Available online at www.sciencedirect.com
ScienceDirect
Linear Algebra and its Applications 429 (2008) 1254–1276
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Maximizing PageRank via outlinks

Cristobald de Kerchove*, Laure Ninove, Paul van Dooren

CESAME, Université catholique de Louvain, Avenue Georges Lemaître 4-6, B-1348 Louvain-la-Neuve, Belgium

Received 30 October 2006; accepted 18 January 2008
Available online 10 March 2008
Submitted by A. Ran

Abstract

We analyze linkage strategies for a set $\mathscr{I}$ of webpages for which the webmaster wants to maximize the sum of Google's PageRank scores. The webmaster can only choose the hyperlinks *starting* from the webpages of $\mathscr{I}$ and has no control on the hyperlinks from other webpages. We provide an optimal linkage strategy under some reasonable assumptions.

AMS classification: 15A18; 15A48; 15A51; 60J15; 68U35

Keywords: PageRank; Google matrix; Markov chain; Perron vector; Optimal linkage strategy

1. Introduction

PageRank, a measure of webpages' relevance introduced by Brin and Page, is at the heart of the well known search engine Google [6,15]. Google classifies the webpages according to the pertinence scores given by PageRank, which are computed from the graph structure of the Web. A page with a high PageRank will appear among the first items in the list of pages corresponding to a particular query.

If we look at the popularity of Google, it is not surprising that some webmasters want to increase the PageRank of their webpages in order to get more visits from websurfers to their website. Since PageRank is based on the link structure of the Web, it is therefore useful to understand how addition or deletion of hyperlinks influence it.

* Corresponding author.
E-mail addresses: c.dekerchove@uclouvain.be (C. de Kerchove), laure.ninove@uclouvain.be (L. Ninove), paul.vandooren@uclouvain.be (P. van Dooren).

0024-3795/$ - see front matter
doi:10.1016/j.laa.2008.01.023

Mathematical analysis of PageRank's sensitivity with respect to perturbations of the matrix describing the webgraph is a topical subject of interest (see for instance [2,5,11,12,13,14] and the references therein). Normwise and componentwise conditioning bounds [11] as well as the derivative [12,13] are used to understand the sensitivity of the PageRank vector. It appears that the PageRank vector is relatively insensitive to small changes in the graph structure, at least when these changes concern webpages with a low PageRank score [5,12]. One could think therefore that trying to modify its PageRank via changes in the link structure of the Web is a waste of time. However, what is important for webmasters is not the values of the PageRank vector but the *ranking* that ensues from it. Lempel and Morel [14] showed that PageRank is not rank-stable, i.e. small modifications in the link structure of the webgraph may cause dramatic changes in the ranking of the webpages. Therefore, the question of how the PageRank of a particular page or set of pages could be increased – even slightly – by adding or removing links to the webgraph remains of interest.

As it is well known [1,9], if a hyperlink from a page i to a page j is added, without no other modification in the Web, then the PageRank of j will increase. But in general, you do not have control on the *inlinks* of your webpage unless you pay another webmaster to add a hyperlink from his/her page to your or you make an *alliance* with him/her by trading a link for a link [3,8]. But it is natural to ask how you could modify your PageRank by yourself. This leads to analyze how the choice of the *outlinks* of a page can influence its own PageRank. Sydow [17] showed via numerical simulations that adding well chosen outlinks to a webpage may increase significantly its PageRank ranking. Avrachenkov and Litvak [2] analyzed theoretically the possible effect of new outlinks on the PageRank of a page and its neighbors. Supposing that a webpage has control only on its outlinks, they gave the optimal linkage strategy for this single page. Bianchini et al. [5] as well as Avrachenkov and Litvak in [1] consider the impact of links between web communities (websites or sets of related webpages), respectively on the sum of the PageRanks and on the individual PageRank scores of the pages of some community. They give general rules in order to have a PageRank as high as possible but they do not provide an optimal link structure for a website.

Our aim in this paper is to find a generalization of Avrachenkov–Litvak's optimal linkage strategy [2] to the case of *a website with several pages*. We consider a given set of pages and suppose we have only control on the *outlinks* of these pages. We are interested in the problem of *maximizing the sum of the PageRanks* of these pages.

Suppose $\mathscr{G} = (\mathscr{N}, \mathscr{E})$ be the webgraph, with a set of nodes $\mathscr{N} = \{1, \dots, n\}$ and a set of links $\mathscr{E} \subseteq \mathscr{N} \times \mathscr{N}$. For a subset of nodes $\mathscr{I} \subseteq \mathscr{N}$, we define

$$\mathscr{E}_{\mathscr{I}} = \{(i, j) \in \mathscr{E}\colon i, j \in \mathscr{I}\} \text{ the set of internal links,}$$

$$\mathscr{E}_{\text{out}(\mathscr{I})} = \{(i, j) \in \mathscr{E}\colon i \in \mathscr{I}, j \notin \mathscr{I}\} \text{ the set of external outlinks,}$$

$$\mathscr{E}_{\text{in}(\mathscr{I})} = \{(i, j) \in \mathscr{E}\colon i \notin \mathscr{I}, j \in \mathscr{I}\} \text{ the set of external inlinks,}$$

$$\mathscr{E}_{\overline{\mathscr{I}}} = \{(i, j) \in \mathscr{E}\colon i, j \notin \mathscr{I}\} \text{ the set of external links.}$$

If we do not impose any condition on $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\text{out}(\mathscr{I})}$, the problem of maximizing the sum of the PageRanks of pages of $\mathscr{I}$ is quite trivial and does not have much interest (see the discussion in Section 4). Therefore, when characterizing optimal link structures, we will make the following *accessibility assumption*: every page of the website must have an access to the rest of the Web.

Our first main result concerns the *optimal outlink structure* for a given website. In the case where the subgraph corresponding to the website is strongly connected, Theorem 10 can be particularized as follows.

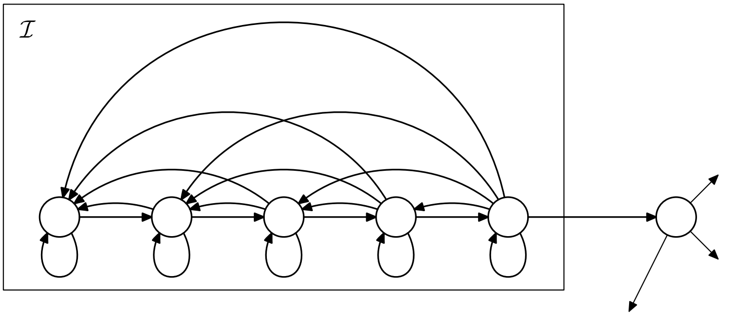

Fig. 1. Every optimal linkage strategy for a set $\mathscr{I}$ of five pages must have this structure.

Theorem 1. *Let $\mathscr{E}_{\mathscr{I}}, \mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Suppose that the subgraph $(\mathscr{I}, \mathscr{E}_{\mathscr{I}})$ is strongly connected and $\mathscr{E}_{\mathscr{I}} \neq \emptyset$. Then every optimal outlink structure $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ is to have only one outlink to a particular page outside of $\mathscr{I}$.*

We are also interested in the optimal *internal* link structure for a website. In the case where there is a unique leaking node in the website, that is only one node linking to the rest of the web, Theorem 11 can be particularized as follows.

Theorem 2. *Let $\mathscr{E}_{\mathrm{out}(\mathscr{I})}, \mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Suppose that there is only one leaking node in $\mathscr{I}$. Then every optimal internal link structure $\mathscr{E}_{\mathscr{I}}$ is composed of a forward chain of links together with every possible backward link.*

Putting together Theorems 10 and 11, we get in Theorem 12 the *optimal link structure* for a website. This optimal structure is illustrated in Fig. 1.

Theorem 3. *Let $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Then, for every optimal link structure, $\mathscr{E}_{\mathscr{I}}$ is composed of a forward chain of links together with every possible backward link, and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ consists of a unique outlink, starting from the last node of the chain.*

This paper is organized as follows. In the following preliminary section, we recall some graph concepts as well as the definition of the PageRank, and we introduce some notations. In Section 3, we develop tools for analysing the PageRank of a set of pages $\mathscr{I}$. Then we come to the main part of this paper: in Section 4 we provide the optimal linkage strategy for a set of nodes. In Section 5, we give some extensions and variants of the main theorems. We end this paper with some concluding remarks.

2. Graphs and PageRank

Let $\mathscr{G} = (\mathscr{N}, \mathscr{E})$ be a directed graph representing the Web. The webpages are represented by the set of nodes $\mathscr{N} = \{1, \ldots, n\}$ and the hyperlinks are represented by the set of directed links $\mathscr{E} \subseteq \mathscr{N} \times \mathscr{N}$. That means that $(i, j) \in \mathscr{E}$ if and only if there exists a hyperlink linking page i to page j.

Let us first briefly recall some usual concepts about directed graphs (see for instance [4]). A link (i, j) is said to be an *outlink* for node i and an *inlink* for node j. If $(i, j) \in \mathscr{E}$, node i is called a *parent* of node j. By

$$j \leftarrow i,$$

we mean that j belongs to the set of *children* of i, that is $j \in \{k \in \mathcal{N} : (i, k) \in \mathcal{E}\}$. The *outdegree* d_i of a node i is its number of children, that is

$$d_i = |\{j \in \mathcal{N} : (i, j) \in \mathcal{E}\}|.$$

A *path* from i_0 to i_s is a sequence of nodes $\langle i_0, i_1, \ldots, i_s \rangle$ such that $(i_k, i_{k+1}) \in \mathcal{E}$ for every $k = 0, 1, \ldots, s-1$. A node i *has an access to a node* j if there exists a path from i to j. In this paper, we will also say that a node i *has an access to a set* $\mathcal{J}$ if i has an access to at least one node $j \in \mathcal{J}$. The graph $\mathcal{G}$ is *strongly connected* if every node of $\mathcal{N}$ has an access to every other node of $\mathcal{N}$. A set of nodes $\mathcal{F} \subseteq \mathcal{N}$ is a *final class* of the graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ if the subgraph $(\mathcal{F}, \mathcal{E}_{\mathcal{F}})$ is strongly connected and moreover $\mathcal{E}_{\text{out}(\mathcal{F})} = \emptyset$ (i.e. nodes of $\mathcal{F}$ do not have an access to $\mathcal{N} \setminus \mathcal{F}$).

Let us now briefly introduce the PageRank score (see [5,6,12,13,15] for background). Without loss of generality (please refer to the book of Langville and Meyer [13] or the survey of Bianchini et al. [5] for details), we can make the assumption that *each node has at least one outlink*, i.e. $d_i \neq 0$ for every $i \in \mathcal{N}$. Therefore the $n \times n$ stochastic matrix $P = [P_{ij}]_{i,j \in \mathcal{N}}$ given by

$$P_{ij} = \begin{cases} d_i^{-1} & \text{if } (i, j) \in \mathcal{E}, \\ 0 & \text{otherwise} \end{cases}$$

is well defined and is a scaling of the adjacency matrix of $\mathcal{G}$. Let also $0 < c < 1$ be a *damping factor* and z be a positive stochastic *personalization vector*, i.e. $z_i > 0$ for all $i = 1, \ldots, n$ and $z^{\mathrm{T}}\mathbf{1} = 1$, where $\mathbf{1}$ denotes the vector of all ones. The *Google matrix* is then defined as

$$G = cP + (1-c)\mathbf{1}z^{\mathrm{T}}.$$

Since $z > 0$ and $c < 1$, this stochastic matrix is positive, i.e. $G_{ij} > 0$ for all i, j. The *PageRank vector* $\boldsymbol{\pi}$ is then defined as the unique invariant measure of the matrix G, that is the unique left Perron vector of G,

$$\begin{aligned} \boldsymbol{\pi}^{\mathrm{T}} &= \boldsymbol{\pi}^{\mathrm{T}} G, \\ \boldsymbol{\pi}^{\mathrm{T}}\mathbf{1} &= 1. \end{aligned} \tag{1}$$

The *PageRank of a node* i is the ith entry $\pi_i = \boldsymbol{\pi}^{\mathrm{T}} e_i$ of the PageRank vector.

The PageRank vector is usually interpreted as the stationary distribution of the following Markov chain (see for instance [13]): a random surfer moves on the webgraph, using hyperlinks between pages with a probability c and *zapping* to some new page according to the personalization vector with a probability $(1-c)$. The Google matrix G is the probability transition matrix of this random walk. In this stochastic interpretation, the PageRank of a node is equal to the inverse of its mean return time, that is π_i^{-1} is the mean number of steps a random surfer starting in node i will take for coming back to i (see [7,10]).

3. PageRank of a website

We are interested in characterizing the *PageRank of a set* $\mathcal{I}$. We define this as the sum

$$\boldsymbol{\pi}^{\mathrm{T}} e_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \pi_i,$$

where $e_{\mathcal{I}}$ denotes the vector with a 1 in the entries of $\mathcal{I}$ and 0 elsewhere. Note that the PageRank of a set corresponds to the notion of energy of a community in [5].

Let $\mathscr{I} \subseteq \mathscr{N}$ be a subset of the nodes of the graph. The PageRank of $\mathscr{I}$ can be expressed as $\boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}} = (1-c)\boldsymbol{z}^{\mathrm{T}}(I - cP)^{-1}\boldsymbol{e}_{\mathscr{I}}$ from PageRank equations (1). Let us then define the vector

$$\boldsymbol{v} = (I - cP)^{-1}\boldsymbol{e}_{\mathscr{I}}. \tag{2}$$

With this, we have the following expression for the PageRank of the set $\mathscr{I}$:

$$\boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}} = (1-c)\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}. \tag{3}$$

The vector $\boldsymbol{v}$ will play a crucial role throughout this paper. In this section, we will first present a probabilistic interpretation for this vector and prove some of its properties. We will then show how it can be used in order to analyze the influence of some page $i \in \mathscr{I}$ on the PageRank of the set $\mathscr{I}$. We will end this section by briefly introducing the concept of basic absorbing graph, which will be useful in order to analyze optimal linkage strategies under some assumptions.

3.1. Mean number of visits before zapping

Let us first see how the entries of the vector $\boldsymbol{v} = (I - cP)^{-1}\boldsymbol{e}_{\mathscr{I}}$ can be interpreted. Let us consider a random surfer on the webgraph $\mathscr{G}$ that, as described in Section 2, follows the hyperlinks of the webgraph with a probability c. But, instead of zapping to some page of $\mathscr{G}$ with a probability $(1-c)$, he *stops* his walk with probability $(1-c)$ at each step of time. This is equivalent to consider a random walk on the extended graph $\mathscr{G}_e = (\mathscr{N} \cup \{n+1\}, \mathscr{E} \cup \{(i, n+1) : i \in \mathscr{N}\})$ with a transition probability matrix

$$P_e = \begin{pmatrix} cP & (1-c)\mathbf{1} \\ 0 & 1 \end{pmatrix}.$$

At each step of time, with probability $1-c$, the random surfer can *disappear* from the original graph, that is he can reach the absorbing node $n+1$.

The nonnegative matrix $(I - cP)^{-1}$ is commonly called the fundamental matrix of the absorbing Markov chain defined by P_e (see for instance [10,16]). In the extended graph $\mathscr{G}_e$, the entry $[(I - cP)^{-1}]_{ij}$ is the expected number of visits to node j before reaching the absorbing node $n+1$ when starting from node i. From the point of view of the standard random surfer described in Section 2, the entry $[(I - cP)^{-1}]_{ij}$ is the expected number of visits to node j before zapping for the first time when starting from node i.

Therefore, the vector $\boldsymbol{v}$ defined in Eq. (2) has the following probabilistic interpretation. The entry $\boldsymbol{v}_i$ is the *expected number of visits to the set* $\mathscr{I}$ *before zapping* for the first time when the random surfer starts his walk in node i.

Now, let us first prove some simple properties about this vector.

Lemma 1. *Let* $\boldsymbol{v} \in \mathbb{R}^n_{\geqslant 0}$ *be defined by* $\boldsymbol{v} = cP\boldsymbol{v} + \boldsymbol{e}_{\mathscr{I}}$. *Then,*

(a) $\max_{i \notin \mathscr{I}} \boldsymbol{v}_i \leqslant c \max_{i \in \mathscr{I}} \boldsymbol{v}_i$;
(b) $\boldsymbol{v}_i \leqslant 1 + c\boldsymbol{v}_i$ *for all* $i \in \mathscr{N}$; *with equality if and only if the node* i *does not have an access to* $\overline{\mathscr{I}}$;
(c) $\boldsymbol{v}_i \geqslant \min_{j \leftarrow i} \boldsymbol{v}_j$ *for all* $i \in \mathscr{I}$; *with equality if and only if the node* i *does not have an access to* $\overline{\mathscr{I}}$.

Proof

(a) Since $c < 1$, for all $i \notin \mathscr{I}$,

$$\max_{i \notin \mathscr{I}} \boldsymbol{v}_i = \max_{i \notin \mathscr{I}} \left(c \sum_{j \leftarrow i} \frac{\boldsymbol{v}_j}{d_i} \right) \leqslant c \max_j \boldsymbol{v}_j.$$

Since $c < 1$, it then follows that $\max_j \boldsymbol{v}_j = \max_{i \in \mathscr{I}} \boldsymbol{v}_i$.

(b) The inequality $\boldsymbol{v}_i \leqslant \frac{1}{1-c}$ follows directly from

$$\max_i \boldsymbol{v}_i \leqslant \max_i \left(1 + c \sum_{j \leftarrow i} \frac{\boldsymbol{v}_j}{d_i} \right) \leqslant 1 + c \max_j \boldsymbol{v}_j.$$

From (a) it then also follows that $\boldsymbol{v}_i \leqslant \frac{c}{1-c}$ for all $i \notin \mathscr{I}$. Now, let $i \in \mathscr{N}$ such that $\boldsymbol{v}_i = \frac{1}{1-c}$. Then $i \in \mathscr{I}$. Moreover,

$$1 + c\boldsymbol{v}_i = \boldsymbol{v}_i = 1 + c \sum_{j \leftarrow i} \frac{\boldsymbol{v}_j}{d_i},$$

that is $\boldsymbol{v}_j = \frac{1}{1-c}$ for every $j \leftarrow i$. Hence node j must also belong to $\mathscr{I}$. By induction, every node k such that i has an access to k must belong to $\mathscr{I}$.

(c) Let $i \in \mathscr{I}$. Then, by (b)

$$1 + c\boldsymbol{v}_i \geqslant \boldsymbol{v}_i = 1 + c \sum_{j \leftarrow i} \frac{\boldsymbol{v}_j}{d_i} \geqslant 1 + c \min_{j \leftarrow i} \boldsymbol{v}_j,$$

so $\boldsymbol{v}_i \geqslant \min_{j \leftarrow i} \boldsymbol{v}_j$ for all $i \in \mathscr{I}$. If $\boldsymbol{v}_i = \min_{j \leftarrow i} \boldsymbol{v}_j$ then also $1 + c\boldsymbol{v}_i = \boldsymbol{v}_i$ and hence, by (b), the node i does not have an access to $\overline{\mathscr{I}}$. $\square$

Let us denote the set of nodes of $\overline{\mathscr{I}}$ which on average give the most visits to $\mathscr{I}$ before zapping by

$$\mathscr{V} = \mathrm{argmax}_{j \in \overline{\mathscr{I}}} \boldsymbol{v}_j.$$

Then the following lemma is quite intuitive. It says that, among the nodes of $\overline{\mathscr{I}}$, those which provide the higher mean number of visits to $\mathscr{I}$ are parents of $\mathscr{I}$, i.e. parents of some node of $\mathscr{I}$.

Lemma 2 (Parents of $\mathscr{I}$). *If $\mathscr{E}_{\mathrm{in}(\mathscr{I})} \neq \emptyset$, then*

$$\mathscr{V} \subseteq \{ j \in \overline{\mathscr{I}} \colon \textit{there exists } \ell \in \mathscr{I} \textit{ such that } (j, \ell) \in \mathscr{E}_{\mathrm{in}(\mathscr{I})} \}.$$

If $\mathscr{E}_{\mathrm{in}(\mathscr{I})} = \emptyset$, then $\boldsymbol{v}_j = 0$ for every $j \in \overline{\mathscr{I}}$.

Proof. Suppose first that $\mathscr{E}_{\mathrm{in}(\mathscr{I})} \neq \emptyset$. Let $k \in \mathscr{V}$ with $\boldsymbol{v} = (I - cP)^{-1} \boldsymbol{e}_{\mathscr{I}}$. If we supposed that there does not exist $\ell \in \mathscr{I}$ such that $(k, \ell) \in \mathscr{E}_{\mathrm{in}(\mathscr{I})}$, then we would have, since $\boldsymbol{v}_k > 0$,

$$\boldsymbol{v}_k = c \sum_{j \leftarrow k} \frac{\boldsymbol{v}_j}{d_k} \leqslant c \max_{j \notin \mathscr{I}} \boldsymbol{v}_j = c\boldsymbol{v}_k < \boldsymbol{v}_k,$$

which is a contradiction. Now, if $\mathscr{E}_{\text{in}(\mathscr{I})} = \emptyset$, then there is no access to $\mathscr{I}$ from $\overline{\mathscr{I}}$, so clearly $\boldsymbol{v}_j = 0$ for every $j \in \overline{\mathscr{I}}$. $\square$

Lemma 2 shows that the nodes $j \in \overline{\mathscr{I}}$ which provide the higher value of $\boldsymbol{v}_j$ must belong to the set of parents of $\mathscr{I}$. The converse is not true, as we will see in the following example: some parents of $\mathscr{I}$ can provide a lower mean number of visits to $\mathscr{I}$ that other nodes which are not parents of $\mathscr{I}$. In other word, Lemma 2 gives a necessary but not sufficient condition in order to maximize the entry $\boldsymbol{v}_j$ for some $j \in \overline{\mathscr{I}}$.

Example 1. Let us see on an example that having $(j, i) \in \mathscr{E}_{\text{in}(\mathscr{I})}$ for some $i \in \mathscr{I}$ is not sufficient to have $j \in \mathscr{V}$. Consider the graph in Fig. 2. Let $\mathscr{I} = \{1\}$ and take a damping factor $c = 0.85$. For $\boldsymbol{v} = (I - cP)^{-1}\boldsymbol{e}_1$, we have

$$\boldsymbol{v}_2 = \boldsymbol{v}_3 = \boldsymbol{v}_4 = 4.359 > \boldsymbol{v}_5 = 3.521 > \boldsymbol{v}_6 = 3.492 > \boldsymbol{v}_7 > \cdots > \boldsymbol{v}_{11},$$

so $\mathscr{V} = \{2, 3, 4\}$. As ensured by Lemma 2, every node of the set $\mathscr{V}$ is a parent of node 1. But here, $\mathscr{V}$ does not contain all parents of node 1. Indeed, the node $6 \notin \mathscr{V}$ while it is a parent of 1 and is moreover its parent with the lowest outdegree. Moreover, we see in this example that node 5, which is a not a parent of node 1 but a parent of node 6, gives a higher value of the expected number of visits to $\mathscr{I}$ before zapping, than node 6, parent of 1. Let us try to get some intuition about that. When starting from node 6, a random surfer has probability one half to reach node 1 in only one step. But he has also a probability one half to move to node 11 and to be send far away from node 1. On the other side, when starting from node 5, the random surfer cannot reach node 1 in only one step. But with probability 3/4 he will reach one of the nodes 2, 3 or 4 in one step. And from these nodes, the websurfer stays very near to node 1 and cannot be sent far away from it.

In the next lemma, we show that from some node $i \in \mathscr{I}$ which has an access to $\overline{\mathscr{I}}$, there always exists what we call a *decreasing path* to $\overline{\mathscr{I}}$. That is, we can find a path such that the mean number of visits to $\mathscr{I}$ is higher when starting from some node of the path than when starting from the successor of this node in the path.

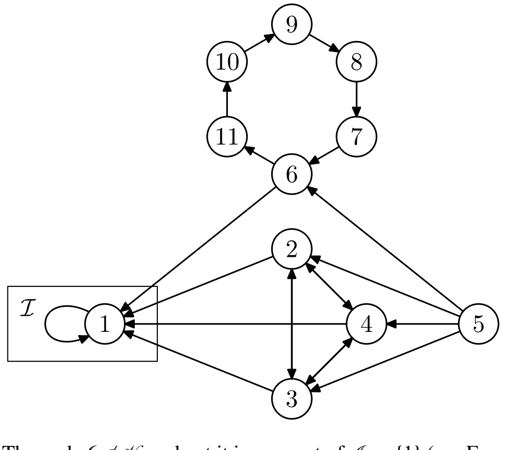

Fig. 2. The node $6 \notin \mathscr{V}$ and yet it is a parent of $\mathscr{I} = \{1\}$ (see Example 1).

Lemma 3 (Decreasing paths to $\overline{\mathscr{I}}$). *For every $i_0 \in \mathscr{I}$ which has an access to $\overline{\mathscr{I}}$, there exists a path $\langle i_0, i_1, \ldots, i_s \rangle$ with $i_1, \ldots, i_{s-1} \in \mathscr{I}$ and $i_s \in \overline{\mathscr{I}}$ such that*

$$\boldsymbol{v}_{i_0} > \boldsymbol{v}_{i_1} > \cdots > \boldsymbol{v}_{i_s}.$$

Proof. Let us simply construct a decreasing path recursively by

$$i_{k+1} \in \operatorname{argmin}_{j \leftarrow i_k} \boldsymbol{v}_j,$$

as long as $i_k \in \mathscr{I}$. If i_k has an access to $\overline{\mathscr{I}}$, then $\boldsymbol{v}_{i_{k+1}} < \boldsymbol{v}_{i_k} < \frac{1}{1-c}$ by Lemma 1(b) and (c), so the node i_{k+1} has also an access to $\overline{\mathscr{I}}$. By assumption, i_0 has an access to $\overline{\mathscr{I}}$. Moreover, the set $\mathscr{I}$ has a finite number of elements, so there must exist an s such that $i_s \in \overline{\mathscr{I}}$. □

3.2. *Influence of the outlinks of a node*

We will now see how a modification of the outlinks of some node $i \in \mathscr{N}$ can change the PageRank of a subset of nodes $\mathscr{I} \subseteq \mathscr{N}$. So we will compare two graphs on $\mathscr{N}$ defined by their set of links, $\mathscr{E}$ and $\widetilde{\mathscr{E}}$, respectively.

Every item corresponding to the graph defined by the set of links $\widetilde{\mathscr{E}}$ will be written with a tilde symbol. So $\widetilde{P}$ denotes its scaled adjacency matrix, $\tilde{\boldsymbol{\pi}}$ the corresponding PageRank vector, $\tilde{d}_i = |\{j : (i, j) \in \widetilde{\mathscr{E}}\}|$ the outdegree of some node i in this graph, $\widetilde{\boldsymbol{v}} = (I - c\widetilde{P})^{-1} \boldsymbol{e}_{\mathscr{I}}$ and $\widetilde{\mathscr{V}} = \operatorname{argmax}_{j \in \overline{\mathscr{I}}} \widetilde{\boldsymbol{v}}_j$. Finally, by $j \,\tilde{\leftarrow}\, i$ we mean $j \in \{k : (i, k) \in \widetilde{\mathscr{E}}\}$.

So, let us consider two graphs defined, respectively, by their set of links $\mathscr{E}$ and $\widetilde{\mathscr{E}}$. Suppose that they differ only in the links starting from some given node i, that is $\{j : (k, j) \in \mathscr{E}\} = \{j : (k, j) \in \widetilde{\mathscr{E}}\}$ for all $k \neq i$. Then their scaled adjacency matrices P and $\widetilde{P}$ are linked by a rank one correction. Let us then define the vector

$$\boldsymbol{\delta} = \sum_{j \tilde{\leftarrow} i} \frac{\boldsymbol{e}_j}{\tilde{d}_i} - \sum_{j \leftarrow i} \frac{\boldsymbol{e}_j}{d_i},$$

which gives the correction to apply to the line i of the matrix P in order to get $\widetilde{P}$.

Now let us first express the difference between the PageRank of $\mathscr{I}$ for two configurations differing only in the links starting from some node i. Note that in the following lemma the personalization vector $\boldsymbol{z}$ does not appear explicitly in the expression of $\tilde{\boldsymbol{\pi}}$.

Lemma 4. *Let two graphs defined respectively by $\mathscr{E}$ and $\widetilde{\mathscr{E}}$ and let $i \in \mathscr{N}$ such that for all $k \neq i$, $\{j : (k, j) \in \mathscr{E}\} = \{j : (k, j) \in \widetilde{\mathscr{E}}\}$. Then*

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}} = \boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}} + c\boldsymbol{\pi}_i \frac{\boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{v}}{1 - c\boldsymbol{\delta}^{\mathrm{T}} (I - cP)^{-1} \boldsymbol{e}_i}.$$

Proof. Clearly, the scaled adjacency matrices are linked by $\widetilde{P} = P + \boldsymbol{e}_i \boldsymbol{\delta}^{\mathrm{T}}$. Since $c < 1$, the matrix $(I - cP)^{-1}$ exists and the PageRank vectors can be expressed as

$$\boldsymbol{\pi}^{\mathrm{T}} = (1 - c)\boldsymbol{z}^{\mathrm{T}} (I - cP)^{-1},$$
$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}} = (1 - c)\boldsymbol{z}^{\mathrm{T}} (I - c(P + \boldsymbol{e}_i \boldsymbol{\delta}^{\mathrm{T}}))^{-1}.$$

Applying the Sherman–Morrison formula to $((I - cP) - ce_i\boldsymbol{\delta}^{\mathrm{T}})^{-1}$, we get

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}} = (1-c)\boldsymbol{z}^{\mathrm{T}}(I - cP)^{-1} + (1-c)\boldsymbol{z}^{\mathrm{T}}(I - cP)^{-1}\boldsymbol{e}_i \frac{c\boldsymbol{\delta}^{\mathrm{T}}(I - cP)^{-1}}{1 - c\boldsymbol{\delta}^{\mathrm{T}}(I - cP)^{-1}\boldsymbol{e}_i}$$

and the result follows immediately. □

Let us now give an equivalent condition in order to increase the PageRank of $\mathscr{I}$ by changing outlinks of some node i. The PageRank of $\mathscr{I}$ increases essentially when the new set of links favors nodes giving a higher mean number of visits to $\mathscr{I}$ before zapping.

Theorem 5 (PageRank and mean number of visits before zapping). *Let two graphs defined respectively by $\mathscr{E}$ and $\widetilde{\mathscr{E}}$ and let $i \in \mathscr{N}$ such that for all $k \neq i$, $\{j : (k, j) \in \mathscr{E}\} = \{j : (k, j) \in \widetilde{\mathscr{E}}\}$. Then*

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} > \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} \quad \text{if and only if } \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{v} > 0$$

and $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ if and only if $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{v} = 0$.

Proof. Let us first show that $\boldsymbol{\delta}^{\mathrm{T}}(I - cP)^{-1}\boldsymbol{e}_i \leqslant 1$ is always verified. Let $\boldsymbol{u} = (I - cP)^{-1}\boldsymbol{e}_i$. Then $\boldsymbol{u} - cP\boldsymbol{u} = \boldsymbol{e}_i$ and, by Lemma 1(a), $\boldsymbol{u}_j \leqslant \boldsymbol{u}_i$ for all j. So

$$\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{u} = \sum_{j \,\tilde{\leftarrow}\, i} \frac{\boldsymbol{u}_j}{\tilde{d}_i} - \sum_{j \leftarrow i} \frac{\boldsymbol{u}_j}{d_i} \leqslant \boldsymbol{u}_i - \sum_{j \leftarrow i} \frac{\boldsymbol{u}_j}{d_i} \leqslant \boldsymbol{u}_i - c\sum_{j \leftarrow i} \frac{\boldsymbol{u}_j}{d_i} = 1.$$

Now, since $c < 1$ and $\boldsymbol{\pi} > 0$, the conclusion follows by Lemma 4. □

The following Proposition 6 shows how to add a new link (i, j) starting from a given node i in order to increase the PageRank of the set $\mathscr{I}$. The PageRank of $\mathscr{I}$ increases as soon as a node $i \in \mathscr{I}$ adds a link to a node j with a larger or equal expected number of visits to $\mathscr{I}$ before zapping.

Proposition 6 (Adding a link). *Let $i \in \mathscr{I}$ and let $j \in \mathscr{N}$ be such that $(i, j) \notin \mathscr{E}$ and $\boldsymbol{v}_i \leqslant \boldsymbol{v}_j$. Let $\widetilde{\mathscr{E}} = \mathscr{E} \cup \{(i, j)\}$. Then*

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} \geqslant \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$$

with equality if and only if the node i does not have an access to $\overline{\mathscr{I}}$.

Proof. Let $i \in \mathscr{I}$ and let $j \in \mathscr{N}$ be such that $(i, j) \notin \mathscr{E}$ and $\boldsymbol{v}_i \leqslant \boldsymbol{v}_j$. Then

$$1 + c\sum_{k \leftarrow i} \frac{\boldsymbol{v}_k}{d_i} = \boldsymbol{v}_i \leqslant 1 + c\boldsymbol{v}_i \leqslant 1 + c\boldsymbol{v}_j$$

with equality if and only if i does not have an access to $\overline{\mathscr{I}}$ by Lemma 1(b). Let $\widetilde{\mathscr{E}} = \mathscr{E} \cup \{(i, j)\}$. Then

$$\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{v} = \frac{1}{d_i + 1}\left(\boldsymbol{v}_j - \sum_{k \leftarrow i} \frac{\boldsymbol{v}_k}{d_i}\right) \geqslant 0$$

with equality if and only if i does not have an access to $\overline{\mathscr{I}}$. The conclusion follows from Theorem 5. □

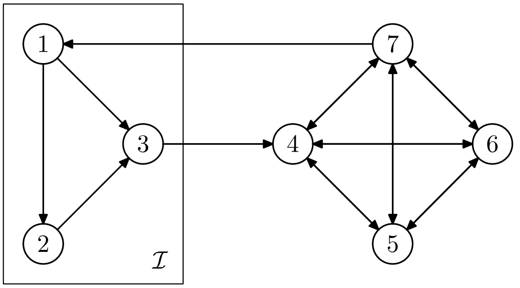

Fig. 3. For $\mathscr{I} = \{1, 2, 3\}$, removing link $(1, 2)$ gives $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} < \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$, even if $\boldsymbol{v}_1 > \boldsymbol{v}_2$ (see Example 2).

Now let us see how to remove a link (i, j) starting from a given node i in order to increase the PageRank of the set $\mathscr{I}$. If a node $i \in \mathscr{N}$ removes a link to its worst child from the point of view of the expected number of visits to $\mathscr{I}$ before zapping, then the PageRank of $\mathscr{I}$ increases.

Proposition 7 (Removing a link). *Let $i \in \mathscr{N}$ and let $j \in \operatorname{argmin}_{k \leftarrow i} \boldsymbol{v}_k$. Let $\widetilde{\mathscr{E}} = \mathscr{E} \setminus \{(i, j)\}$. Then*

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} \geqslant \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$$

with equality if and only if $\boldsymbol{v}_k = \boldsymbol{v}_j$ for every k such that $(i, k) \in \mathscr{E}$.

Proof. Let $i \in \mathscr{N}$ and let $j \in \operatorname{argmin}_{k \leftarrow i} \boldsymbol{v}_k$. Let $\widetilde{\mathscr{E}} = \mathscr{E} \setminus \{(i, j)\}$. Then

$$\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{v} = \sum_{k \leftarrow i} \frac{\boldsymbol{v}_k - \boldsymbol{v}_j}{d_i(d_i - 1)} \geqslant 0$$

with equality if and only if $\boldsymbol{v}_k = \boldsymbol{v}_j$ for all $k \leftarrow i$. The conclusion follows by Theorem 5. □

In order to increase the PageRank of $\mathscr{I}$ with a new link (i, j), Proposition 6 only requires that $\boldsymbol{v}_j \leqslant \boldsymbol{v}_i$. On the other side, Proposition 7 requires that $\boldsymbol{v}_j = \min_{k \leftarrow i} \boldsymbol{v}_k$ in order to increase the PageRank of $\mathscr{I}$ by deleting link (i, j). One could wonder whether or not this condition could be weakened to $\boldsymbol{v}_j < \boldsymbol{v}_i$, so as to have symmetric conditions for the addition or deletion of links. In fact, this cannot be done as shown in the following example.

Example 2. Let us see by an example that the condition $j \in \operatorname{argmin}_{k \leftarrow i} \boldsymbol{v}_k$ in Proposition 7 cannot be weakened to $\boldsymbol{v}_j < \boldsymbol{v}_i$. Consider the graph in Fig. 3 and take a damping factor $c = 0.85$. Let $\mathscr{I} = \{1, 2, 3\}$. We have

$$\boldsymbol{v}_1 = 2.63 > \boldsymbol{v}_2 = 2.303 > \boldsymbol{v}_3 = 1.533.$$

As ensured by Proposition 7, if we remove the link $(1, 3)$, the PageRank of $\mathscr{I}$ increases (e.g. from 0.199 to 0.22 with a uniform personalization vector $\boldsymbol{z} = \frac{1}{n}\mathbf{1}$), since $3 \in \operatorname{argmin}_{k \leftarrow 1} \boldsymbol{v}_k$. But, if we remove instead the link $(1, 2)$, the PageRank of $\mathscr{I}$ decreases (from 0.199 to 0.179 with $\boldsymbol{z}$ uniform) even if $\boldsymbol{v}_2 < \boldsymbol{v}_1$.

Remark 1. Let us note that, if the node i does not have an access to the set $\overline{\mathscr{I}}$, then for every *deletion* of a link starting from i, the PageRank of $\mathscr{I}$ will not be modified. Indeed, in this case $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{v} = 0$ since by Lemma 1(b), $\boldsymbol{v}_j = \frac{1}{1-c}$ for every $j \leftarrow i$.

3.3. Basic absorbing graph

Now, let us introduce briefly the notion of basic absorbing graph (see Chapter III about absorbing Markov chains in Kemeny and Snell's book [10]).

For a given graph $(\mathscr{N}, \mathscr{E})$ and a specified subset of nodes $\mathscr{I} \subseteq \mathscr{N}$, the *basic absorbing graph* is the graph $(\mathscr{N}, \mathscr{E}^0)$ defined by $\mathscr{E}^0_{\text{out}(\mathscr{I})} = \emptyset$, $\mathscr{E}^0_{\mathscr{I}} = \{(i,i): i \in \mathscr{I}\}$, $\mathscr{E}^0_{\text{in}(\mathscr{I})} = \mathscr{E}_{\text{in}(\mathscr{I})}$ and $\mathscr{E}^0_{\overline{\mathscr{I}}} = \mathscr{E}_{\overline{\mathscr{I}}}$. In other words, the basic absorbing graph $(\mathscr{N}, \mathscr{E}^0)$ is a graph constructed from $(\mathscr{N}, \mathscr{E})$, keeping the same sets of external inlinks and external links $\mathscr{E}_{\text{in}(\mathscr{I})}, \mathscr{E}_{\overline{\mathscr{I}}}$, removing the external outlinks $\mathscr{E}_{\text{out}(\mathscr{I})}$ and changing the internal link structure $\mathscr{E}_{\mathscr{I}}$ in order to have only self-links for nodes of $\mathscr{I}$.

Like in the previous subsection, every item corresponding to the basic absorbing graph will have a zero symbol. For instance, we will write $\boldsymbol{\pi}_0$ for the PageRank vector corresponding to the basic absorbing graph and $\mathscr{V}_0 = \operatorname{argmax}_{j \in \overline{\mathscr{I}}}[(I - cP_0)^{-1}\boldsymbol{e}_{\mathscr{I}}]_j$.

Proposition 8 (PageRank for a basic absorbing graph). *Let a graph defined by a set of links $\mathscr{E}$ and let $\mathscr{I} \subseteq \mathscr{N}$. Then*

$$\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} \leqslant \boldsymbol{\pi}_0^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$$

with equality if and only if $\mathscr{E}_{\text{out}(\mathscr{I})} = \emptyset$.

Proof. Up to a permutation of the indices, Eq. (2) can be written as

$$\begin{pmatrix} I - cP_{\mathscr{I}} & -cP_{\text{out}(\mathscr{I})} \\ -cP_{\text{in}(\mathscr{I})} & I - cP_{\overline{\mathscr{I}}} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_{\mathscr{I}} \\ \boldsymbol{v}_{\overline{\mathscr{I}}} \end{pmatrix} = \begin{pmatrix} \mathbf{1} \\ 0 \end{pmatrix},$$

so we get

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{v}_{\mathscr{I}} \\ c(I - cP_{\overline{\mathscr{I}}})^{-1} P_{\text{in}(\mathscr{I})} \boldsymbol{v}_{\mathscr{I}} \end{pmatrix}. \tag{4}$$

By Lemma 1(b) and since $(I - cP_{\overline{\mathscr{I}}})^{-1}$ is a nonnegative matrix (see for instance the chapter on M-matrices in Berman and Plemmons's book [4]), we then have

$$\boldsymbol{v} \leqslant \begin{pmatrix} \frac{1}{1-c}\mathbf{1} \\ \frac{c}{1-c}(I - cP_{\overline{\mathscr{I}}})^{-1} P_{\text{in}(\mathscr{I})}\mathbf{1} \end{pmatrix} = \boldsymbol{v}_0$$

with equality if and only if no node of $\mathscr{I}$ has an access to $\overline{\mathscr{I}}$, that is $\mathscr{E}_{\text{out}(\mathscr{I})} = \emptyset$. The conclusion now follows from Eq. (3) and $\boldsymbol{z} > 0$. $\square$

Let us finally prove a nice property of the set $\mathscr{V}$ when $\mathscr{I} = \{i\}$ is a singleton: it is independent of the outlinks of i. In particular, it can be found from the basic absorbing graph.

Lemma 9. *Let a graph defined by a set of links $\mathscr{E}$ and let $\mathscr{I} = \{i\}$ Then there exists an $\alpha \neq 0$ such that $(I - cP)^{-1}\boldsymbol{e}_i = \alpha(I - cP_0)^{-1}\boldsymbol{e}_i$. As a consequence,*

$$\mathscr{V} = \mathscr{V}_0.$$

Proof. Let $\mathscr{I} = \{i\}$. Since $\boldsymbol{v}_{\mathscr{I}} = \boldsymbol{v}_i$ is a scalar, it follows from Eq. (4) that the direction of the vector $\boldsymbol{v}$ does not depend on $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\text{out}(\mathscr{I})}$ but only on $\mathscr{E}_{\text{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$. $\square$

4. Optimal linkage strategy for a website

In this section, we consider a set of nodes $\mathscr{I}$. For this set, we want to choose the sets of internal links $\mathscr{E}_{\mathscr{I}} \subseteq \mathscr{I} \times \mathscr{I}$ and external outlinks $\mathscr{E}_{\mathrm{out}(\mathscr{I})} \subseteq \mathscr{I} \times \overline{\mathscr{I}}$ in order to maximize the PageRank score of $\mathscr{I}$, that is $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$.

Let us first discuss about the constraints on $\mathscr{E}$ we will consider. If we do not impose any condition on $\mathscr{E}$, the problem of maximizing $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ is quite trivial. As shown by Proposition 8, you should take in this case $\mathscr{E}_{\mathrm{out}(\mathscr{I})} = \emptyset$ and $\mathscr{E}_{\mathscr{I}}$ an arbitrary subset of $\mathscr{I} \times \mathscr{I}$ such that each node has at least one outlink. You just try to lure the random walker to your pages, not allowing him to leave $\mathscr{I}$ except by zapping according to the preference vector. Therefore, it seems sensible to impose that $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ *must be nonempty*.

Now, let us show that, in order to avoid trivial solutions to our maximization problem, it is not enough to assume that $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ must be nonempty. Indeed, with this single constraint, in order to lose as few as possible visits from the random walker, you should take a unique leaking node $k \in \mathscr{I}$ (i.e. $\mathscr{E}_{\mathrm{out}(\mathscr{I})} = \{(k, \ell)\}$ for some $\ell \in \overline{\mathscr{I}}$) and isolate it from the rest of the set $\mathscr{I}$ (i.e. $\{i \in \mathscr{I} : (i, k) \in \mathscr{E}_{\mathscr{I}}\} = \emptyset$).

Moreover, it seems reasonable to imagine that Google penalizes (or at least tries to penalize) such behavior in the context of spam alliances [8].

All this discussion leads us to make the following assumption.

Assumption A (*Accessibility*). Every node of $\mathscr{I}$ has an access to at least one node of $\overline{\mathscr{I}}$.

Let us now explain the basic ideas we will use in order to determine an optimal linkage strategy for a set of webpages $\mathscr{I}$. We determine some forbidden patterns for an optimal linkage strategy and deduce the only possible structure an optimal strategy can have. In other words, we assume that we have a configuration which gives an optimal PageRank $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$. Then we prove that if some particular pattern appeared in this optimal structure, then we could construct another graph for which the PageRank $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ is strictly higher than $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$.

We will firstly determine the shape of an optimal external outlink structure $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$, when the internal link structure $\mathscr{E}_{\mathscr{I}}$ is given, in Theorem 10. Then, given the external outlink structure $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ we will determine the possible optimal internal link structure $\mathscr{E}_{\mathscr{I}}$ in Theorem 11. Finally, we will put both results together in Theorem 12 in order to get the general shape of an optimal linkage strategy for a set $\mathscr{I}$ when $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ are given.

Proofs of this section will be illustrated by several figures for which we take the following drawing convention.

Convention. When nodes are drawn from left to right on the same horizontal line, they are arranged by decreasing value of $\boldsymbol{v}_j$. Links are represented by continuous arrows and paths by dashed arrows.

The first result of this section concerns the optimal *outlink* structure $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ for the set $\mathscr{I}$, while its internal structure $\mathscr{E}_{\mathscr{I}}$ is given. An example of optimal outlink structure is given after the theorem.

Theorem 10 (Optimal outlink structure). *Let $\mathscr{E}_{\mathscr{I}}$, $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Let $\mathscr{F}_1, \ldots, \mathscr{F}_r$ be the final classes of the subgraph $(\mathscr{I}, \mathscr{E}_{\mathscr{I}})$. Let $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ such that the PageRank $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ is maximal under Assumption A. Then $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ has the following structure*:

$$\mathcal{E}_{\mathrm{out}(\mathcal{I})} = \mathcal{E}_{\mathrm{out}(\mathcal{F}_1)} \cup \cdots \cup \mathcal{E}_{\mathrm{out}(\mathcal{F}_r)},$$

where for every $s = 1, \ldots, r$,

$$\mathcal{E}_{\mathrm{out}(\mathcal{F}_s)} \subseteq \{(i, j) \colon i \in \operatorname{argmin}_{k \in \mathcal{F}_s} \boldsymbol{v}_k \text{ and } j \in \mathcal{V}\}.$$

Moreover for every $s = 1, \ldots, r$, *if* $\mathcal{E}_{\mathcal{F}_s} \neq \emptyset$, *then* $|\mathcal{E}_{\mathrm{out}(\mathcal{F}_s)}| = 1$.

Proof. Let $\mathcal{E}_{\mathcal{I}}$, $\mathcal{E}_{\mathrm{in}(\mathcal{I})}$ and $\mathcal{E}_{\overline{\mathcal{I}}}$ be given. Suppose $\mathcal{E}_{\mathrm{out}(\mathcal{I})}$ is such that $\boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathcal{I}}$ is maximal under Assumption A.

We will determine the possible leaking nodes of $\mathcal{I}$ by analyzing three different cases.

Firstly, let us consider some node $i \in \mathcal{I}$ such that i does not have children in $\mathcal{I}$, i.e. $\{k \in \mathcal{I} \colon (i, k) \in \mathcal{E}_{\mathcal{I}}\} = \emptyset$. Then clearly we have $\{i\} = \mathcal{F}_s$ for some $s = 1, \ldots, r$, with $i \in \operatorname{argmin}_{k \in \mathcal{F}_s} \boldsymbol{v}_k$ and $\mathcal{E}_{\mathcal{F}_s} = \emptyset$. From Assumption A, we have $\mathcal{E}_{\mathrm{out}(\mathcal{F}_s)} \neq \emptyset$, and from Theorem 5 and the optimality assumption, we have $\mathcal{E}_{\mathrm{out}(\mathcal{F}_s)} \subseteq \{(i, j) \colon j \in \mathcal{V}\}$ (see Fig. 4).

Secondly, let us consider some $i \in \mathcal{I}$ such that i has children in $\mathcal{I}$, i.e. $\{k \in \mathcal{I} \colon (i, k) \in \mathcal{E}_{\mathcal{I}}\} \neq \emptyset$ and

$$\boldsymbol{v}_i \leqslant \min_{\substack{k \leftarrow i \\ k \in \mathcal{I}}} \boldsymbol{v}_k.$$

Let $j \in \operatorname{argmin}_{k \leftarrow i} \boldsymbol{v}_k$. Then $j \in \overline{\mathcal{I}}$ and $\boldsymbol{v}_j < \boldsymbol{v}_i$ by Lemma 1(*c*). Suppose by contradiction that the node i would keep an access to $\overline{\mathcal{I}}$ if we took $\widetilde{\mathcal{E}}_{\mathrm{out}(\mathcal{I})} = \mathcal{E}_{\mathrm{out}(\mathcal{I})} \setminus \{(i, j)\}$ instead of $\mathcal{E}_{\mathrm{out}(\mathcal{I})}$. Then, by Proposition 7, considering $\widetilde{\mathcal{E}}_{\mathrm{out}(\mathcal{I})}$ instead of $\mathcal{E}_{\mathrm{out}(\mathcal{I})}$ would increase strictly the PageRank of $\mathcal{I}$ while Assumption A remains satisfied (see Fig. 5). This would contradict the optimality assumption for $\mathcal{E}_{\mathrm{out}(\mathcal{I})}$. From this, we conclude that

- the node i belongs to final class $\mathcal{F}_s$ of the subgraph $(\mathcal{I}, \mathcal{E}_{\mathcal{I}})$ with $\mathcal{E}_{\mathcal{F}_s} \neq \emptyset$ for some $s = 1, \ldots, r$;
- there does not exist another $\ell \in \overline{\mathcal{I}}$, $\ell \neq j$ such that $(i, \ell) \in \mathcal{E}_{\mathrm{out}(\mathcal{I})}$;
- there does not exist another k in the same final class $\mathcal{F}_s$, $k \neq i$ such that such that $(k, \ell) \in \mathcal{E}_{\mathrm{out}(\mathcal{I})}$ for some $\ell \in \overline{\mathcal{I}}$.

Again, by Theorem 5 and the optimality assumption, we have $j \in \mathcal{V}$ (see Fig. 4).

Let us now notice that

$$\max_{k \in \overline{\mathcal{I}}} \boldsymbol{v}_k < \min_{k \in \mathcal{I}} \boldsymbol{v}_k. \tag{5}$$

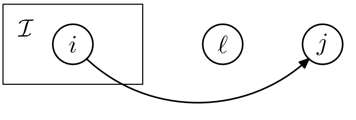

Fig. 4. If $\boldsymbol{v}_j < \boldsymbol{v}_\ell$, then $\tilde{\boldsymbol{\pi}}^{\mathrm{T}} \boldsymbol{e}_{\mathcal{I}} > \boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathcal{I}}$ with $\widetilde{\mathcal{E}}_{\mathrm{out}(\mathcal{I})} = \mathcal{E}_{\mathrm{out}(\mathcal{I})} \cup \{(i, \ell)\} \setminus \{(i, j)\}$.

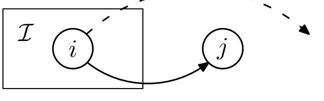

Fig. 5. If $\boldsymbol{v}_j = \min_{k \leftarrow i} \boldsymbol{v}_k$ and i has another access to $\overline{\mathcal{I}}$, then $\tilde{\boldsymbol{\pi}}^{\mathrm{T}} \boldsymbol{e}_{\mathcal{I}} > \boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathcal{I}}$ with $\widetilde{\mathcal{E}}_{\mathrm{out}(\mathcal{I})} = \mathcal{E}_{\mathrm{out}(\mathcal{I})} \setminus \{(i, j)\}$.

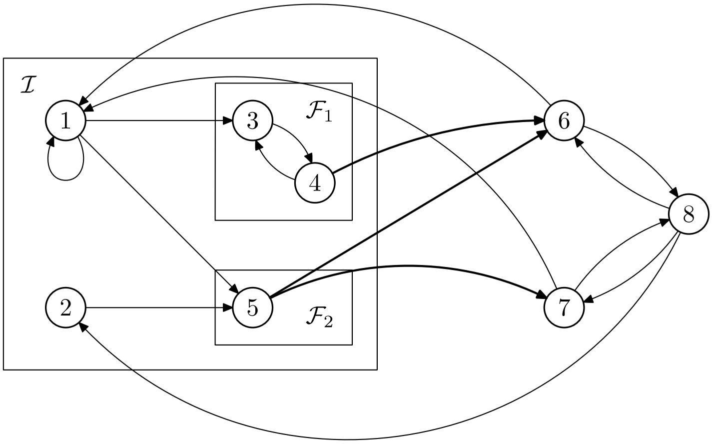

Fig. 6. Bold arrows represent one of the six optimal *outlink* structures for this configuration with two final classes (see Example 3).

Indeed, with $i \in \operatorname{argmin}_{k\in\mathscr{I}} \boldsymbol{v}_k$, we are in one of the two cases analyzed above for which we have seen that $\boldsymbol{v}_i > \boldsymbol{v}_j = \operatorname{argmax}_{k\in\overline{\mathscr{I}}} \boldsymbol{v}_k$.

Finally, consider a node $i \in \mathscr{I}$ that does not belong to any of the final classes of the subgraph $(\mathscr{I}, \mathscr{E}_{\mathscr{I}})$. Suppose by contradiction that there exists $j \in \overline{\mathscr{I}}$ such that $(i, j) \in \mathscr{E}_{\mathrm{out}(\mathscr{I})}$. Let $\ell \in \operatorname{argmin}_{k\leftarrow i} \boldsymbol{v}_k$. Then it follows from inequality (5) that $\ell \in \overline{\mathscr{I}}$. But the same argument as above shows that the link $(i, \ell) \in \mathscr{E}_{\mathrm{out}(\mathscr{I})}$ must be removed since $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ is supposed to be optimal (see Fig. 5 again). So, there does not exist $j \in \overline{\mathscr{I}}$ such that $(i, j) \in \mathscr{E}_{\mathrm{out}(\mathscr{I})}$ for a node $i \in \mathscr{I}$ which does not belong to any of the final classes $\mathscr{F}_1, \ldots, \mathscr{F}_r$. $\square$

Example 3. Let us consider the graph given in Fig. 6. The internal link structure $\mathscr{E}_{\mathscr{I}}$, as well as $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ are given. The subgraph $(\mathscr{I}, \mathscr{E}_{\mathscr{I}})$ has two final classes $\mathscr{F}_1$ and $\mathscr{F}_2$. With $c = 0.85$ and $\boldsymbol{z}$ the uniform probability vector, this configuration has six optimal outlink structures (one of these solutions is represented by bold arrows in Fig. 6). Each one can be written as $\mathscr{E}_{\mathrm{out}(\mathscr{I})} = \mathscr{E}_{\mathrm{out}(\mathscr{F}_1)} \cup \mathscr{E}_{\mathrm{out}(\mathscr{F}_2)}$, with $\mathscr{E}_{\mathrm{out}(\mathscr{F}_1)} = \{(4, 6)\}$ or $\mathscr{E}_{\mathrm{out}(\mathscr{F}_1)} = \{(4, 7)\}$ and $\emptyset \neq \mathscr{E}_{\mathrm{out}(\mathscr{F}_2)} \subseteq \{(5, 6), (5, 7)\}$. Indeed, since $\mathscr{E}_{\mathscr{F}_1} \neq \emptyset$, as stated by Theorem 10, the final class $\mathscr{F}_1$ has exactly one external outlink in every optimal outlink structure. On the other hand, the final class $\mathscr{F}_2$ may have several external outlinks, since it is composed of a unique node and moreover this node does not have a self-link. Note that $\mathscr{V} = \{6, 7\}$ in each of these six optimal configurations, but this set $\mathscr{V}$ cannot be determined a priori since it depends on the chosen outlink structure.

Now, let us determine the optimal *internal* link structure $\mathscr{E}_{\mathscr{I}}$ for the set $\mathscr{I}$, while its outlink structure $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ is given. Examples of optimal internal structure are given after the proof of the theorem.

Theorem 11 (Optimal internal link structure). *Let $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$, $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Let $\mathscr{L} = \{i \in \mathscr{I} : (i, j) \in \mathscr{E}_{\mathrm{out}(\mathscr{I})}$ for some $j \in \overline{\mathscr{I}}\}$ be the set of leaking nodes of $\mathscr{I}$ and let $n_{\mathscr{L}} = |\mathscr{L}|$ be the number of leaking nodes. Let $\mathscr{E}_{\mathscr{I}}$ such that the PageRank $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ is maximal under Assumption A. Then there exists a permutation of the indices such that $\mathscr{I} = \{1, 2, \ldots, n_{\mathscr{I}}\}$, $\mathscr{L} = \{n_{\mathscr{I}} - n_{\mathscr{L}} + 1, \ldots, n_{\mathscr{I}}\}$*

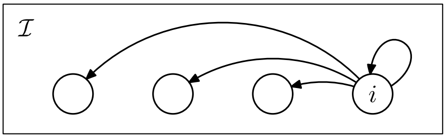

Fig. 7. Every $i \in \mathscr{I}$ must link to every $j \in \mathscr{I}$ with $\boldsymbol{v}_j \geqslant \boldsymbol{v}_i$.

$$\boldsymbol{v}_1 > \cdots > \boldsymbol{v}_{n_\mathscr{I}-n_\mathscr{L}} > \boldsymbol{v}_{n_\mathscr{I}-n_\mathscr{L}+1} \geqslant \cdots \geqslant \boldsymbol{v}_{n_\mathscr{I}}$$

and $\mathscr{E}_\mathscr{I}$ *has the following structure*:

$$\mathscr{E}_\mathscr{I}^L \subseteq \mathscr{E}_\mathscr{I} \subseteq \mathscr{E}_\mathscr{I}^U,$$

where

$$\mathscr{E}_\mathscr{I}^L = \{(i,j) \in \mathscr{I} \times \mathscr{I}: j \leqslant i\} \cup \{(i,j) \in (\mathscr{I} \setminus \mathscr{L}) \times \mathscr{I}: j = i+1\},$$
$$\mathscr{E}_\mathscr{I}^U = \mathscr{E}_\mathscr{I}^L \cup \{(i,j) \in \mathscr{L} \times \mathscr{L}: i < j\}.$$

Proof. Let $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$, $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Suppose $\mathscr{E}_\mathscr{I}$ is such that $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_\mathscr{I}$ is maximal under Assumption A.

Firstly, by Proposition 6 and since every node of $\mathscr{I}$ has an access to $\overline{\mathscr{I}}$, every node $i \in \mathscr{I}$ links to every node $j \in \mathscr{I}$ such that $\boldsymbol{v}_j \geqslant \boldsymbol{v}_i$ (see Fig. 7), that is

$$\{(i,j) \in \mathscr{E}_\mathscr{I}: \boldsymbol{v}_i \leqslant \boldsymbol{v}_j\} = \{(i,j) \in \mathscr{I} \times \mathscr{I}: \boldsymbol{v}_i \leqslant \boldsymbol{v}_j\}. \tag{6}$$

Secondly, let $(k,i) \in \mathscr{E}_\mathscr{I}$ such that $k \neq i$ and $k \in \mathscr{I} \setminus \mathscr{L}$. Let us prove that, if the node i has an access to $\overline{\mathscr{I}}$ by a path $\langle i, i_1, \ldots, i_s\rangle$ such that $i_j \neq k$ for all $j = 1, \ldots, s$ and $i_s \in \overline{\mathscr{I}}$, then $\boldsymbol{v}_i < \boldsymbol{v}_k$ (see Fig. 8). Indeed, if we had $\boldsymbol{v}_k \leqslant \boldsymbol{v}_i$ then, by Lemma 1(c), there would exists $\ell \in \mathscr{I}$ such that $(k,\ell) \in \mathscr{E}_\mathscr{I}$ and $\boldsymbol{v}_\ell = \min_{j \leftarrow k} \boldsymbol{v}_j < \boldsymbol{v}_i \leqslant \boldsymbol{v}_k$. But, with $\widetilde{\mathscr{E}}_\mathscr{I} = \mathscr{E}_\mathscr{I} \setminus \{(k,\ell)\}$, we would have $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_\mathscr{I} > \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_\mathscr{I}$ by Proposition 7 while Assumption A remains satisfied since the node k would keep access to $\overline{\mathscr{I}}$ via the node i (see Fig. 9). That contradicts the optimality assumption. This leads us to the conclusion that $\boldsymbol{v}_k > \boldsymbol{v}_i$ for every $k \in \mathscr{I} \setminus \mathscr{L}$ and $i \in \mathscr{L}$. Moreover $\boldsymbol{v}_i \neq \boldsymbol{v}_k$ for every $i, k \in \mathscr{I} \setminus \mathscr{L}$, $i \neq k$. Indeed, if we had $\boldsymbol{v}_i = \boldsymbol{v}_k$, then $(k,i) \in \mathscr{E}_\mathscr{I}$ by (6) while by Lemma 3, the node i would have an access to $\overline{\mathscr{I}}$ by a path independant from k. So we should have $\boldsymbol{v}_i < \boldsymbol{v}_k$.

We conclude from this that we can relabel the nodes of $\mathscr{N}$ such that $\mathscr{I} = \{1, 2, \ldots, n_\mathscr{I}\}$, $\mathscr{L} = \{n_\mathscr{I} - n_\mathscr{L} + 1, \ldots, n_\mathscr{I}\}$ and

$$\boldsymbol{v}_1 > \boldsymbol{v}_2 > \cdots > \boldsymbol{v}_{n_\mathscr{I}-n_\mathscr{L}} > \boldsymbol{v}_{n_\mathscr{I}-n_\mathscr{L}+1} \geqslant \cdots \geqslant \boldsymbol{v}_{n_\mathscr{I}}. \tag{7}$$

It follows also that, for $i \in \mathscr{I} \setminus \mathscr{L}$ and $j > i$, $(i,j) \in \mathscr{E}_\mathscr{I}$if and only if $j = i+1$. Indeed, suppose first $i < n_\mathscr{I} - n_\mathscr{L}$. Then, we cannot have $(i,j) \in \mathscr{E}_\mathscr{I}$ with $j > i+1$ since in this case we would contradict the ordering of the nodes given by Eq. (7) (see Fig. 8 again with $k = i+1$ and remember that by Lemma 3, node j has an access to $\overline{\mathscr{I}}$ by a decreasing path). Moreover, node i must link to some node $j > i$ in order to satisfy Assumption A, so $(i, i+1)$ must belong to $\mathscr{E}_\mathscr{I}$. Now, consider the case $i = n_\mathscr{I} - n_\mathscr{L}$. Suppose we had $(i,j) \in \mathscr{E}_\mathscr{I}$ with $j > i+1$. Let us first note that there cannot exist two or more different links (i,ℓ) with $\ell \in \mathscr{L}$ since in this case we could remove one of these links and increase strictly the PageRank of the set $\mathscr{I}$. If $\boldsymbol{v}_j = \boldsymbol{v}_{i+1}$, we could relabel the nodes by permuting these two indices. If $\boldsymbol{v}_j < \boldsymbol{v}_{i+1}$, then with $\widetilde{\mathscr{E}}_\mathscr{I} = \mathscr{E}_\mathscr{I} \cup \{(i, i+1)\} \setminus \{(i,j)\}$, we would have $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_\mathscr{I} > \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_\mathscr{I}$ by Theorem 5 while Assumption A remains satisfied since the i

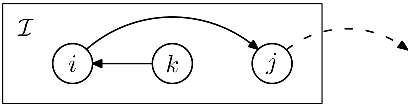

Fig. 8. The node i cannot have an access to $\overline{\mathscr{I}}$ without crossing k since in this case we should then have $\boldsymbol{v}_i < \boldsymbol{v}_k$.

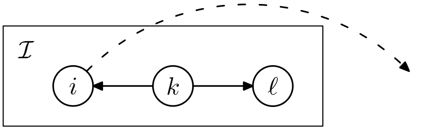

Fig. 9. If $\boldsymbol{v}_\ell = \min_{j \leftarrow k} \boldsymbol{v}_j$, then $\tilde{\boldsymbol{\pi}}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}} > \boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}}$ with $\widetilde{\mathscr{E}}_{\mathrm{out}(\mathscr{I})} = \mathscr{E}_{\mathrm{out}(\mathscr{I})} \setminus \{(k, \ell)\}$.

would keep access to $\overline{\mathscr{I}}$ via node $i+1$. That contradicts the optimality assumption. So we have proved that

$$\{(i, j) \in \mathscr{E}_{\mathscr{I}} : i < j \text{ and } i \in \mathscr{I} \setminus \mathscr{L}\} = \{(i, i+1) : i \in \mathscr{I} \setminus \mathscr{L}\}. \tag{8}$$

Thirdly, it is obvious that

$$\{(i, j) \in \mathscr{E}_{\mathscr{I}} : i < j \text{ and } i \in \mathscr{L}\} \subseteq \{(i, j) \in \mathscr{L} \times \mathscr{L} : i < j\}. \tag{9}$$

The announced structure for a set $\mathscr{E}_{\mathscr{I}}$ giving a maximal PageRank score $\boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}}$ under Assumption A now follows directly from Eqs. (6), (8) and (9). $\square$

Example 4. Let us consider the graphs given in Fig. 10. For both cases, the external outlink structure $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ with two leaking nodes, as well as $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ are given. With $c = 0.85$ and $\boldsymbol{z}$ the uniform probability vector, the optimal internal link structure for configuration (a) is given by $\mathscr{E}_{\mathscr{I}} = \mathscr{E}^L_{\mathscr{I}}$, while in configuration (b) we have $\mathscr{E}_{\mathscr{I}} = \mathscr{E}^U_{\mathscr{I}}$ (bold arrows), with $\mathscr{E}^L_{\mathscr{I}}$ and $\mathscr{E}^U_{\mathscr{I}}$ defined in Theorem 11.

Finally, combining the optimal outlink structure and the optimal internal link structure described in Theorems 10 and 11, we find the *optimal linkage strategy* for a set of webpages. Let us note that, since we have here control on both $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$, there are no more cases of several final classes or several leaking nodes to consider. For an example of optimal link structure, see Fig. 1.

Theorem 12 (Optimal link structure). *Let $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Let $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ such that $\boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}}$ is maximal under Assumption A. Then there exists a permutation of the indices such that $\mathscr{I} = \{1, 2, \ldots, n_{\mathscr{I}}\}$,*

$$\boldsymbol{v}_1 > \cdots > \boldsymbol{v}_{n_{\mathscr{I}}} > \boldsymbol{v}_{n_{\mathscr{I}}+1} \geqslant \cdots \geqslant \boldsymbol{v}_n$$

and $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ have the following structure:

$$\mathscr{E}_{\mathscr{I}} = \{(i, j) \in \mathscr{I} \times \mathscr{I} : j \leqslant i \textit{ or } j = i+1\},$$
$$\mathscr{E}_{\mathrm{out}(\mathscr{I})} = \{(n_{\mathscr{I}}, n_{\mathscr{I}}+1)\}.$$

Proof. Let $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given and suppose $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ are such that $\boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{e}_{\mathscr{I}}$ is maximal under Assumption A. Let us relabel the nodes of $\mathscr{N}$ such that $\mathscr{I} = \{1, 2, \ldots, n_{\mathscr{I}}\}$ and $\boldsymbol{v}_1 \geqslant \cdots \geqslant \boldsymbol{v}_{n_{\mathscr{I}}} > \boldsymbol{v}_{n_{\mathscr{I}}+1} = \max_{j \in \overline{\mathscr{I}}} \boldsymbol{v}_j$. By Theorem 11, $(i, j) \in \mathscr{E}_{\mathscr{I}}$ for every nodes $i, j \in \mathscr{I}$ such that $j \leqslant i$.

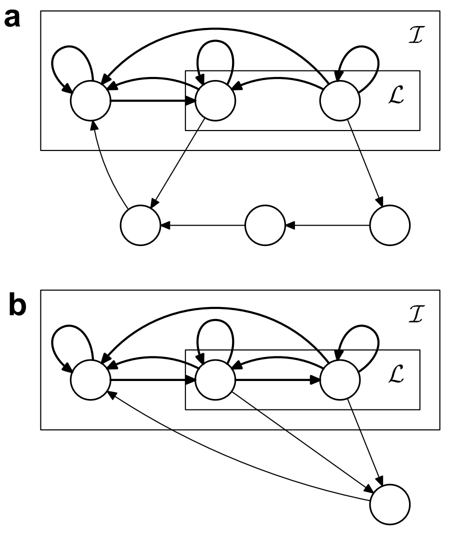

Fig. 10. Bold arrows represent optimal *internal* link structures. In (a) we have $\mathscr{E}_{\mathscr{I}} = \mathscr{E}^L_{\mathscr{I}}$, while $\mathscr{E}_{\mathscr{I}} = \mathscr{E}^U_{\mathscr{I}}$ in (b).

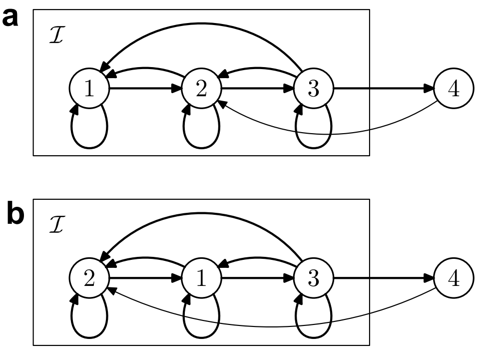

Fig. 11. For $\mathscr{I} = \{1, 2, 3\}$, $c = 0.85$ and $\boldsymbol{z}$ uniform, the link structure in (a) is not optimal and yet it satisfies the necessary conditions of Theorem 12 (see Example 5).

In particular, every node of $\mathscr{I}$ has an access to node 1. Therefore, there is a unique final class $\mathscr{F}_1 \subseteq \mathscr{I}$ in the subgraph $(\mathscr{I}, \mathscr{E}_{\mathscr{I}})$. So, by Theorem 10, $\mathscr{E}_{\mathrm{out}(\mathscr{I})} = \{(k, \ell)\}$ for some $k \in \mathscr{F}_1$ and $\ell \in \overline{\mathscr{I}}$. Without loss of generality, we can suppose that $\ell = n_{\mathscr{I}} + 1$. By Theorem 11 again, the leaking node $k = n_{\mathscr{I}}$ and therefore $(i, i+1) \in \mathscr{E}_{\mathscr{I}}$ for every node $i \in \{1, \ldots, n_{\mathscr{I}} - 1\}$. $\square$

Let us note that having a structure like described in Theorem 12 is a *necessary but not sufficient* condition in order to have a maximal PageRank.

Example 5. Let us show by an example that the graph structure given in Theorem 12 is not sufficient to have a maximal PageRank. Consider for instance the graphs in Fig. 11. Let $c = 0.85$ and a uniform personalization vector $\boldsymbol{z} = \frac{1}{n}\mathbf{1}$. Both graphs have the link structure required Theorem 12 in order to have a maximal PageRank, with $\boldsymbol{v}_{(a)} = (6.484 \quad 6.42 \quad 6.224 \quad 5.457)^{\mathrm{T}}$ and $\boldsymbol{v}_{(b)} = \begin{pmatrix}6.432 & 6.494 & 6.247 & 5.52\end{pmatrix}^{\mathrm{T}}$. But the configuration (a) is not optimal since in this case, the PageRank $\pi_{(a)}^{\mathrm{T}} e_{\mathscr{I}} = 0.922$ is strictly less than the PageRank $\pi_{(b)}^{\mathrm{T}} e_{\mathscr{I}} = 0.926$ obtained by the configuration (b). Let us nevertheless note that, with a non uniform personalization vector $z = (0.7 \quad 0.1 \quad 0.1 \quad 0.1)^{\mathrm{T}}$, the link structure (a) would be optimal.

5. Extensions and variants

Let us now present some extensions and variants of the results of the previous section. We will first emphasize the role of parents of $\mathscr{I}$. Secondly, we will briefly talk about Avrachenkov–Litvak's optimal link structure for the case where $\mathscr{I}$ is a singleton. Then we will give variants of Theorem 12 when self-links are forbidden or when a minimal number of external outlinks is required. Finally, we will make some comments of the influence of external *inlinks* on the PageRank of $\mathscr{I}$.

5.1. Linking to parents

If some node of $\mathscr{I}$ has at least one parent in $\overline{\mathscr{I}}$ then the optimal linkage strategy for $\mathscr{I}$ is to have an internal link structure like described in Theorem 12 together with a single link to one of the parents of $\mathscr{I}$.

Corollary 13 (Necessity of linking to parents). *Let* $\mathscr{E}_{\mathrm{in}(\mathscr{I})} \neq \emptyset$ *and* $\mathscr{E}_{\overline{\mathscr{I}}}$ *be given. Let* $\mathscr{E}_{\mathscr{I}}$ *and* $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ *such that* $\pi^{\mathrm{T}} e_{\mathscr{I}}$ *is maximal under Assumption A. Then* $\mathscr{E}_{\mathrm{out}(\mathscr{I})} = \{(i, j)\}$, *for some* $i \in \mathscr{I}$ *and* $j \in \overline{\mathscr{I}}$ *such that* $(j, k) \in \mathscr{E}_{\mathrm{in}(\mathscr{I})}$ *for some* $k \in \mathscr{I}$.

Proof. This is a direct consequence of Lemma 2 and Theorem 12. □

Let us nevertheless remember that not every parent of nodes of $\mathscr{I}$ will give an optimal link structure, as we have already discussed in Example 1 and we develop now.

Example 6. Let us continue Example 1. We consider the graph in Fig. 2 as basic absorbing graph for $\mathscr{I} = \{1\}$, that is $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ are given. We take $c = 0.85$ as damping factor and a uniform personalization vector $z = \frac{1}{n}\mathbf{1}$. We have seen in Example 1 than $\mathscr{V}_0 = \{2, 3, 4\}$. Let us consider the value of the PageRank π_1 for different sets $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$:

	$\mathscr{E}_{\mathrm{out}(\mathscr{I})}$				
	$\emptyset$	$\{(1,2)\}$	$\{(1,5)\}$	$\{(1,6)\}$	$\{(1,2),(1,3)\}$
$\mathscr{E}_{\mathscr{I}} = \emptyset$	/	0.1739	0.1402	0.1392	0.1739
$\mathscr{E}_{\mathscr{I}} = \{(1,1)\}$	0.5150	0.2600	0.2204	0.2192	0.2231

As expected from Corollary 15, the optimal linkage strategy for $\mathscr{I} = \{1\}$ is to have a self-link and a link to one of the nodes 2, 3 or 4. We note also that a link to node 6, which is a parent of node 1 provides a lower PageRank that a link to node 5, which is not parent of 1. Finally, if we suppose self-links are forbidden (see below), then the optimal linkage strategy is to link to one *or more* of the nodes 2–4.

In the case where no node of $\mathscr{I}$ has a parent in $\overline{\mathscr{I}}$, then *every* structure like described in Theorem 12 will give an optimal link structure.

Proposition 14 (No external parent). *Let $\mathcal{E}_{\text{in}(\mathcal{I})}$ and $\mathcal{E}_{\overline{\mathcal{I}}}$ be given. Suppose that $\mathcal{E}_{\text{in}(\mathcal{I})} = \emptyset$. Then the PageRank $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathcal{I}}$ is maximal under Assumption A if and only if*

$$\mathcal{E}_{\mathcal{I}} = \{(i, j) \in \mathcal{I} \times \mathcal{I} : j \leqslant i \text{ or } j = i + 1\},$$
$$\mathcal{E}_{\text{out}(\mathcal{I})} = \{(n_{\mathcal{I}}, n_{\mathcal{I}} + 1)\}$$

for some permutation of the indices such that $\mathcal{I} = \{1, 2, \ldots, n_{\mathcal{I}}\}$.

Proof. This follows directly from $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathcal{I}} = (1 - c)\boldsymbol{z}^{\mathrm{T}}\boldsymbol{v}$ and the fact that, if $\mathcal{E}_{\text{in}(\mathcal{I})} = \emptyset$

$$\boldsymbol{v} = (I - cP)^{-1}\boldsymbol{e}_{\mathcal{I}} = \begin{pmatrix} (I - cP_{\mathcal{I}})^{-1}\mathbf{1} \\ 0 \end{pmatrix},$$

up to a permutation of the indices. □

5.2. Optimal linkage strategy for a singleton

The optimal outlink structure for a single webpage has already been given by Avrachenkov and Litvak in [2]. Their result becomes a particular case of Theorem 12. Note that in the case of a single node, the possible choices for $\mathcal{E}_{\text{out}(\mathcal{I})}$ can be found a priori by considering the basic absorbing graph, since $\mathcal{V} = \mathcal{V}_0$.

Corollary 15 (Optimal link structure for a single node). *Let $\mathcal{I} = \{i\}$ and let $\mathcal{E}_{\text{in}(\mathcal{I})}$ and $\mathcal{E}_{\overline{\mathcal{I}}}$ be given. Then the PageRank $\boldsymbol{\pi}_i$ is maximal under Assumption A if and only if $\mathcal{E}_{\mathcal{I}} = \{(i, i)\}$ and $\mathcal{E}_{\text{out}(\mathcal{I})} = \{(i, j)\}$ for some $j \in \mathcal{V}_0$.*

Proof. This follows directly from Lemma 9 and Theorem 12. □

5.3. Optimal linkage strategy under additional assumptions

Let us consider the problem of maximizing the PageRank $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathcal{I}}$ when *self-links are forbidden*. Indeed, it seems to be often supposed that Google's PageRank algorithm does not take self-links into account. In this case, Theorem 12 can be adapted readily for the case where $|\mathcal{I}| \geqslant 2$. When $\mathcal{I}$ is a singleton, we must have $\mathcal{E}_{\mathcal{I}} = \emptyset$, so $\mathcal{E}_{\text{out}(\mathcal{I})}$ can contain *several* links, as stated in Theorem 10.

Corollary 16 (Optimal link structure with no self-links). *Suppose $|\mathcal{I}| \geqslant 2$. Let $\mathcal{E}_{\text{in}(\mathcal{I})}$ and $\mathcal{E}_{\overline{\mathcal{I}}}$ be given. Let $\mathcal{E}_{\mathcal{I}}$ and $\mathcal{E}_{\text{out}(\mathcal{I})}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathcal{I}}$ is maximal under Assumption A and assumption that there does not exist $i \in \mathcal{I}$ such that $\{(i, i)\} \in \mathcal{E}_{\mathcal{I}}$. Then there exists a permutation of the indices such that $\mathcal{I} = \{1, 2, \ldots, n_{\mathcal{I}}\}$, $\boldsymbol{v}_1 > \cdots > \boldsymbol{v}_{n_{\mathcal{I}}} > \boldsymbol{v}_{n_{\mathcal{I}}+1} \geqslant \cdots \geqslant \boldsymbol{v}_n$, and $\mathcal{E}_{\mathcal{I}}$ and $\mathcal{E}_{\text{out}(\mathcal{I})}$ have the following structure*:

$$\mathcal{E}_{\mathcal{I}} = \{(i, j) \in \mathcal{I} \times \mathcal{I} : j < i \text{ or } j = i + 1\},$$
$$\mathcal{E}_{\text{out}(\mathcal{I})} = \{(n_{\mathcal{I}}, n_{\mathcal{I}} + 1)\}.$$

Corollary 17 (Optimal link structure for a single node with no self-link). *Suppose $\mathcal{I} = \{i\}$. Let $\mathcal{E}_{\text{in}(\mathcal{I})}$ and $\mathcal{E}_{\overline{\mathcal{I}}}$ be given. Suppose $\mathcal{E}_{\mathcal{I}} = \emptyset$. Then the PageRank $\boldsymbol{\pi}_i$ is maximal under Assumption A if and only if $\emptyset \neq \mathcal{E}_{\text{out}(\mathcal{I})} \subseteq \mathcal{V}_0$.*

Let us now consider the problem of maximizing the PageRank $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ when *several external outlinks are required*. Then the proof of Theorem 10 can be adapted readily in order to have the following variant of Theorem 12.

Corollary 18 (Optimal link structure with several external outlinks). *Let $\mathscr{E}_{\mathrm{in}(\mathscr{I})}$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Let $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ is maximal under Assumption A and assumption that $|\mathscr{E}_{\mathrm{out}(\mathscr{I})}| \geqslant r$. Then there exists a permutation of the indices such that $\mathscr{I} = \{1, 2, \ldots, n_{\mathscr{I}}\}$, $\boldsymbol{v}_1 > \cdots > \boldsymbol{v}_{n_{\mathscr{I}}} > \boldsymbol{v}_{n_{\mathscr{I}}+1} \geqslant \cdots \geqslant \boldsymbol{v}_n$, and $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ have the following structure*:

$$\mathscr{E}_{\mathscr{I}} = \{(i, j) \in \mathscr{I} \times \mathscr{I} : j < i \text{ or } j = i + 1\},$$
$$\mathscr{E}_{\mathrm{out}(\mathscr{I})} = \{(n_{\mathscr{I}}, j_k) : j_k \in \mathscr{V} \text{ for } k = 1, \ldots, r\}.$$

5.4. *External inlinks*

Finally, let us make some comments about the addition of external inlinks to the set $\mathscr{I}$. It is well known that adding an inlink to a particular page always increases the PageRank of this page [1,9]. This can be viewed as a direct consequence of Theorem 5 and Lemma 1. The case of a set of several pages $\mathscr{I}$ is not so simple. We prove in the following theorem that, if the set $\mathscr{I}$ has a link structure as described in Theorem 12 then adding an inlink to a page of $\mathscr{I}$ from a page $j \in \overline{\mathscr{I}}$ which is *not* a parent of some node of $\mathscr{I}$ will increase the PageRank of $\mathscr{I}$. But in general, adding an inlink to some page of $\mathscr{I}$ from $\overline{\mathscr{I}}$ may *decrease* the PageRank of the set $\mathscr{I}$, as shown in Examples 7 and 8.

Theorem 19 (External inlinks). *Let $\mathscr{I} \subseteq \mathscr{N}$ and a graph defined by a set of links $\mathscr{E}$. If*

$$\min_{i \in \mathscr{I}} \boldsymbol{v}_i > \max_{j \notin \mathscr{I}} \boldsymbol{v}_j$$

then, for every $j \in \overline{\mathscr{I}}$ which is not a parent of $\mathscr{I}$, and for every $i \in \mathscr{I}$, the graph defined by $\widetilde{\mathscr{E}} = \mathscr{E} \cup \{(j, i)\}$ gives $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} > \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$.

Proof. This follows directly from Theorem 5. □

Example 7. Let us show by an example that a new external inlink is not always profitable for a set $\mathscr{I}$ in order to improve its PageRank, even if $\mathscr{I}$ has an optimal linkage strategy. Consider for instance the graph in Fig. 12. With $c = 0.85$ and $\boldsymbol{z}$ uniform, we have $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} = 0.8481$. But if we consider the graph defined by $\widetilde{\mathscr{E}}_{\mathrm{in}(\mathscr{I})} = \mathscr{E}_{\mathrm{in}(\mathscr{I})} \cup \{(3, 2)\}$, then we have $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} = 0.8321 < \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$.

Example 8. A new external inlink does not not always increase the PageRank of a set $\mathscr{I}$ in even if this new inlink comes from a page which is not already a parent of some node of $\mathscr{I}$. Consider

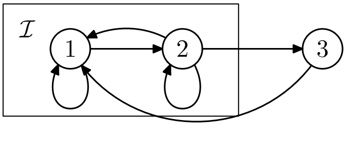

Fig. 12. For $\mathscr{I} = \{1, 2\}$, adding the external inlink $(3, 2)$ gives $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} < \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ (see Example 7).

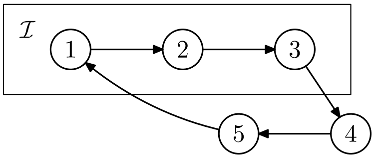

Fig. 13. For $\mathscr{I} = \{1, 2, 3\}$, adding the external inlink (4, 3) gives $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} < \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ (see Example 8).

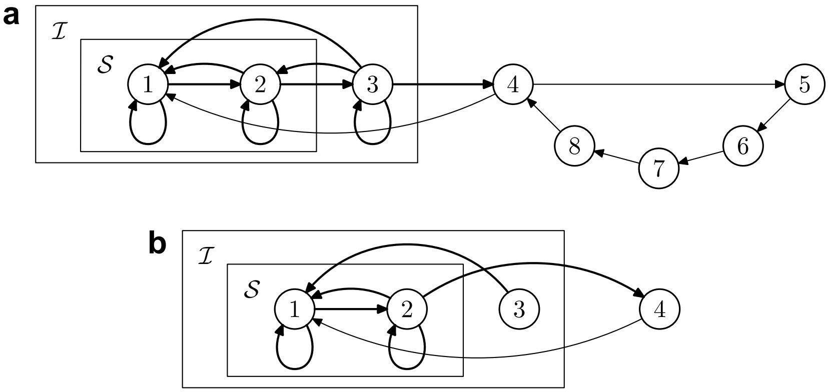

Fig. 14. In (a) and (b), bold arrows represent optimal link structures for $\mathscr{I} = \{1, 2, 3\}$ with respect to a target set $\mathscr{S} = \{1, 2\}$ (see Example 9).

for instance the graph in Fig. 13. With $c = 0.85$ and $\boldsymbol{z}$ uniform, we have $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} = 0.6$. But if we consider the graph defined by $\widetilde{\mathscr{E}}_{\mathrm{in}(\mathscr{I})} = \mathscr{E}_{\mathrm{in}(\mathscr{I})} \cup \{(4, 3)\}$, then we have $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}} = 0.5897 < \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$.

6. Conclusions

In this paper we provide the general shape of an optimal link structure for a website in order to maximize its PageRank. This structure with a forward chain and every possible backward links may be not intuitive. At our knowledge, it has never been mentioned, while topologies like a clique, a ring or a star are considered in the literature on collusion and alliance between pages [3,8]. Moreover, this optimal structure gives new insight into the affirmation of Bianchini et al. [5] that, in order to maximize the PageRank of a website, hyperlinks to the rest of the webgraph "should be in pages with a small PageRank and that have many internal hyperlinks". More precisely, we have seen that the leaking pages must be choosen with respect to the mean number of visits before zapping they give to the website, rather than their PageRank.

Let us now present some possible directions for future work.

We have noticed in Example 5 that the first node of $\mathscr{I}$ in the forward chain of an optimal link structure is not necessarily *a child* of some node of $\overline{\mathscr{I}}$. In the example we gave, the personalization vector was not uniform. We wonder if this could occur with a uniform personalization vector and make the following conjecture.

Conjecture. Let $\mathscr{E}_{\mathrm{in}(\mathscr{I})} \neq \emptyset$ and $\mathscr{E}_{\overline{\mathscr{I}}}$ be given. Let $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{I}}$ is maximal under Assumption A. If $\boldsymbol{z} = \frac{1}{n}\mathbf{1}$, then there exists $j \in \overline{\mathscr{I}}$ such that $(j, i) \in \mathscr{E}_{\mathrm{in}(\mathscr{I})}$, where $i \in \operatorname{argmax}_k \boldsymbol{v}_k$.

If this conjecture was true we could also ask if the node $j \in \overline{\mathscr{I}}$ such that $(j, i) \in \mathscr{E}_{\mathrm{in}(\mathscr{I})}$ where $i \in \operatorname{argmax}_k \boldsymbol{v}_k$ belongs to $\mathscr{V}$.

Another question concerns the optimal linkage strategy in order to maximize an arbitrary linear combination of the PageRanks of the nodes of $\mathscr{I}$. In particular, we could want to maximize the PageRank $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{S}}$ of *a target subset* $\mathscr{S} \subseteq \mathscr{I}$ by choosing $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ as usual. A general shape for an optimal link structure seems difficult to find, as shown in the following example.

Example 9. Consider the graphs in Fig. 14. In both cases, let $c = 0.85$ and $\boldsymbol{z} = \frac{1}{n}\mathbf{1}$. Let $\mathscr{I} = \{1, 2, 3\}$ and let $\mathscr{S} = \{1, 2\}$ be the target set. In the configuration (a), the optimal sets of links $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ for maximizing $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{e}_{\mathscr{S}}$ has the link structure described in Theorem 12. But in (a), the optimal $\mathscr{E}_{\mathscr{I}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{I})}$ do not have this structure. Let us note nevertheless that, by Theorem 12, the subsets $\mathscr{E}_{\mathscr{S}}$ and $\mathscr{E}_{\mathrm{out}(\mathscr{S})}$ must have the link structure described in Theorem 12.

Acknowledgements

This paper presents research supported by a grant "Actions de recherche concertées – Large Graphs and Networks" of the "Communauté Française de Belgique" and by the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The second author was supported by a research fellow Grant of the "Fonds de la Recherche Scientifique – FNRS" (Belgium). The scientific responsibility rests with the authors.

References

[1] Konstantin Avrachenkov, Nelly Litvak, Decomposition of the Google PageRank and optimal linking strategy, Tech. Report, INRIA, 2004. <http://www.inria.fr/rrrt/rr-5101.html>.

[2] Konstantin Avrachenkov, Nelly Litvak, The effect of new links on Google PageRank,Stoch. Models, 22 (2) (2006) 319–331.

[3] Ricardo Baeza-Yates, Carlos Castillo, Vicente López, PageRank increase under different collusion topologies, in: First International Workshop on Adversarial Information Retrieval on the Web, 2005. <http://airweb.cse.lehigh.edu/2005/baeza-yates.pdf>.

[4] Abraham Berman, Robert J. Plemmons, Nonnegative Matrices in the Mathematical Sciences, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), vol. 9, Philadelphia, PA, 1994.

[5] Monica Bianchini, Marco Gori, Franco Scarselli, Inside PageRank, ACM Trans. Inter. Tech. 5 (1) (2005) 92–128.

[6] Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual web search engine, Comput. Network ISDN Syst. 30 (1–7) (1998) 107–117. Proceedings of the Seventh International World Wide Web Conference, April 1998.

[7] Grace E. Cho, Carl D. Meyer, Markov chain sensitivity measured by mean first passage times, Linear Algebra Appl. 316 (1–3) (2000) 21–28. Conference Celebrating the 60th Birthday of Robert J. Plemmons (Winston–Salem, NC, 1999).

[8] Zoltán Gyöngyi, Hector Garcia-Molina, Link spam alliances, in: VLDB '05: Proceedings of the 31st International Conference on Very Large Data Bases, VLDB Endowment, 2005, pp. 517–528, <http://portal.acm.org/citation.cfm?id=1083654>.

[9] Ilse C.F. Ipsen, Rebecca S. Wills, Mathematical properties and analysis of Google's PageRank, Bol. Soc. Esp. Mat. Apl. 34 (2006) 191–196.

[10] John G. Kemeny, J. Laurie Snell, Finite Markov Chains, The University Series in Undergraduate Mathematics, D. Van Nostrand Co., Inc., Princeton, NJ–Toronto–London–New York, 1960.

[11] Steve Kirkland, Conditioning of the entries in the stationary vector of a Google-type matrix, Linear Algebra Appl. 418 (2–3) (2006) 665–681.

[12] Amy N. Langville, Carl D. Meyer, Deeper inside PageRank, Internet Math. 1 (3) (2004) 335–380.

[13] Amy N. Langville, Carl D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, Princeton, NJ, 2006.

[14] Ronny Lempel, Shlomo Moran, Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs, Inf. Retr. 8 (2) (2005) 245–264.

[15] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, The PageRank citation ranking: bringing order to the Web, Tech. Report, Computer Science Department, Stanford University, 1998. <http://dbpubs.stanford.edu:8090/pub/1999-66>.

[16] Eugene Seneta, Nonnegative matrices and Markov chains, Springer Series in Statistics, second ed., Springer-Verlag, New York, 1981.

[17] Marcin Sydow, Can One Out-link Change Your PageRank?, vol. 3528, Springer, Berlin Heidelberg, 2005, pp. 408–414.