Stochastic Models, 22:319–331, 2006
Copyright © Taylor & Francis Group, LLC
ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1080/15326340600649052

THE EFFECT OF NEW LINKS ON GOOGLE PAGERANK

Konstantin Avrachenkov □ *INRIA Sophia Antipolis, France*

Nelly Litvak □ *Department of Applied Mathematics, University of Twente, Enschede, The Netherlands*

□ *PageRank is one of the principle criteria according to which Google ranks Web pages. PageRank can be interpreted as the frequency that a random surfer visits a Web page, and thus it reflects the popularity of a Web page. We study the effect of newly created links on Google PageRank. We discuss to what extent a page can control its PageRank. Using asymptotic analysis we provide simple conditions that show whether or not new links result in increased PageRank for a Web page and its neighbors. Furthermore, we show that there exists an optimal (although impractical) linking strategy. We conclude that a Web page benefits from links inside its Web community and on the other hand irrelevant links penalize the Web pages and their Web communities.*

Keywords Asymptotic analysis; Optimal linking; PageRank; Rank one update.

Mathematics Subject Classification Primary 60J10, 60J20, 90B12; Secondary 90B18, 91D30, 68M10.

1. INTRODUCTION

Surfers on the Internet frequently use search engines to find pages satisfying their query. However, there are typically hundreds or thousands of relevant pages available on the Web. Thus, listing them in an adequate order is a crucial and non-trivial task. The original idea that Google founders presented in Ref.[4] was to list pages according to their PageRank, which is a measure of page popularity. The PageRank is defined in the following way. Denote by n the total number of pages on the Web and define the $n \times n$ hyperlink matrix P as follows. Suppose that page i has $k > 0$ outgoing links. Then $p_{ij} = 1/k$ if j is one of the outgoing links and

Received July 2004; Accepted August 2005

Address correspondence to Nelly Litvak, Faculty of Electrical Engineering, Mathematics, and Computer Science, Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; E-mail: n.litvak@math.utwente.nl

$p_{ij} = 0$ otherwise. If a page does not have outgoing links, the probability is spread among all pages of the Web, namely, $p_{ij} = 1/n$ for all $j = 1, \ldots, n$. Further, it is assumed that a random surfer goes with some probability to an arbitrary Web page with the uniform distribution. Thus, 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

$$\widehat{P} = cP + (1 - c)(1/n)E, \tag{1}$$

where E is a matrix whose all entries are equal to one, n is the number of Web pages, and $c \in (0, 1)$ is the probability of not jumping to a random page (Google originally used $c = 0.85$). The Google matrix $\widehat{P}$ is stochastic, aperiodic, and irreducible, so there exists a unique row vector π such that

$$\pi\widehat{P} = \pi, \quad \pi\underline{1} = 1, \tag{2}$$

where $\underline{1}$ is a column vector of ones. The row vector π satisfying (2) is called a PageRank vector, or simply PageRank. If a surfer follows a hyperlink with probability c and jumps to a random page with probability $1 - c$, then π_i can be interpreted as a stationary probability that the surfer is at page i.

The factor c serves several purposes. The most apparent purposes, which are widely discussed in the literature (Ref.[12]) are as follows: (i) if $c < 1$ then the matrix $\widehat{P}$ is irreducible, and thus the PageRank distribution exists and is uniquely defined; (ii) choosing the value of c not too close to 1, one can guarantee a fast convergence of the power iteration method in PageRank computations (Refs.[7,12]). As we show in Section 2, the parameter c is also responsible for robustness of PageRank π_i with respect to outgoing links of page i. We further discuss the impact of the factor c in Conclusions (Section 6).

In order to keep up with constant modifications of the Web structure, Google regularly updates its PageRank. According to publicly available information Google uses power iterations to compute the PageRank. Several proposals such as Refs.[1,3,8–10,13–16] (see also an extensive survey paper Ref.[12]) have recently been put forward to accelerate the PageRank computation. This research direction can be regarded as taking a system point of view. On the contrary, here we take a user point of view and try to answer the following questions: When do new links benefit the Web page from which they emanate? When do the pages from the same Web community benefit from the link creation? Is there an optimal linking strategy? As one can see from numerous articles and forums (see e.g., Refs.[20,21]), these questions are highly relevant in the search engine optimization community. Our paper is an attempt to suggest rigorous mathematical arguments that may help to provide solid practical recommendations as well as to dismiss many common misconceptions.

2. TO WHAT EXTENT CAN A PAGE CONTROL ITS PAGERANK?

The PageRank defined in (2) clearly depends on both incoming and outgoing links of a page. Thus, the easiest way for a page to change its ranking is to modify the outgoing links. Below we shall show that the PageRank can be written as a product of three terms where only one term depends on outgoing links. It will allow us to estimate the extent to which a page can control its PageRank. To this end, we first recall the following useful expression for the PageRank (Refs.[2,12,16]), which in fact follows directly from (1), (2):

$$\pi = \frac{1-c}{n}\underline{1}^T[I-cP]^{-1}. \tag{3}$$

Let z_{ij} denote the (i,j) element of the matrix $Z=[I-cP]^{-1}$. Namely, we have

$$z_{ij} = e_i^T[I-cP]^{-1}e_j, \quad i,j=1,\dots,n, \tag{4}$$

where e_k is the kth column of the identity matrix I. Now, define a discrete-time absorbing Markov chain $\{X_t, t=0,1,\dots\}$ with the state space $\{0,1\dots,n\}$, where transitions between the states $1,\dots,n$ are conducted by the matrix cP, and the state 0 is absorbing. Let N_j be the number of visits to state $j=1,\dots,n$ before absorption including the visit at time $t=0$ if X_0 is j. Formally,

$$N_j = \sum_{t=0}^{\infty}\mathbf{1}_{\{X_t=j\}}, \quad j=1,\dots,n,$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. Note that z_{ij} equals the conditional expectation of N_j given that the initial state is i:

$$\begin{aligned} z_{ij} &= e_i^T\left[\sum_{t=0}^{\infty}c^tP^t\right]e_j = \sum_{t=0}^{\infty}e_i^Tc^tP^te_j = \sum_{t=0}^{\infty}\mathbb{P}(X_t=j \mid X_0=i) \\ &= \sum_{t=0}^{\infty}\mathbb{E}(\mathbf{1}_{\{X_t=j\}}|X_0=i) = \mathbb{E}(N_j \mid X_0=i). \end{aligned}$$

Let q_{ij} be the probability of reaching the state j before absorption if the initial state is i. Using the strong Markov property (Ref.[17]), we can establish the following decomposition result.

Proposition 2.1. *The PageRank of page i is given by*

$$\pi_i = \frac{1-c}{n}z_{ii}\left(1+\sum_{\substack{j=1\\ j\neq i}}^{n}q_{ji}\right), \quad i=1,\dots,n. \tag{5}$$

Proof. It follows from (3) that

$$\pi_i = \frac{1-c}{n}\underline{1}^T[I - cP]^{-1}e_i = \frac{1-c}{n}\sum_{j=1}^{n} z_{ji}. \tag{6}$$

Further, we have $\mathbb{P}(N_i = 0 \mid X_0 = j) = 1 - q_{ji}$, and using the strong Markov property, we can write for any $k \geq 1$

$$\begin{aligned}\mathbb{P}(N_i = k \mid X_0 = j) &= \mathbb{P}(N_i = k, N_i \geq 1 \mid X_0 = j)\\ &= \mathbb{P}(N_i \geq 1 | X_0 = j)\mathbb{P}(N_i = k \mid X_0 = j, N_i \geq 1)\\ &= q_{ji}\mathbb{P}(N_i = k \mid X_0 = i).\end{aligned}$$

Consequently, for any $i, j = 1, \ldots, n$; $i \neq j$, we have

$$z_{ji} = \mathbb{E}(N_i | X_0 = j) = q_{ji}\mathbb{E}(N_i \mid X_0 = i) = q_{ji}z_{ii}. \tag{7}$$

Substituting (7) in (6) we immediately obtain (5). □

The decomposition formula (5) represents the PageRank of page i as a product of three multipliers where only the term z_{ii} depends on the outgoing links of page i. Hence, by changing the outgoing links, a page can control its PageRank up to multiplication by a factor $z_{ii} = 1/(1 - q_{ii}) \in [1, (1 - c^2)^{-1}]$, where $q_{ii} \in [0, c^2]$ is the probability of return back to i starting from i (the upper bound $(1 - c^2)^{-1}$ is approximately 3.6 for $c = .85$). We note that even a threefold increase of the PageRank might not be considered as a significant improvement, since Google measures the PageRank on a logarithmic scale (Ref.[19]). An increase of this order could be helpful for unimportant pages but in practice the upper bound $(1 - c^2)^{-1}$ is hardly possible to achieve. Indeed, in order to ensure the highest possible return probability c^2, a page i must point to pages that link only to i. Such a policy makes i and its neighbors isolated from the rest of the Web. In fact, our results suggest once a page i provides a number of "natural" links, such as a link from a user's homepage to his/her department homepage, the multiplication factor z_{ii} becomes quite robust and hardly subject to major changes. The conclusion is that the PageRank of a Web page cannot be improved considerably by manipulating its outgoing links. The greatest possible increase is not very significant, and it can be achieved only by damaging a logical link structure, which will not pay off in the end.

If page i does not have outgoing links, then z_{ii} is very close to the lower bound 1, since there is almost no chance to return from i back to i before absorption. Ref.[2] presents a circuit analysis to study in detail the influence of pages without outgoing links (leaves, dangling nodes) on the PageRank of a Web community. The authors of Ref.[2] came to the same conclusion that dangling causes a considerable loss in ranking. Our formula (5) helps to quantify this loss.

In the ensuing sections we shall obtain exact formulae that further quantify the changes in the PageRank distribution when new links are added by one of the pages.

3. RANK ONE UPDATE OF GOOGLE PAGERANK

Let us consider a Web page with k_0 old hyperlinks and k_1 newly created hyperlinks. Without loss of generality, we assume that the page with new links has index 1 and the pages towards which new links are pointed have indices from 2 to k_1+1. Put $k=k_0+k_1$ and let p_1^T be the first row of matrix P. Then after adding the new links the first row becomes $(k_0/k)p_1^T + (1/k)\sum_{i=2}^{k_1+1} e_i^T$, and thus the addition of new links can be regarded as rank one update of the hyperlink matrix

$$\widetilde{P} = P + e_1 u^T, \tag{8}$$

with

$$u^T = \frac{1}{k}\sum_{i=2}^{k_1+1} e_i^T - \frac{k_1}{k} p_1^T.$$

In Ref.[13] the authors use updating formulae to accelerate the PageRank computation. By restricting ourselves to the case of a rank one update, we are able to perform a more comprehensive analysis. The next theorem provides updating formulae for the PageRank elements.

Theorem 3.1. *Let k_1 new links emanating from page 1 be added. Then, the elements of new PageRank vector are given by the following updating formulae*

$$\tilde{\pi}_1 = \frac{\pi_1}{1 - \frac{k_1}{k}\left(1 + \frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{i1} - z_{11}\right)}, \tag{9}$$

$$\tilde{\pi}_j = \pi_j + \pi_1 \frac{\frac{k_1}{k}\left(\frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{ij} - z_{1j}\right)}{1 - \frac{k_1}{k}\left(1 + \frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{i1} - z_{11}\right)}, \quad j = 2, \ldots, n. \tag{10}$$

Proof. Applying the Sherman-Morrison-Woodbury updating formula (Ref.[6]) to $[I - c\widetilde{P}]^{-1}$, we can write

$$[I - c\widetilde{P}]^{-1} = [I - cP]^{-1} + c\frac{[I - cP]^{-1} e_1 u^T [I - cP]^{-1}}{1 - cu^T [I - cP]^{-1} e_1}.$$

Then, premultiplying the above equation by $\frac{1-c}{n}\underline{1}^T$ and using (3), we get

$$\tilde{\pi} = \pi + \pi_1 \frac{cu^T[I - cP]^{-1}}{1 - cu^T[I - cP]^{-1}e_1},$$

and consequently,

$$\tilde{\pi}_1 = \pi_1 \frac{1}{1 - cu^T[I - cP]^{-1}e_1}, \tag{11}$$

$$\tilde{\pi}_j = \pi_j + \pi_1 \frac{cu^T[I - cP]^{-1}e_j}{1 - cu^T[I - cP]^{-1}e_1}, \quad j = 2, \dots, n. \tag{12}$$

Next, we evaluate $cu^T[I - cP]^{-1}e_j$ for $j = 1, \dots, n$,

$$cu^T[I - cP]^{-1}e_j = cu^T Ze_j = c\frac{k_1}{k}\left(\frac{1}{k_1}\sum_{i=2}^{k_1+1} e_i^T - p_1^T\right)Ze_j$$

$$= \frac{k_1}{k}\left(\frac{c}{k_1}\sum_{i=2}^{k_1+1} e_i^T Ze_j - cp_1^T Ze_j\right).$$

Since $cPZ = Z - I$, we have $cp_1^T Z = z_1^T - e_1^T$, where z_1^T is the 1st row of the matrix Z, and hence $cp_1^T Ze_j = z_{1j} - e_1^T e_j$. Thus, we get

$$cu^T[I - cP]^{-1}e_j = \frac{k_1}{k}\left(\frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{ij} - (z_{1j} - e_1^T e_j)\right).$$

Substituting the above expression for $cu^T[I - cP]^{-1}e_j$, $j = 1, \dots, n$, into (11) and (12), we obtain (9) and (10). □

The results in Theorem 3.1 are in line with formula (5). If page 1 updates its outgoing links then in decomposition (5) for π_1 only the second multiplier will be affected. In the new situation, the probability $\tilde{q}_{11}$ to return to page 1 starting from this page, is given by

$$\tilde{q}_{11} = \frac{k - k_1}{k} q_{11} + \frac{c}{k}\sum_{i=2}^{k_1+1} q_{i1}.$$

Substituting this expression in

$$\tilde{\pi}_1 = \frac{\tilde{z}_{11}}{z_{11}}\pi_1 = \frac{1 - q_{11}}{1 - \tilde{q}_{11}}\pi_1,$$

we get the updating formula (9). According to (9) the ranking of page 1 increases when

$$1 + \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{i1} - z_{11} > 0, \tag{13}$$

which is equivalent to

$$\frac{1}{k_1} \sum_{i=2}^{k_1+1} q_{i1} > q_{11}.$$

Hence, the page 1 increases its ranking when it refers to pages that are characterized by a high value of q_{i1}. These must be the pages that refer to page 1 or at least belong to the same Web community. Here by a Web community we mean a set of Web pages that a surfer can reach from one to another in a relatively small number of steps.

Let us now consider formula (10). First, we see that the difference between the old and the new ranking of page j is proportional to π_1. Naturally, hyperlink references from pages with high ranking have a greater impact on other pages. Furthermore, the PageRank of page j increases if

$$\frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{ij} > z_{1j}. \tag{14}$$

Indeed, if (14) holds then the increase of PageRank for page j follows from (6) since z_{kj} increases for each page k that has a path to j via page 1, and the other z_{kj}'s remain unaffected. Naturally, it is most beneficial for page j to receive one of the new links. Formally, it follows from (7) that $z_{ij} = q_{ij} z_{jj}$ where $q_{ij} < 1$, so that z_{jj} constitutes the maximal possible contribution in the left-hand side of (14). On the other hand, if several new links are added then the PageRank of page j might actually decrease even if this page receives one of the new links. Such situation occurs when most of newly created links point to "irrelevant" pages. For instance, let $j = 2$ and assume that there is no hyperlink path from pages $3, \dots, k+1$ to page 2. Then z_{ij} is close to zero for $i = 3, \dots, k+1$, and the PageRank of page 2 will increase only if $(c/k_1) z_{22} > z_{12}$, which is not necessarily true, especially if z_{12} and k_1 are considerably large. The asymptotic analysis in the next section allows us to further clarify this issue.

4. ASYMPTOTIC ANALYSIS

Let us apply an asymptotic analysis to formulae (9) and (10) when c is close to one and the Markov chain induced by the hyperlink matrix P

is irreducible. The asymptotic approach allows us to derive simple natural conditions that show if a Web page with newly created links and its neighbors benefit from these new links.

Let m_{ij} be the average time needed to reach j starting from i when the random walk follows the original hyperlink matrix P, i.e., $c = 1$. We refer to the m_{ij}'s as mean first passage times (Ref.[11]). Note that $m_{ii} > 1$ is the mean return time to page i starting from this page. The following theorem allows us to predict changes in PageRank using the mean first passage times.

Theorem 4.1. *Let c be sufficiently close to one and let page 1 have k_1 new links to pages $\{2, \ldots, k_1 + 1\}$. Assume that the Markov chain induced by the hyperlink matrix P is irreducible. Then, we have the following conditions:*

1. *if $m_{11} > 1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1} m_{i1}$, the creation of new links $\{1 \to 2, \ldots, 1 \to k_1 + 1\}$ increases the PageRank of page 1;*
2. *if $m_{1j} > 1 + \frac{1}{k_1}\sum_{i=2, i\neq j}^{k_1+1} m_{ij}$, the creation of new links $\{1 \to 2, \ldots, 1 \to k_1 + 1\}$ increases the PageRank of page j from the set $\{2, \ldots, k_1 + 1\}$;*
3. *if $m_{1l} > 1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1} m_{il}$, the creation of new links $\{1 \to 2, \ldots, 1 \to k_1 + 1\}$ increases the PageRank of page l, for $l > k_1 + 1$;*

Proof. First, we make a change of variable $c = 1/(1+\rho)$, with $\rho > 0$ in the formula for $Z(c)$ as follows:

$$Z(c) = [I - cP]^{-1} = \left[I - \frac{1}{1+\rho}P\right]^{-1} = (1+\rho)[\rho I - (P - I)]^{-1}.$$

Then, we use the resolvent Laurent series expansion for the Markov chain generator (see e.g., Ref.[18], Theorem 8.2.3)

$$\begin{aligned} Z(c) &= [I - cP]^{-1} = (1+\rho)[\rho I - (P - I)]^{-1} \\ &= (1+\rho)\left[\frac{1}{\rho}\Pi + H + \sum_{k=1}^{\infty} \rho^k (-1)^k H^{k+1}\right], \end{aligned}$$

where $\Pi = \underline{1}\pi$ is the ergodic projection and H is the deviation matrix of the Markov chain induced by P. Since we consider a finite state Markov chain, the above series has a non-zero radius of convergence.

Let us now prove the first statement of the theorem. It is enough to show that Condition 1 guarantees that

$$1 + \frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{i1}(c) - z_{11}(c) > 0$$

for all c sufficiently close to one, or equivalently, for all ρ sufficiently close to zero,

$$\begin{aligned}
&1 + \frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{i1}(c) - z_{11}(c) \\
&\quad = 1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1}\left(\frac{1}{\rho}\pi_1 + h_{i1} + O(\rho)\right) - (1+\rho)\left(\frac{1}{\rho}\pi_1 + h_{11} + O(\rho)\right) \\
&\quad = 1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1} h_{i1} - h_{11} - \pi_1 + O(\rho).
\end{aligned}$$

Since $m_{ij} = (h_{jj} - h_{ij})/\pi_j$ for $i \neq j$ (Ref.[11]), we have

$$\begin{aligned}
1 + \frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{i1}(c) - z_{11}(c) &= 1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1}(h_{11} - m_{i1}\pi_1) - h_{11} - \pi_1 + O(\rho) \\
&= 1 - \pi_1\left(1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1} m_{i1}\right).
\end{aligned}$$

Then, the condition

$$1 - \pi_1\left(1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1} m_{i1}\right) > 0$$

is equivalent to

$$m_{11} > 1 + \frac{1}{k_1}\sum_{i=2}^{k_1+1} m_{i1},$$

since $m_{11} = 1/\pi_1$.

Next we prove the second statement of the theorem. It is enough to show that Condition 2 implies that for a given $j \in \{2, \ldots, k_1 + 1\}$,

$$\frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{ij}(c) - z_{1j}(c) > 0$$

for all c sufficiently close to one, or equivalently, for all ρ sufficiently close to zero. We write

$$\begin{aligned}
&\frac{c}{k_1}\sum_{i=2}^{k_1+1} z_{ij}(c) - z_{1j}(c) \\
&\quad = \frac{1}{k_1}\sum_{i=2}^{k_1+1}\left(\frac{1}{\rho}\pi_j + h_{ij} + O(\rho)\right) - (1+\rho)\left(\frac{1}{\rho}\pi_j + h_{1j} + O(\rho)\right)
\end{aligned}$$

$$= \frac{1}{k_1}\sum_{i=2}^{k_1+1} h_{ij} - h_{1j} - \pi_j + O(\rho)$$

$$= \frac{1}{k_1}\sum_{i=2,i\neq j}^{k_1+1} (h_{ij} - h_{1j} - \pi_j) + \frac{1}{k_1}(h_{jj} - h_{1j} - \pi_j) + O(\rho)$$

$$= \frac{1}{k_1}\sum_{i=2,i\neq j}^{k_1+1} (h_{jj} - m_{ij}\pi_j - h_{jj} + m_{1j}\pi_j - \pi_j) + \frac{1}{k_1}(m_{1j}\pi_j - \pi_j) + O(\rho)$$

$$= \frac{\pi_j}{k_1}\left(\sum_{i=2,i\neq j}^{k_1+1} (m_{1j} - m_{ij} - 1) + (m_{1j} - 1)\right) + O(\rho).$$

The latter expression is positive for all sufficiently small ρ, if

$$\sum_{i=2,i\neq j}^{k_1+1} (m_{1j} - m_{ij} - 1) + (m_{1j} - 1) > 0,$$

which is equivalent to Condition 2. The proof of Condition 3 is very similar to the proofs of Conditions 1 and 2. □

All conditions of Theorem 4.1 have very clear probabilistic interpretations. These interpretations become even more transparent in the case when only one new link is added. For instance, in the case of a single new link Condition 1 takes the form $m_{11} > m_{21} + 1$. The latter means that the average path from page 2 to page 1 should be shorter at least by one hop than the average return path from page 1 to itself. Condition 2 becomes $m_{12} > 1$, which is always true. This is not surprising as we know that an addition of a *single* new link pointing to a Web page always increases its PageRank (see e.g., Ref.[2]). However, as mentioned at the end of the previous section, if *several* new links are added simultaneously, then the pages receiving a new link do not necessarily have increased PageRank. Writing the inequality in Condition 2 for some $j = 2, \ldots, k+1$, we see that the right-hand side may become quite large when several new links point to pages that, on average, have very long paths to page j. To summarize, incoming links are especially valuable if they are received from popular pages that are dedicated to a particular Web community.

There is a striking similarity between Conditions 1–3 from Theorem 4.1 and conditions (13) and (14) derived in the previous section. However, in the asymptotic case, the conditions of improving the PageRank are expressed in terms of mean first passage times rather than mean number of visits before absorption.

5. OPTIMAL LINKING STRATEGY

In this section we show that there exists in fact an optimal linking strategy. Consider a page $i = 1, \ldots, n$ and assume that i has links to pages $i_1, \ldots, i_k$ distinct from i. Further, let $m_{ij}(c)$ be the mean first passage time from page i to page j for the Google transition matrix $\widehat{P}$ with parameter c. Then writing the linear equations for the mean first passage times (Ref. [11]), we obtain

$$m_{ii}(c) = 1 + \frac{c}{k} \sum_{l=1}^{k} m_{i_l i}(c) + \frac{1}{n}(1-c) \sum_{\substack{j=1 \\ j \neq i}}^{n} m_{ji}(c), \tag{15}$$

The objective now is to choose k and $i_1, \ldots, i_k$ such that $m_{ii}(c)$ becomes as small as possible and consequently $\pi_i = 1/m_{ii}(c)$ becomes as large as possible. From (15) one can see that $m_{ii}(c)$ is a linear function of the values $m_{ji}(c)$, $j \neq i$. Moreover, outgoing links from i do not affect $m_{ji}(c)$ for any $j \neq i$. Thus, by linking from i to j, one can only alter the coefficients in the right-hand side of (15). As was also noted in Section 2, this means that the owner of the page i has very little control over its PageRank. The best that he/she can do is to link only to one page j^* such that

$$m_{j^* i}(c) = \min_j \{m_{ji}(c)\}.$$

Note that (surprisingly) the PageRank of j^* plays no role here. Thus, we have the following statement.

Theorem 5.1. *The optimal linking strategy for a Web page is to have only one outgoing link pointing to a Web page with a shortest mean first passage time back to the original page.*

This matches the observation in the end of Section 2 that the improvement in the PageRank is maximal when page i links to pages that have hyperlinks to i only. Definitely, such pages have the smallest value of the mean first passage time to i. Of course, linking to only one page as suggested by Theorem 5.1 most likely will result in poor content quality of the Web page. However, the message from the above theorem is that one has to link to pages that are relevant and belong to the same Web community. Interestingly, the discussion on optimal linking strategy partially explains the "practical" advice according to which, a Web site owner should view his/her site as a set of pages and maintain a good interlink structure and to refer to his/her colleagues (Ref.[20]). Indeed, it follows from our arguments that such a policy will certainly increase the PageRank of all pages in the group.

6. CONCLUSIONS

Our main conclusion is that a Web page cannot significantly manipulate its PageRank by changing its outgoing links. Furthermore, keeping a logical hyperlink structure and linking to a relevant Web community is the most sensible and rewarding policy. This statement has often been uttered by many leading search engine optimization specialists, and it has now received a rigorous mathematical discussion in the present paper. The multiplication factor that is still subject to control is bounded by $(1-c^2)^{-1}$, which gives a new interpretation for the "Google constant" c: this parameter ensures that the PageRank is robust to manipulations. Apart from the speed of convergence of the power iteration method, this is yet another reason to choose the value of c not too close to 1.

We also would like to mention one more meaning of c that we have not encountered in the literature so far. It is well known that the Web has a so-called bow-tie structure (Ref.[5]) with one gigantic Strongly Connected Component accompanied by In and Out components. Roughly, the pages in the Out component receive links from other pages but do not link back. Such a bow-tie structure induces a Markov chain where only some pages in the Out component constitute the set of recurrent states. Note that even if the stationary distribution of such a Markov chain were uniquely defined, one cannot use $c = 1$ for ranking the pages. Indeed, if $c = 1$, then the PageRank of transient states, including all pages in the strongly connected component, is simply their stationary probabilities, which are equal to zero. Such a ranking obviously does not make sense. Thus, the parameter $c < 1$ is needed not only to ensure the fast convergence of the power iteration method but also for obtaining reasonable values of the PageRank. The effect of the value of c on PageRank is a promising future research direction.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of the Netherlands Organization for Scientific Research (NWO) under the grants VGP 61-520 and Meervoud 632.002.401, and the French Organization EGIDE under Van Gogh grant No. 05433UD.

We would like to thank Peter Taylor and the anonymous referees for their interesting and constructive comments that helped to improve the presentation of the paper.

REFERENCES

1. Arasu, A.; Novak, J.; Tomkins, A.; Tomlin, J. PageRank computation and the structure of the Web: experiments and algorithms. In *Proceedings of the 11th International World Wide Web Conference,* Honolulu, USA, May 2002.

2. Bianchini, M.; Gori, M.; Scarselli, F. Inside PageRank. ACM Trans. Internet Technology **2005**, *5*, 92–128.
3. Bradley, J.T.; De Jager, D.V.; Knottenbeltand, W.J.; Trifunovic, A. Hypergraph Partitioning for Faster Parallel PageRank Computation. Available at http://www.doc.ic.ac.uk/dvd03.
4. Brin, S.; Page, L.; Motwami, R.; Winograd, T. The PageRank citation ranking: bringing order to the web. Stanford University Technical Report, 1998.
5. Broder, A.Z.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J.L. Graph structure in the Web. Computer Networks **2000**, *33*, 309–320.
6. Golub, G.H.; Van Loan, C.F. *Matrix Computations*, 3rd ed.; Johns Hopkins, Baltimore, 1996.
7. Haveliwala, T.H.; Kamvar, S.D. The second eigenvalue of the Google matrix, Stanford University Technical Report, March, 2003.
8. Ipsen, I.C.F.; Kirkland, S. Convergence analysis of a PageRank updating algorithm by Langville and Meyer. SIAM J. Matrix Anal. Appl. **2006**, *27* (4), 952–967.
9. Kamvar, S.D.; Haveliwala, T.H.; Manning, C.D.; Golub, G.H. Exploiting the block structure of the Web for computing PageRank. Stanford University Technical Report, 2003.
10. Kamvar, S.D.; Haveliwala, T.H.; Manning, C.D.; Golub, G.H. Extrapolation methods for accelerating PageRank computations. In *Proceedings of the 12th International World Wide Web Conference,* Budapest, Hungary, May, 2003.
11. Kemeny, J.G.; Snell, J.L. Finite Markov Chains; *The University Series in Undergraduate Mathematics;* Van Nostrand, Princeton, NJ, 1960.
12. Langville, A.N.; Meyer, C.D. Deeper inside PageRank. Internet Mathematics Journal **2003**, *1* (3), 335–380.
13. Langville, A.N.; Meyer, C.D. Updating PageRank with iterative aggregation. In *Proceedings of the 13th World Wide Web Conference,* New York, USA, May 2004.
14. Langville, A.N.; Meyer, C.D. A reordering for the PageRank problem. NCSU CRSC Technical Report, 2004.
15. Lee, C.P.-C.; Golub, G.H.; Zenios, S.A. A fast two-stage algorithm for computing PageRank. Stanford University Technical Report, 2004.
16. Moler, C.D.; Moler, K.A. *Numerical Computing with MATLAB*; SIAM: Philadelphia, PA, 2003.
17. Norris, J.R. *Markov Chains*; Cambridge University Press, 1997.
18. Puterman, M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*; Wiley, New York, 1994.
19. www.toolbar.google.com, Accessed July 2004.
20. www.searchenginewatch.com, Accessed July 2004.
21. www.webmasterworld.com, Accessed July 2004.