Table 1). Such small c however leads to ranking that takes into account only local information about the Web graph (see e.g. [20]). The choice $\mathbf{v} = \hat{\pi}_{\text{ESCC}}$ does not seem to represent the dynamics of the system; probably because the “easily bored surfer” random walk that is used in PageRank computations never follows a quasi-stationary distribution since it often restarts itself from the uniform probability vector.

For the uniform vector $\mathbf{v} = \mathbf{u}_{\text{ESCC}}$, we have $||\mathbf{v}T|| = p_1$, which gives c_1, c_2, c^* presented in Table 1. We have obtained a higher upper bound but the values of c^* are still much smaller than 0.85.

Finally, consider the normalized PageRank vector $\mathbf{v}(c) = \pi_{\text{ESCC}}(c)/||\pi_{\text{ESCC}}(c)||$. This choice of $\mathbf{v}$ can also be justified as follows. Consider the derivative of the total PageRank mass of ESCC. Since $[I-cT]^{-1}$ and $[I-T]$ commute, we can write

$$\frac{d}{dc}||\pi_{\text{ESCC}}(c)|| = -\gamma \mathbf{u}_{\text{ESCC}}[I-cT]^{-1}[I-T][I-cT]^{-1}\mathbf{1},$$

or, equivalently,

$$\begin{aligned}\frac{d}{dc}||\pi_{\text{ESCC}}(c)|| &= -\frac{1}{1-c}\pi_{\text{ESCC}}[I-T][I-cT]^{-1}\mathbf{1}\\ &= -\frac{1}{1-c}\left(\pi_{\text{ESCC}} - ||\pi_{\text{ESCC}}||\frac{\pi_{\text{ESCC}}}{||\pi_{\text{ESCC}}||}T\right)[I-cT]^{-1}\mathbf{1}\\ &= -\frac{1}{1-c}\left(\pi_{\text{ESCC}} - ||\pi_{\text{ESCC}}||\mathbf{v}(c)T\right)[I-cT]^{-1}\mathbf{1},\end{aligned}$$

with $\mathbf{v}(c) = \pi_{\text{ESCC}}/||\pi_{\text{ESCC}}||$. It is easy to see that

$$||\pi_{\text{ESCC}}(c)|| = \gamma - \gamma(1-\mathbf{u}_{\text{ESCC}}T\mathbf{1})c + \mathrm{o}(c).$$

Consequently, we obtain

$$\frac{d}{dc}||\pi_{\text{ESCC}}(c)|| = -\frac{1}{1-c}\left(\pi_{\text{ESCC}} - \gamma\mathbf{v}(c)T + \gamma(1-\mathbf{u}_{\text{ESCC}}T\mathbf{1})c\mathbf{v}(c)T + \mathrm{o}(c)\right)[I-cT]^{-1}\mathbf{1}.$$

Since in practice T is very close to stochastic, we have

$$1-\mathbf{u}_{\text{ESCC}}T\mathbf{1} \approx 0, \quad \text{and} \quad [I-cT]^{-1}\mathbf{1} \approx \frac{1}{1-c}\mathbf{1}.$$

The latter approximation follows from Lemma 2. Thus, satisfying condition (22) means keeping the value of the derivative small.

Let us now solve (22) for $\mathbf{v}(c) = \pi_{\text{ESCC}}(c)/||\pi_{\text{ESCC}}(c)||$. Using (18), we rewrite (22) as

$$||\pi_{\text{ESCC}}(c)|| = \frac{\gamma}{||\pi_{\text{ESCC}}(c)||}\pi_{\text{ESCC}}(c)T\mathbf{1} = \frac{\gamma^2(1-c)}{||\pi_{\text{ESCC}}(c)||}\mathbf{u}_{\text{IN+SCC}}[I-cT]^{-1}T\mathbf{1},$$

Multiplying by $||\pi_{\text{ESCC}}(c)||$, after some algebra we obtain

$$||\pi_{\text{ESCC}}(c)||^2 = \frac{\gamma}{c}\,||\pi_{\text{ESCC}}(c)|| - \frac{(1-c)\gamma^2}{c}.$$

Solving the quadratic equation for $||\pi_{\text{ESCC}}(c)||$, we get

$$||\pi_{\text{ESCC}}(c)|| = r(c) = \begin{cases}\gamma & \text{if } c \le 1/2,\\ \frac{\gamma(1-c)}{c} & \text{if } c > 1/2.\end{cases}$$

Hence, the value c^* solving (22) corresponds to the point where the graphs of $||\pi_{\text{ESCC}}(c)||$ and $r(c)$ cross each other. There is only one such point on (0,1), and since $||\pi_{\text{ESCC}}(c)||$ decreases very slowly unless c is close to one, whereas $r(c)$ decreases relatively fast for $c > 1/2$, we expect that c^* is only slightly larger than $1/2$. Under conditions of Proposition 3, $r(c)$ first crosses the line $\gamma(1-c)/(1-\lambda_1 c)$, then $||\pi_{\text{ESCC}}(c)||_1$, and then $\gamma(1-c)/(1-p_1 c)$. Thus, we yield $(1+\lambda_1)^{-1} < c^* < (1+p_1)^{-1}$. Since both λ_1 and p_1 are large, this suggests that c should be chosen around $1/2$. This is also reflected in Table 1.

Last but not least, to support our theoretical argument about the undeserved high ranking of pages from Pure OUT, we carry out the following experiment. In the $INRIA$ dataset we have chosen an absorbing component in Pure OUT consisting just of two nodes. We have added an artificial link from one of these nodes to a node in the Giant SCC and recomputed the PageRank. In Table 2 in the column "PR rank w/o link" we give a ranking of a page according to the PageRank value computed before the addition of the artificial link and in the column "PR rank with link" we give a ranking of a page according to the PageRank value computed after the addition of the artificial link. We have also analyzed the log file of the site INRIA Sophia Antipolis (`www-sop.inria.fr`) and ranked the pages according to the number of clicks for the period of one year up to May 2007. We note that since we have the access only to the log file of the INRIA Sophia Antipolis site, we use the PageRank ranking also only for the pages from the INRIA Sophia Antipolis site. For instance, for $c = 0.85$, the ranking of Page A without an artificial link is 731 (this means that 730 pages are ranked better than Page A among the pages of INRIA Sophia Antipolis). However, its ranking according to the number of clicks is much lower, 2588. This confirms our conjecture that the nodes in Pure OUT obtain unjustifiably high ranking. Next we note that the addition of an artificial link significantly diminishes the ranking. In fact, it brings it close to the ranking provided by the number of clicks. Finally, we draw the attention of the reader to the fact that choosing $c = 1/2$ also significantly reduces the gap between the ranking by PageRank and the ranking by the number of clicks.

c	PR rank w/o link	PR rank with link	rank by no. of clicks
Node A			
0.5	1648	2307	2588
0.85	731	2101	2588
0.95	226	2116	2588
Node B			
0.5	1648	4009	3649
0.85	731	3279	3649
0.95	226	3563	3649

Table 2. Comparison between PR and click based rankings.

To summarize, our results indicate that with $c = 0.85$, the Pure OUT component receives an unfairly large share of the PageRank mass. Remarkably, in order to satisfy any of the three intuitive criteria of fairness presented above, the value of c should be drastically reduced. The experiment with the log files confirms the same. Of course, a drastic reduction of c also considerably accelerates the computation of PageRank by numerical methods [21,5,22].

Acknowledgments

This work is supported by EGIDE ECO-NET grant no. 10191XC and by NWO Meervoud grant no. 632.002.401.

References

1. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University (1998)
2. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5) (1999) 604–632
3. Lempel, R., Moran, S.: The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Comput. Networks **33**(1–6) (2000) 387–401
4. Langville, A.N., Meyer, C.D.: Deeper inside PageRank. Internet Math. **1** (2003) 335–380
5. Langville, A.N., Meyer, C.D.: Google's PageRank and beyond: the science of search engine rankings. Princeton University Press, Princeton, NJ (2006)
6. Boldi, P., Santini, M., Vigna, S.: PageRank as a function of the damping factor. In: Proc. of the Fourteenth International World Wide Web Conference, Chiba, Japan, ACM Press (2005)
7. Avrachenkov, K., Litvak, N.: The effect of new links on Google PageRank. Stoch. Models **22**(2) (2006) 319–331
8. Bianchini, M., Gori, M., Scarselli, F.: Inside PageRank. ACM Trans. Inter. Tech. **5**(1) (2005) 92–128
9. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Statac, R., Tomkins, A., J.Wiener: Graph structure in the Web. Computer Networks **33** (2000) 309–320
10. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E.: The Web as a graph. In: Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS, ACM Press (15–17 2000) 1–10
11. Avrachenkov, K.: Analytic Perturbation Theory and its Applications. PhD thesis, University of South Australia (1999)
12. Korolyuk, V.S., Turbin, A.F.: Mathematical foundations of the state lumping of large systems. Volume 264 of Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht (1993)
13. Pervozvanskii, A.A., Gaitsgori, V.G.: Theory of suboptimal decisions. Volume 12 of Mathematics and its Applications (Soviet Series). Kluwer Academic Publishers Group, Dordrecht (1988)
14. Yin, G.G., Zhang, Q.: Discrete-time Markov chains. Volume 55 of Applications of Mathematics (New York). Springer-Verlag, New York (2005)
15. Chen, P., Xie, H., Maslov, S., Redner, S.: Finding scientific gems with google's pagerank algorithm. Arxiv preprint Physics 0604130 (2006)
16. Dill, S., Kumar, R., Mccurley, K.S., Rajagopalan, S., Sivakumar, D., Tomkins, A.: Self-similarity in the web. ACM Trans. Inter. Tech. **2**(3) (2002) 205–223
17. Moler, C., Moler, K.: Numerical Computing with MATLAB. SIAM (2003)
18. Eiron, N., McCurley, K., Tomlin, J.: Ranking the Web frontier. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM Press (2004) 309–318

19. Seneta, E.: Non-negative matrices and Markov chains. Springer Series in Statistics. Springer, New York (2006) Revised reprint of the second (1981) edition [Springer-Verlag, New York; MR0719544].
20. Fortunato, S., Flammini, A.: Random walks on directed networks: the case of PageRank. Technical Report 0604203, arXiv/physics (2006)
21. Avrachenkov, K., Litvak, N., Nemirovsky, D., Osipova, N.: Monte Carlo methods in PageRank computation: When one iteration is sufficient. SIAM J. Numer. Anal. **45**(2) (2007) 890–904
22. Berkhin, P.: A survey on PageRank computing. Internet Math. **2** (2005) 73–120
23. Avrachenkov, K., Haviv, M., Howlett, P.: Inversion of analytic matrix functions that are singular at the origin. SIAM Journal on Matrix Analysis and Applications **22**(4) (2001) 1175–1189

Appendix: Results from Singular Perturbation Theory

Lemma 1. *Let $A(\varepsilon) = A + \varepsilon C$ be a transition matrix of perturbed Markov chain.*

The perturbed Markov chain is assumed to be ergodic for sufficiently small ε different from zero. Let the unperturbed Markov chain ($\varepsilon = 0$) have m ergodic classes. Namely, the transition matrix A can be written in the form

$$A = \begin{bmatrix} A_1 & & 0 & 0 \\ & \ddots & & \\ 0 & & A_m & 0 \\ L_1 & \cdots & L_m & E \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Then, the stationary distribution of the perturbed Markov chain has a limit

$$\lim_{\varepsilon \to 0} \pi(\varepsilon) = [\nu_1 \mu_1 \ \cdots \ \nu_m \mu_m \ 0],$$

where zeros correspond to the set of transient states in the unperturbed Markov chain, μ_i is a stationary distribution of the unperturbed Markov chain corresponding to the i-th ergodic set, and ν_i is the i-th element of the aggregated stationary distribution vector that can be found by solution

$$\nu D = \nu, \quad \nu \mathbf{1} = 1,$$

where $D = MCB$ is the generator of the aggregated Markov chain and

$$M = \begin{bmatrix} \mu_1 & & 0 & 0 \\ & \ddots & & \\ 0 & & \mu_m & 0 \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad B = \begin{bmatrix} \mathbf{1} & & 0 \\ & \ddots & \\ 0 & & \mathbf{1} \\ \phi_1 & \cdots & \phi_m \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

with $\phi_i = [I - E]^{-1} L_i \mathbf{1}$.

The proof of this lemma can be found in [11,12,14].

Lemma 2. *Let $A(\varepsilon) = A - \varepsilon C$ be a perturbation of irreducible stochastic matrix A such that $A(\varepsilon)$ is substochastic. Then, for sufficiently small ε the following Laurent series expansion holds*

$$[I - A(\varepsilon)]^{-1} = \frac{1}{\varepsilon} X_{-1} + X_0 + \varepsilon X_1 + \ldots \; , \tag{23}$$

with

$$X_{-1} = \frac{1}{\mu C \mathbf{1}} \mathbf{1} \mu, \tag{24}$$

where μ is the stationary distribution of A. It follows that

$$[I - A(\varepsilon)]^{-1} = \frac{1}{\mu \varepsilon C \mathbf{1}} \mathbf{1} \mu + O(1) \quad \textit{as } \varepsilon \to 0. \tag{25}$$

Proof. The proof of this result is based on the approach developed in [11,23]. The existence of the Laurent series (23) is a particular case of more general results on the inversion of analytic matrix functions [23]. To calculate the terms of the Laurent series, let us equate the terms with the same powers of ε in the following identity

$$(I - A + \varepsilon C)(\frac{1}{\varepsilon} X_{-1} + X_0 + \varepsilon X_1 + \ldots) = I,$$

which results in

$$(I - A) X_{-1} = 0, \tag{26}$$

$$(I - A) X_0 + C X_{-1} = I, \tag{27}$$

$$(I - A) X_1 + C X_0 = 0. \tag{28}$$

From equation (26) we conclude that

$$X_{-1} = \mathbf{1} \mu_{-1}, \tag{29}$$

where μ_{-1} is some vector. We find this vector from the condition that the equation (27) has a solution. In particular, equation (27) has a solution if and only if

$$\mu (I - C X_{-1}) = 0.$$

By substituting into the above equation the expression (29), we obtain

$$\mu - \mu C \mathbf{1} \mu_{-1} = 0,$$

and, consequently,

$$\mu_{-1} = \frac{1}{\mu C \mathbf{1}} \mu,$$

which together with (29) gives (24).