Influence of the outlinks of a page on its PageRank

Cristobald de Kerchove [a] Laure Ninove [a] Paul van Dooren [a]

[a] *CESAME, Université catholique de Louvain, Av. Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium.*

Abstract

We analyze linkage strategies for one and several pages (forming a clique) willing to maximize their PageRank. The latter is given by the Perron vector of an irreducible stochastic matrix, namely the Google matrix. We study the case where the user can only control the outlinks of its own page or the outlinks of its own site consisting in several pages. In that way, this paper deals with the sensibility of the PageRank and formalize previous simulations that shows significant changes in the PageRank of a page that modifies its own outlinks. An optimal linkage strategy is given for a user having one page or a site where every page points to every page forming then a clique.

Key words: Google Matrix, Markov chain, Perron vector, Optimal linkage strategy

Introduction

Google has established its well-known PageRank, introduced by Page and Brin in [9], that classifies the pages of the World Wide Web by scoring each of them. A page with a high PageRank will be among the first items in the list. No surprise then that everyone wishes to maximize its own PageRank [1,5,2]. However, the only control the user has on its own page or site are the outlinks pointing to some external pages. Modifying these outlinks leads to perturbations in the PageRank and therefore to a special case of the sensitivity of the PageRank [4,7,3,6]. Ilse-Wills [6] inform us that new inlinks to some page u always increase the PageRank of u, but the adding of new outlinks from u is not necessarily decreasing for the PageRank. Sydow [11] shows in

Email addresses: dekerchove@inma.ucl.ac.be (Cristobald de Kerchove), ninove@inma.ucl.ac.be (Laure Ninove), vdooren@inma.ucl.ac.be (Paul van Dooren).

some simulations that good chosen outlinks may increase its own PageRank. That motivates the search of an optimal linkage strategies for one or several pages.

The paper is divided into three parts. First the preliminaries introduce the notations, the PageRank equations and the reduced PageRank equations for a subset of nodes. Using these equations, the second part exposes some result of optimality for one page and for a clique of pages. Finally, some remarks and open problems are given for further analysis in the topic.

1 Preliminaries

1.1 PageRank equations

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a directed webgraph, with a set of nodes $\mathcal{N} = \{1, 2, \ldots, n\}$. Without loss of generality (see for instance Bianchini et al. [4] for preconditioning concerns), we can suppose that each node of $\mathcal{N}$ has at least one outlink and no self-loop. Therefore, the adjacency matrix A of $\mathcal{G}$ has no zero row but a zero diagonal. Let $P = D^{-1}A$ be the stochastic $n \times n$ matrix which is obtained by scaling the adjacency matrix A with the inverse of the diagonal matrix D of outdegrees of all nodes. Let also $c \in]0, 1]$ be a *damping factor*, and $\mathbf{z}$ be a stochastic *personalization vector*, *i.e.* $\mathbf{z^T 1} = \mathbf{1}$. The *Google matrix* G is then defined as

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

where $\mathbf{1}$ denotes the vector of all ones, its dimension usually follows from the context. We suppose that G is irreducible, which is the case as soon as A is irreducible, or $c < 1$ and $\mathbf{z} > \mathbf{0}$. The *PageRank vector* π is then defined by the following equations

$$\pi^{\mathbf{T}}\mathbf{G} = \pi^{\mathbf{T}}, \tag{1}$$
$$\pi^{\mathbf{T}}\mathbf{1} = 1, \tag{2}$$

and is usually interpreted as the stationary distribution of a random surfer using hyperlinks between pages with a probability c and jumping to some new page according to the personalization vector with a probability $(1 - c)$. Hence, $\pi_{\mathbf{i}}$ is the weighted sum of the scores of its own parents times c, plus its personalized score $\mathbf{z_i}$ times $(1 - c)$.

1.2 Reduced PageRank equations for a subset of nodes

Throughout the paper, we will be interested in the characterization of the PageRank for a subset $\mathcal{I}$ of $n_{\mathcal{I}}$ nodes of $\mathcal{N}$. In order to write the basic equations for this subset of nodes, we need to introduce the concepts of mean return time and reduced transitions probabilities for $\mathcal{I}$.

Let $X_0, X_1, \ldots$ be a random walk on the graph defined by the Google matrix G.

Definition 1 *The $n_{\mathcal{I}} \times n_{\mathcal{I}}$ reduced transition matrix $S_{\mathcal{I}}$ of is defined by*

$$S_{ij} = \mathbb{P}(X_t = j, X_1, \ldots, X_{t-1} \in \bar{\mathcal{I}} \text{ for some } t \geq 1 \mid X_0 = i) \quad \text{for all } i, j \in \mathcal{I}.$$

Definition 2 *The $n_{\mathcal{I}}$-vector $\mathbf{t}_{\mathcal{I}}$ of mean return times in $\mathcal{I}$ is defined by*

$$\mathbf{t_i} = \mathbb{E}(\mathbf{t} \mid \mathbf{X_0} = \mathbf{i}, \mathbf{X_1}, \ldots, \mathbf{X_{t-1}} \in \bar{\mathcal{I}}, \mathbf{X_t} \in \mathcal{I}) \quad \text{for all } \mathbf{i} \in \mathcal{I}.$$

The reduced transition matrix $S_{\mathcal{I}}$ is usually called the Perron or stochastic complement. It represents the transition matrix for the reduced Markov chain on $\mathcal{I}$ which is built by hiding the transitions outside the states of $\mathcal{I}$. Its expression and properties are well known (see [8]).

Proposition 3 (Meyer [8]) *Let $S_{\mathcal{I}}$ be the reduced transition matrix of a irreducible stochastic matrix G for $\mathcal{I}$. Then $S_{\mathcal{I}}$ is an irreducible stochastic matrix and can be expressed as*

$$S_{\mathcal{I}} = G_{\mathcal{I}} + G_{\mathrm{out}(\mathcal{I})}(I - G_{\bar{\mathcal{I}}})^{-1} G_{\mathrm{in}(\mathcal{I})},$$

with $(I - G_{\bar{\mathcal{I}}})$ invertible and having a non-negative inverse usually called the fundamental matrix.

The mean return times vector $\mathbf{t}_{\mathcal{I}}$ can also be written in function of the entries of the Google matrix G.

Proposition 4 *Let $\mathbf{t}_{\mathcal{I}}$ be the mean return times vector for $\mathcal{I}$. Then*

$$\mathbf{t}_{\mathcal{I}} = \mathbf{1} + \mathbf{G}_{\mathrm{out}(\mathcal{I})}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}.$$

Proof. It is well known (see for instance [10]) that, for $i, j \in \bar{\mathcal{I}}$, the entry $((I - G_{\bar{\mathcal{I}}})^{-1})_{ij}$ is the mean number of visits to node j before reaching the set $\mathcal{I}$ when starting from i. Moreover, for all $i \in \bar{\mathcal{I}}$,

$$\begin{aligned}
&\mathbb{E}(t \mid X_0 = i, X_1, \ldots, X_{t-1} \in \bar{\mathcal{I}}, X_t \in \mathcal{I}) \\
&\quad = \sum_{j \in \bar{\mathcal{I}}} \mathbb{E}(\text{number of visits to node } j \text{ without leaving } \bar{\mathcal{I}} \mid X_0 = i) \\
&\quad = ((I - G_{\bar{\mathcal{I}}})^{-1}\mathbf{1})_{\mathbf{i}}.
\end{aligned}$$

Therefore, for all $i \in \mathcal{I}$,

$$\begin{aligned}\mathbf{t_i} &= \mathbb{E}(t \mid X_0 = i, X_1, \ldots, X_{t-1} \in \bar{\mathcal{I}}, X_t \in \mathcal{I}) \\ &= 1 + \sum_{\ell \in \bar{\mathcal{I}}} \mathbb{P}(X_1 = \ell \mid X_0 = i)\, \mathbb{E}(t \mid X_0 = \ell, X_1, \ldots, X_{t-1} \in \bar{\mathcal{I}}, X_t \in \mathcal{I}),\end{aligned}$$

hence it follows that $\mathbf{t}_\mathcal{I} = \mathbf{1} + \mathbf{G}_{\mathrm{out}(\mathcal{I})}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}$.

Now we can write *reduced PageRank equations* for the subset of node $\mathcal{I}$. The first one is well known [8], and expresses that the PageRank subvector $\pi_\mathcal{I}$ is proportional to the stationary distribution vector for the reduced transition matrix $S_\mathcal{I}$. The second one gives the proportionality ratio thanks to the mean return times vector $\mathbf{t}_\mathcal{I}$.

Proposition 5 (Meyer [8]) *The PageRank subvector $\pi_\mathcal{I}$ for the nodes of $\mathcal{I}$ satisfies*

$$\pi_\mathcal{I}^\mathbf{T}\mathbf{S}_\mathcal{I} = \pi_\mathcal{I}^\mathbf{T}.$$

Proposition 6 *The PageRank subvector $\pi_\mathcal{I}$ for the nodes of $\mathcal{I}$ satisfies*

$$\pi_\mathcal{I}^\mathbf{T}\mathbf{t}_\mathcal{I} = \mathbf{1}.$$

Proof. By Proposition 4,

$$\pi_\mathcal{I}^\mathbf{T}\mathbf{t}_\mathcal{I} = \pi_\mathcal{I}^\mathbf{T}\mathbf{1} + \pi_\mathcal{I}^\mathbf{T}\mathbf{G}_{\mathrm{out}(\mathcal{I})}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1} = \pi_\mathcal{I}^\mathbf{T}\mathbf{1} + \pi_{\bar{\mathcal{I}}}^\mathbf{T}\mathbf{1} = \mathbf{1},$$

since $\pi_{\bar{\mathcal{I}}}^\mathbf{T} = \pi_\mathcal{I}\mathbf{G}_{\mathrm{out}(\mathcal{I})}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}$ by equation 1, and $\pi_\mathcal{I}\mathbf{1} + \pi_{\bar{\mathcal{I}}}^\mathbf{T}\mathbf{1} = \mathbf{1}$, by equation 2.

2 Optimal outlink structure

We are interested in how given pages can modify their PageRank by changing their outlink structure. We will consider two cases: first, a single page wants to maximize its PageRank, and second, a set of $n_\mathcal{I}$ pages, organized in a clique, wants to maximize the sum of the PageRanks of its pages. In these two cases, the only variables are the links from the node or the clique to the rest of the graph.

We make the assumption that there exists at least one outlink from the considered set of pages $\mathcal{I}$ to the rest of the graph. Clearly, the optimal structure is reached when no outlink points to the rest of the graph turning the set of pages $\mathcal{I}$ into an absorbing state. However we can imagine that Google penalize such selfish behavior while is looking for link spam alliances [5].

The adjacency matrix A of the webgraph can therefore be written as

$$A = \begin{pmatrix} A_{\mathcal{I}} & A_{\mathrm{out}(\mathcal{I})} \\ A_{\mathrm{in}(\mathcal{I})} & A_{\bar{\mathcal{I}}} \end{pmatrix},$$

where $A_{\mathcal{I}} = \mathbf{1}\mathbf{1}^{\mathbf{T}} - \mathbf{I}$, $A_{\mathrm{in}(\mathcal{I})}$ and $A_{\bar{\mathcal{I}}}$ are given, and $A_{\mathrm{out}(\mathcal{I})} \in \{0,1\}^{n_{\mathcal{I}} \times n_{\bar{\mathcal{I}}}}$, $A_{\mathrm{out}(\mathcal{I})} \neq 0$ has to be determined.

2.1 Optimal set of outlinks for a particular node

We first consider the case where $\mathcal{I}$ is a singleton, say $\mathcal{I} = \{1\}$. We want to maximize its PageRank score $\pi_{\mathbf{1}}$. The only control we have on the webgraph is the choice of the outlinks of page 1, that is the choice of $\mathcal{L} \subseteq \bar{\mathcal{I}} = \{2, \ldots, n\}$, $\mathcal{L} \neq \emptyset$, for which

$$A_{\mathrm{out}(\{1\})} = \mathbf{e}_{\mathcal{L}}^{\mathbf{T}},$$

where $(\mathbf{e}_{\mathcal{L}})_{\mathbf{i}} = \mathbf{1}$ if $i \in \mathcal{L}$, 0 otherwise.

The expression of $\pi_{\mathbf{1}}$ follows directly from Proposition 6.

Corollary 7 *The PageRank score $\pi_{\mathbf{1}}(\mathcal{L})$ of the node 1 is given by*

$$\pi_{\mathbf{1}}(\mathcal{L}) = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{c}\,\frac{\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}}{|\mathcal{L}|}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1} + (\mathbf{1} - \mathbf{c})\,\mathbf{1}\mathbf{z}_{\bar{\mathcal{I}}}^{\mathbf{T}}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}}.$$

Proposition 8 *The PageRank $\pi_{\mathbf{1}}(\mathcal{L})$ is maximal if and only if $\mathcal{L}$ is a non-empty subset of $\mathcal{L}^*$, where*

$$\mathcal{L}^* = \arg\min_i \mathbf{e}_{\mathbf{i}}^{\mathbf{T}}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}. \tag{3}$$

Proof. Proof. Clearly, by Corollary 7, $\pi_{\mathbf{1}}(\mathcal{L})$ is maximal if and only if $\frac{\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}}{|\mathcal{L}|}(I - G_{\bar{\mathcal{I}}})^{-1}\mathbf{1}$ is minimal.

The following proposition claims that the best one node can do is to point to some of its own parents. If it has no parent, its PageRank does not depend on its outlinks.

Proposition 9 *Suppose that $A_{\mathrm{in}(\{1\})} \neq 0$. Then,*

$$\mathcal{L}^* \subseteq \mathrm{support}(A_{\mathrm{in}(\{1\})}).$$

Proof. Proof. Let $\mathbf{v} = (\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}$. From $G_{\mathrm{in}(\{1\})} + G_{\bar{\mathcal{I}}}\mathbf{1} = \mathbf{1}$, we obtain $(1-c)\mathbf{z_1}\mathbf{1} = (\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})\mathbf{1} - \mathbf{c}\mathbf{A}_{\mathrm{in}(\{1\})}$. Therefore, since $A_{\mathrm{in}(\{1\})} \neq 0$, $(1-c)\mathbf{z_1}\mathbf{v} = \mathbf{1} - \mathbf{c}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{A}_{\mathrm{in}(\{1\})} \lneqq \mathbf{1}$, that is

$$(1-c)\mathbf{z_1} \min_{\mathbf{j}} \mathbf{v_j} < \mathbf{1}. \tag{4}$$

Let $i \notin \mathrm{support}(A_{\mathrm{in}(\{1\})})$. Then $\mathbf{e_i^T}\mathbf{G}_{\bar{\mathcal{I}}}\mathbf{1} = \mathbf{1} - (\mathbf{1} - \mathbf{c})\mathbf{z_1}$. So

$$\begin{aligned}
\mathbf{v_i} &= \mathbf{e_i^T}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1} \\
&= \mathbf{e_i^T}(\mathbf{I} + \mathbf{G}_{\bar{\mathcal{I}}}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}})\mathbf{1} = \mathbf{1} + \mathbf{e_i^T}\mathbf{G}_{\bar{\mathcal{I}}}\mathbf{v} \\
&\geq 1 + \mathbf{e_i^T}\mathbf{G}_{\bar{\mathcal{I}}}\mathbf{1} \min_{\mathbf{j}} \mathbf{v_j} \\
&\geq 1 + (1 - (1-c)\mathbf{z_1}) \min_{\mathbf{j}} \mathbf{v_j} > \min_{\mathbf{j}} \mathbf{v_j},
\end{aligned}$$

by equation 4. It follows that if $i \notin \mathrm{support}(A_{\mathrm{in}(\{1\})})$, then $i \notin \mathcal{L}^*$.

Proposition 9 shows that, in order to maximize the PageRank of a node by choosing its outlinks, this node must link to nodes which link to itself, i.e. some of its parents. But, as shown in Figure 1, all its parents are not necessarily optimal, and moreover, it can be better for this node to link to some nodes which are not its parents than to link to some of its particular parents.

In [1], the authors shows that the optimal linkage strategy for one node i is

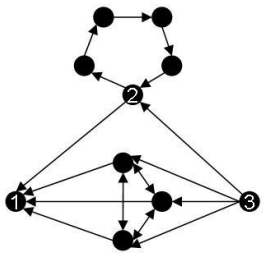

Fig. 1. Node 1 has advantage to link to node 3 (grand-parent) rather than node 2 (parent) where the random surfer is more likely to waste time in the loop before going back to node 1.

reached when it points to one particular node. We add that this particular node belongs to parents of node i. However, all parents are not interesting (Figure 1).

2.2 Optimal set of outlinks for a clique

We now consider the case where $\mathcal{I}$ is a set of $n_{\mathcal{I}} \geq 2$ nodes, organized in a clique. We want maximize the sum of their PageRanks, that is $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$.

We make the conjecture that $A_{\text{out}(\mathcal{I})}$ must be of rank one in order to maximize the sum of the PageRank of $\mathcal{I}$ for any $A_{\mathcal{I}}$. Further, only one node $\in \mathcal{I}$ must have no more than one external child.

Definition 10 *An external child/parent/node is a child/parent/node in $\bar{\mathcal{I}}$.*

That conjecture is intuitive: you should chose a node of the clique which has a low PageRank to have a leak in your clique, so you do not lose to much score. And moreover, it is better for you to point to an external node which will soon give you PageRank back.

Proposition 14 proves our conjecture for $A_{\mathcal{I}}$ being a clique with at least one node without external parent (section 2.2.1) and a minimal zapping in the clique. Proposition 16 assumes that $A_{\text{out}(\mathcal{I})}$ is of rank one (section 2.2.2) and proves then that $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$ is optimal with $A_{\text{out}(\mathcal{I})}$ having only one nonzero entry. Therefore Corollary 17 claims that when a clique has some node without external parent and a minimal zapping in the clique, then that node must point to one particular external parent of the clique in order to maximize $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$.

2.2.1 Some node of the clique does not have any external parent

We will first analyze the case where at least one of the nodes of the clique, say 1, has no external parent.

We introduce the following lemma which is similar to Corollary 7 and Proposition 8.

Lemma 11 *The sum of PageRanks $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$ is maximal if and only if $\frac{\pi_{\mathcal{I}}^{\mathbf{T}}}{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}} G_{\text{out}(\mathcal{I})}(I - G_{\bar{\mathcal{I}}})^{-1}\mathbf{1}$ is minimal.*

Proof. Proof. This follows simply from Propositions 6 and 4:

$$\frac{1}{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}} = \frac{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{t}_{\mathcal{I}}}{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}} = 1 + \frac{\pi_{\mathcal{I}}^{\mathbf{T}}}{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}} G_{\text{out}(\mathcal{I})}(I - G_{\bar{\mathcal{I}}})^{-1}\mathbf{1}.$$

The idea of the following lemma is just to express the relation between the PageRank scores of the nodes of the clique which have external children (say such nodes $\in \mathcal{R}$), the PageRanks sum of the clique ($\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$), the personalization vector of nodes in $\mathcal{R}$ ($\mathbf{z}_{\mathcal{R}}$) and the flow given by the external parents to nodes in $\mathcal{R}$.

Lemma 12 *Let $A_{\text{out}(\mathcal{I})} = \sum_{i\in\mathcal{R}} \mathbf{e_i}\mathbf{e}_{\mathcal{L}_\mathbf{i}}$, for some $i \in \mathcal{R} \subseteq \mathcal{I}$, $\mathcal{L}_i \subseteq \bar{\mathcal{I}}$. Let $r = |\mathcal{R}|$ and $\ell_i = |\mathcal{L}_i|$ for all $i \in \mathcal{R}$. Then*

$$\sum_{i\in\mathcal{R}} \pi_{\mathbf{i}} \left(\mathbf{1} + \frac{\mathbf{c}}{\mathbf{n}_{\mathcal{I}} - \mathbf{1}} \frac{\mathbf{n}_{\mathcal{I}} - \mathbf{1} + \ell_\mathbf{i}\mathbf{r}}{\mathbf{n}_{\mathcal{I}} - \mathbf{1} + \ell_\mathbf{i}} \right) = \frac{\mathbf{cr}}{\mathbf{n}_{\mathcal{I}} - \mathbf{1}} \pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1} + (\mathbf{1} - \mathbf{c})\mathbf{z}_{\mathcal{R}}^{\mathbf{T}}\mathbf{1} + \mathbf{c}\pi_{\bar{\mathcal{I}}}^{\mathbf{T}}\mathbf{D}_{\bar{\mathcal{I}}}^{-\mathbf{1}}\mathbf{A}_{\text{in}(\mathcal{I})}\mathbf{e}_{\mathcal{R}}.$$

Proof. Proof. For all $i \in \mathcal{R}$, we have from PageRank equation 1,

$$\begin{aligned}\pi_{\mathbf{i}} &= \pi^{\mathbf{T}}(\mathbf{cD^{-1}A} + \mathbf{(1-c)1z^{T}})\mathbf{e_i} \\ &= c\left(\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{D_{\mathcal{I}}^{-1}A}_{\mathcal{I}}\mathbf{e_i} + \pi_{\bar{\mathcal{I}}}^{\mathbf{T}}\mathbf{D_{\bar{\mathcal{I}}}^{-1}A}_{\mathrm{in}(\mathcal{I})}\mathbf{e_i}\right) + (1-c)\mathbf{z_i} \\ &= c\Big(\sum_{j\in\mathcal{I}\setminus\mathcal{R}} \frac{\pi_{\mathbf{j}}}{n_{\mathcal{I}}-1} + \sum_{j\in\mathcal{R}, j\neq i} \frac{\pi_{\mathbf{j}}}{n_{\mathcal{I}}-1+\ell_j} + \pi_{\bar{\mathcal{I}}}^{\mathbf{T}}\mathbf{D_{\bar{\mathcal{I}}}^{-1}A}_{\mathrm{in}(\mathcal{I})}\mathbf{e_i}\Big) + \mathbf{(1-c)z_i}.\end{aligned}$$

Therefore,

$$\pi_{\mathcal{R}}^{\mathbf{T}}\mathbf{1} = \mathbf{cr}\frac{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1} - \pi_{\mathcal{R}}^{\mathbf{T}}\mathbf{1}}{\mathbf{n_{\mathcal{I}} - 1}} + \mathbf{c(r-1)}\sum_{\mathbf{i}\in\mathcal{R}} \frac{\pi_{\mathbf{i}}}{\mathbf{n_{\mathcal{I}} - 1 + \ell_i}} + \mathbf{c}\pi_{\bar{\mathcal{I}}}^{\mathbf{T}}\mathbf{D_{\bar{\mathcal{I}}}^{-1}A}_{\mathrm{in}(\mathcal{I})}\mathbf{e}_{\mathcal{R}} + \mathbf{(1-c)}\mathbf{z}_{\mathcal{R}}^{\mathbf{T}}\mathbf{1},$$

and the conclusion follows by rearranging terms.

We now get directly a lower bound on the ratio between the PageRanks sum for the nodes of the clique which have outlinks to the rest of the graph, and the PageRanks sum for all the nodes of the clique.

Lemma 13 *Let $A_{\mathrm{out}(\mathcal{I})} = \sum_{i\in\mathcal{R}} \mathbf{e_i}\mathbf{e}_{\mathcal{L}_i}$, for some $i \in \mathcal{R}$, $\mathcal{L}_i$. Let $r = |\mathcal{R}|$ and let $\delta = \max_{i\in\mathcal{R}} \frac{n_{\mathcal{I}}-1+|\mathcal{L}_i|\,r}{n_{\mathcal{I}}-1+|\mathcal{L}_i|}$. Then*

$$\frac{\pi_{\mathcal{R}}^{\mathbf{T}}\mathbf{1}}{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}} \geq \frac{cr}{n_{\mathcal{I}} - 1 + c\delta} + \frac{(1-c)(n_{\mathcal{I}}-1)}{n_{\mathcal{I}} - 1 + c\delta}\frac{\mathbf{z}_{\mathcal{R}}^{\mathbf{T}}\mathbf{1}}{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}}$$

with equality if and only if $r = 1$ and $A_{\mathrm{in}(\mathcal{I})}\mathbf{e}_{\mathcal{R}} = \mathbf{0}$.

Proof. to do

The following proposition shows that, if there exists some node in the clique which does not have any external parent, and if this node also has a minimal probabilty of zapping in the clique, then the outlink matrix $A_{\mathrm{out}(\mathcal{I})}$ must be of rank one in order to maximize the sum of the PageRank scores of the clique.

Let us note that decreasing the number of external children to make $A_{\mathrm{out}(\mathcal{I})}$ of rank one does not necessarily improve the sum of the PageRanks of the clique as shown in Figure 2.

Proposition 14 *Suppose that $\mathcal{F} = \{f \in \mathcal{I} : A_{\mathrm{in}(\mathcal{I})}\mathbf{e_f} = \mathbf{0}$ and $\mathbf{z_f} = \min_{\mathbf{i}\in\mathcal{I}} \mathbf{z_i}\} \neq \emptyset$. Then $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$ can be maximal only if $A_{\mathrm{out}(\mathcal{I})}$ is of rank one.*

Proof. We will compare the PageRank sum for two cases. For the first case, we suppose that rank $A_{\mathrm{out}(\mathcal{I})} \geq 2$. Then, without loss of generality, $A_{\mathrm{out}(\mathcal{I})} =$

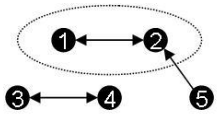

Fig. 2. Two nodes pointing outside may be better than one node pointing outside. For example, with $c = .85$ we have

$$A_{\mathrm{out}(\mathcal{I})} = \begin{pmatrix} 1\ 0\ 0 \\ 0\ 0\ 0 \end{pmatrix} <_{PR} A_{\mathrm{out}(\mathcal{I})} = \begin{pmatrix} 1\ 0\ 0 \\ 0\ 0\ 1 \end{pmatrix} <_{PR} A_{\mathrm{out}(\mathcal{I})} = \begin{pmatrix} 0\ 0\ 1 \\ 0\ 0\ 0 \end{pmatrix} \text{ (best).}$$

$\sum_{i\in\mathcal{R}} \mathbf{e_i}\mathbf{e}_{\mathcal{L}_\mathbf{i}}$, for some $i \in \mathcal{R}$, $\mathcal{L}_i$. Let $r = |\mathcal{R}| \geq 2$ and let

$$\delta = \max_{i\in\mathcal{R}} \frac{n_\mathcal{I} - 1 + |\mathcal{L}_i|\, r}{n_\mathcal{I} - 1 + |\mathcal{L}_i|} < r.$$

For the second case, we consider the particular case where $\tilde{A}_{\mathrm{out}(\mathcal{I})} = \mathbf{e_f}\mathbf{e}_\ell^\mathbf{T}$, for some node $f \in \mathcal{I}$ such that $A_{\mathrm{in}(\mathcal{I})}\mathbf{e_f} = \mathbf{0}$, and $\ell \in \mathcal{L}^*$ (definition of $\mathcal{L}^*$ in Proposition 8). Let π and $\tilde{\pi}$ be respectively the PageRanks for $A_{\mathrm{out}(\mathcal{I})}$ and $\tilde{A}_{\mathrm{out}(\mathcal{I})}$. Let

$$\alpha = \frac{\mathbf{z}_\mathcal{R}^\mathbf{T}\mathbf{1}}{\delta \mathbf{z_f}}.$$

Since $\delta < r$ and $\mathbf{z_f} = \min_{\mathbf{i}\in\mathcal{I}} \mathbf{z_i}$, we have $\alpha > 1$, and hence there exists some $p \in \mathbb{N}$ such that $\alpha^p \tilde{\pi}_\mathcal{I}^T \mathbf{1} \geq \pi_\mathcal{I}^\mathbf{T}\mathbf{1}$. We want to prove that, for all $k \in \mathbb{N}$,

$$\alpha^{k+1}\tilde{\pi}_\mathcal{I}^T\mathbf{1} \geq \pi_\mathcal{I}^\mathbf{T}\mathbf{1} \quad \text{implies that} \quad \alpha^\mathbf{k}\tilde{\pi}_\mathcal{I}^\mathbf{T}\mathbf{1} > \pi_\mathcal{I}^\mathbf{T}\mathbf{1},$$

and finally $\tilde{\pi}_\mathcal{I}^T\mathbf{1} > \pi_\mathcal{I}^\mathbf{T}\mathbf{1}$.

Let $k \in \mathbb{N}$, and suppose that $\alpha^{k+1}\tilde{\pi}_\mathcal{I}^T\mathbf{1} \geq \pi_\mathcal{I}^\mathbf{T}\mathbf{1}$. Then, by Lemma 13,

$$\begin{aligned}
\alpha^k \frac{\pi_\mathcal{R}^\mathbf{T}\mathbf{1}}{\pi_\mathcal{I}^\mathbf{T}\mathbf{1}} &> \alpha^k \frac{cr}{n_\mathcal{I} - 1 + c\delta} + \frac{(1-c)(n_\mathcal{I}-1)}{\frac{n_\mathcal{I}-1}{\delta} + c} \frac{\alpha^k}{\delta} \frac{\mathbf{z}_\mathcal{R}^\mathbf{T}\mathbf{1}}{\pi_\mathcal{I}^\mathbf{T}\mathbf{1}} \\
&> \frac{c}{n_\mathcal{I} - 1 + c} + \frac{(1-c)(n_\mathcal{I}-1)}{n_\mathcal{I} - 1 + c} \alpha^{k+1} \frac{\mathbf{z_f}}{\pi_\mathcal{I}^\mathbf{T}\mathbf{1}} \\
&\geq \frac{c}{n_\mathcal{I} - 1 + c} + \frac{(1-c)(n_\mathcal{I}-1)}{n_\mathcal{I} - 1 + c} \frac{\mathbf{z_f}}{\tilde{\pi}_\mathcal{I}^T\mathbf{1}} \quad = \frac{\tilde{\pi}_f}{\tilde{\pi}_\mathcal{I}^T\mathbf{1}}.
\end{aligned}$$

So, since $\ell \in \mathcal{L}^*$,

$$\begin{aligned}
&\alpha^k \frac{\pi_\mathcal{I}^\mathbf{T}}{\pi_\mathcal{I}^\mathbf{T}\mathbf{1}} G_{\mathrm{out}(\mathcal{I})}(I - G_{\bar{\mathcal{I}}})^{-1}\mathbf{1} \\
&\qquad = \alpha^k \Big(c \sum_{i\in\mathcal{R}} \frac{\pi_\mathbf{i}}{\pi_\mathcal{I}^\mathbf{T}\mathbf{1}} \frac{\mathbf{e}_{\mathcal{L}_\mathbf{i}}^\mathbf{T}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}}{n_\mathcal{I} - 1 + |\mathcal{L}_i|} + (1-c)\mathbf{z}_{\bar{\mathcal{I}}}^\mathbf{T}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1} \Big)
\end{aligned}$$

$$> \Big(c\frac{\tilde{\pi}_f}{\tilde{\pi}_{\mathcal{I}}^T\mathbf{1}}\frac{\mathbf{e}_\ell^\mathbf{T}(\mathbf{I}-\mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}}{n_{\mathcal{I}}} + (1-c)\mathbf{z}_{\bar{\mathcal{I}}}^\mathbf{T}(\mathbf{I}-\mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}\Big)$$
$$= \frac{\tilde{\pi}_{\mathcal{I}}^T}{\tilde{\pi}_{\mathcal{I}}^T\mathbf{1}}\tilde{G}_{\mathrm{out}(\mathcal{I})}(I-G_{\bar{\mathcal{I}}})^{-1}\mathbf{1}.$$

Now, by Lemma 11,

$$\frac{\alpha^k}{\pi_{\mathcal{I}}^\mathbf{T}\mathbf{1}} = \alpha^k\Big(1+\frac{\pi_{\mathcal{I}}^\mathbf{T}}{\pi_{\mathcal{I}}^\mathbf{T}\mathbf{1}}G_{\mathrm{out}(\mathcal{I})}(I-G_{\bar{\mathcal{I}}})^{-1}\mathbf{1}\Big)$$
$$> 1+\frac{\tilde{\pi}_{\mathcal{I}}^T}{\tilde{\pi}_{\mathcal{I}}^T\mathbf{1}}\tilde{G}_{\mathrm{out}(\mathcal{I})}(I-G_{\bar{\mathcal{I}}})^{-1}\mathbf{1} \quad = \frac{\mathbf{1}}{\tilde{\pi}_{\mathcal{I}}^\mathbf{T}\mathbf{1}},$$

and therefore $\alpha^k\tilde{\pi}_{\mathcal{I}}^T\mathbf{1} > \pi_{\mathcal{I}}^\mathbf{T}\mathbf{1}$. By induction on decreasing k, we conclude that $\tilde{\pi}_{\mathcal{I}}^T\mathbf{1} > \pi_{\mathcal{I}}^\mathbf{T}\mathbf{1}$, that is the second case where $\tilde{A}_{\mathrm{out}(\mathcal{I})} = \mathbf{e_f}\mathbf{e}_\ell^\mathbf{T}$, for some node $f \in \mathcal{I}$ such that $A_{\mathrm{in}(\mathcal{I})}\mathbf{e_f} = \mathbf{0}$, and $\ell \in \mathcal{L}^*$ is better than any configuration where $A_{\mathrm{out}(\mathcal{I})}$ has a rank strictly larger than one.

2.2.2 The outlink structure is of rank one

With Proposition 14, we have seen that under some assumptions on the structure of the graph and on the preference vector, the optimal outlink structure is of rank one. Now, we will not impose conditions on the inlinks of the clique, nor on the preference vector, but we will suppose that the adjacency matrix block $A_{\mathrm{out}(\mathcal{I})}$ corresponding to the externam outlinks to the rest of the graph is of rank one, that is

$$A_{\mathrm{out}(\mathcal{I})} = \mathbf{e}_\mathcal{M}\mathbf{e}_\mathcal{L}^\mathbf{T},$$

for some nonempty $\mathcal{M} \subseteq \mathcal{I}$ and $\mathcal{L} \subseteq \bar{\mathcal{I}}$.

Lemma 15 *Let $a, b \geq 0$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\mathbf{n}$ such that $\mathbf{v} \geq \mathbf{0}$ and $a + \mathbf{u}^\mathbf{T}\mathbf{e_i} \geq \mathbf{0}$ for all $i \in \{1, \ldots, n\}$. Let, for all $\emptyset \neq \mathcal{R} \subseteq \{1, \ldots, n\}$,*

$$f(\mathcal{R}) = \frac{a + \mathbf{u}^\mathbf{T}\mathbf{e}_\mathcal{R}}{(a + \mathbf{u}^\mathbf{T}\mathbf{e}_\mathcal{R})\mathbf{b} + \mathbf{v}^\mathbf{T}\mathbf{e}_\mathcal{R}}.$$

Then, the maximum of f is obtained for some $\mathcal{R}$ with $|\mathcal{R}^*| = 1$. Moreover, if $a > 0$ and $\mathbf{v} > \mathbf{0}$, then $f(\mathcal{R}^*) > f(\mathcal{R})$ for all $\mathcal{R}$ such that $|\mathcal{R}| \geq 2$.

Proof. Let $i \in \{1, \ldots, n\}$ such that $f(\{i\}) \geq f(\{j\})$ for all $j \in \{1, \ldots, n\}$. Let $\emptyset \neq \mathcal{R} \subseteq \{1, \ldots, n\}$. Then

$$f(\{i\}) = \frac{a + \mathbf{u}^\mathbf{T}\mathbf{e_i}}{(a + \mathbf{u}^\mathbf{T}\mathbf{e_i})\mathbf{b} + \mathbf{v}^\mathbf{T}\mathbf{e_i}} \geq \frac{\sum_{j\in\mathcal{R}}(a + \mathbf{u}^\mathbf{T}\mathbf{e_j})}{\sum_{j\in\mathcal{R}}((a + \mathbf{u}^\mathbf{T}\mathbf{e_j})\mathbf{b} + \mathbf{v}^\mathbf{T}\mathbf{e_j})}$$

$$\geq \frac{a + \sum_{j\in\mathcal{R}} \mathbf{u^T e_j}}{(a + \sum_{j\in\mathcal{R}} \mathbf{u^T e_j})\mathbf{b} + \sum_{\mathbf{j}\in\mathcal{R}} \mathbf{v^T e_j}} = f(\mathcal{R}).$$

Moreover, if $|\mathcal{R}| \geq 2$, $a > 0$ and $\mathbf{v} > \mathbf{0}$, the last inequality is strict.

The following proposition assumes that $A_{\mathrm{out}(\mathcal{I})}$ is of rank one and proves then that $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$ is optimal with $A_{\mathrm{out}(\mathcal{I})}$ having only one nonzero entry. In other words, it is enough to have one node in the clique pointing to one external child.

Proposition 16 *Under the constraint that $A_{\mathrm{out}(\mathcal{I})} = \mathbf{e}_{\mathcal{M}}\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}$ for some non-empty $\mathcal{M} \subseteq \mathcal{I}$ and $\mathcal{L} \subseteq \bar{\mathcal{I}}$, the sum of PageRanks $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$ can be maximal only if $|\mathcal{L}| = |\mathcal{M}| = 1$.*

Proof. With $A_{\mathrm{out}(\mathcal{I})} = \mathbf{e}_{\mathcal{M}}\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}$, we have

$$\begin{aligned} G_{\mathrm{out}(\mathcal{I})} &= c\frac{\mathbf{e}_{\mathcal{M}}\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}}{n_{\mathcal{I}} - 1 + |\mathcal{L}|} + (1-c)\mathbf{1}\mathbf{z}_{\bar{\mathcal{I}}}^{\mathbf{T}}, \\ G_{\mathcal{I}}\mathbf{e}_{\mathcal{M}} &= \left(\frac{c}{n_{\mathcal{I}} - 1}|\mathcal{M}| + (1-c)\mathbf{z}_{\mathcal{I}}^{\mathbf{T}}\mathbf{e}_{\mathcal{M}}\right)\mathbf{1} - \frac{\mathbf{c}}{\mathbf{n_{\mathcal{I}} - 1}}\frac{\mathbf{n_{\mathcal{I}} - 1} + |\mathcal{L}||\mathcal{M}|}{\mathbf{n_{\mathcal{I}} - 1} + |\mathcal{L}|}\mathbf{e}_{\mathcal{M}}. \end{aligned}$$

From Propositions 5 and 3,

$$\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{e}_{\mathcal{M}} = \pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{G}_{\mathcal{I}}\mathbf{e}_{\mathcal{M}} + \pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{G}_{\mathrm{out}(\mathcal{I})}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{G}_{\mathrm{in}(\mathcal{I})}\mathbf{e}_{\mathcal{M}}.$$

Substituting $G_{\mathrm{out}(\mathcal{I})}$ and $G_{\mathcal{I}}\mathbf{e}_{\mathcal{M}}$, we get

$$\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}\,(\mathbf{w^T e}_{\mathcal{M}}) = \frac{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{e}_{\mathcal{M}}}{\mathbf{n_{\mathcal{I}} - 1} + |\mathcal{L}|}\left(\alpha + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{1} + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{M}\mathbf{e}_{\mathcal{M}}\right), \tag{5}$$

where

$$\begin{aligned} \alpha &= n_{\mathcal{I}} - 1 + c, \\ \mathbf{w^T} &= \frac{c\mathbf{1^T}}{n_{\mathcal{I}} - 1} + (1-c)\mathbf{z}_{\mathcal{I}}^{\mathbf{T}} + (\mathbf{1} - \mathbf{c})\mathbf{z}_{\bar{\mathcal{I}}}^{\mathbf{T}}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{G}_{\mathrm{in}(\mathcal{I})}, \\ M &= \frac{c\mathbf{1}\mathbf{1^T}}{n_{\mathcal{I}} - 1} - c(I - G_{\bar{\mathcal{I}}})^{-1}G_{\mathrm{in}(\mathcal{I})}. \end{aligned}$$

Moreover, from Propositions 6 and 4,

$$1 = \pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1} + \pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{G}_{\mathrm{out}(\mathcal{I})}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1}.$$

Substituting $G_{\mathrm{out}(\mathcal{I})}$, we get

$$\beta\,\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1} + \frac{\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{e}_{\mathcal{M}}}{\mathbf{n_{\mathcal{I}} - 1} + |\mathcal{L}|}\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{q} = \mathbf{1}, \tag{6}$$

where

$$\begin{aligned}\beta &= 1 + (1-c)\mathbf{z}_{\bar{\mathcal{I}}}^{\mathbf{T}}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{1},\\ \mathbf{q} &= c(I - G_{\bar{\mathcal{I}}})^{-1}\mathbf{1}.\end{aligned}$$

Combining equations 5 and 6, we finally have

$$\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1} = \frac{\alpha + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{1} + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{M}\mathbf{e}_{\mathcal{M}}}{(\alpha + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{1} + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{M}\mathbf{e}_{\mathcal{M}})\beta + \mathbf{w}^{\mathbf{T}}\mathbf{e}_{\mathcal{M}}\,\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{q}}.$$

Suppose first that $\mathcal{L}$ is fixed. We can apply Lemma 15 with

$$\begin{aligned}a &= \alpha + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{1} > \mathbf{0},\\ b &= \beta \geq 0,\\ \mathbf{v}^{\mathbf{T}} &= \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{q}\mathbf{w}^{\mathbf{T}} > \mathbf{0},\\ a + \mathbf{u}^{\mathbf{T}}\mathbf{e}_{\mathbf{i}} &= \alpha + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{1} + \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}\mathbf{M}\mathbf{e}_{\mathbf{i}} > \mathbf{e}_{\mathcal{L}}^{\mathbf{T}}(\mathbf{1} - \mathbf{c}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{G}_{\mathrm{in}(\mathcal{I})}\mathbf{e}_{\mathbf{i}}) \geq \mathbf{0} \text{ for all } \mathbf{i},\end{aligned}$$

since $(I - G_{\bar{\mathcal{I}}})^{-1}G_{\mathrm{in}(\mathcal{I})}\mathbf{e}_{\mathbf{i}} \leq (\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{G}_{\mathrm{in}(\mathcal{I})}\mathbf{1} = \mathbf{1}$. It follows that, whatever $\mathcal{L}$, the set $\mathcal{M}$ must be a singleton in order to maximize $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$. Suppose that $\mathcal{M}$ is fixed. By the same argument with

$$\begin{aligned}a &= \alpha > 0,\\ b &= \beta \geq 0,\\ \mathbf{v} &= \mathbf{q}\mathbf{w}^{\mathbf{T}}\mathbf{e}_{\mathcal{M}} > \mathbf{0},\\ a + \mathbf{u}^{\mathbf{T}}\mathbf{e}_{\mathbf{i}} &= \alpha + \mathbf{e}_{\mathbf{i}}^{\mathbf{T}}(\mathbf{1} + \mathbf{M}\mathbf{e}_{\mathcal{M}}) > \mathbf{1} - \mathbf{c}\mathbf{e}_{\mathbf{i}}^{\mathbf{T}}(\mathbf{I} - \mathbf{G}_{\bar{\mathcal{I}}})^{-\mathbf{1}}\mathbf{G}_{\mathrm{in}(\mathcal{I})}\mathbf{e}_{\mathcal{M}} \geq \mathbf{0} \text{ for all } \mathbf{i},\end{aligned}$$

it follows that $\mathcal{L}$ must be a singleton to maximize $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$, whatever $\mathcal{M}$. We can therefore conclude that, if $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$ is maximal under the constraint that $A_{\mathrm{out}(\mathcal{I})} = \mathbf{e}_{\mathcal{M}}\mathbf{e}_{\mathcal{L}}^{\mathbf{T}}$, with $\mathcal{M}, \mathcal{L} \neq \emptyset$, then we must have $|\mathcal{M}| = |\mathcal{L}| = 1$.

The following corollary is deducted from proposition 14 and 16, it uses the notation $\mathcal{L}^*$ defined in proposition 8.

Corollary 17 *Suppose that* $\mathcal{F} = \{f \in \mathcal{I} \colon A_{\mathrm{in}(\mathcal{I})}\mathbf{e}_{\mathbf{f}} = \mathbf{0}$ *and* $\mathbf{z}_{\mathbf{f}} = \min_{\mathbf{i}\in\mathcal{I}} \mathbf{z}_{\mathbf{i}}\} \neq \emptyset$*. Then* $\pi_{\mathcal{I}}^{\mathbf{T}}\mathbf{1}$ *can be maximal if and only if* $A_{\mathrm{out}(\mathcal{I})} = \mathbf{e}_{\mathbf{i}}\mathbf{e}_{\mathbf{j}}^{\mathbf{T}}$ *where* $i \in \mathcal{F}$ *and* $j \in \mathcal{L}^*$*.*

The example in Figure 3 shows that the condition $\mathbf{z}_{\mathbf{f}} = \min_{\mathbf{i}\in\mathcal{I}} \mathbf{z}_{\mathbf{i}}$ is necessary, indeed the vector $\mathbf{z}$ may modify the *leaking* node (the node in $\mathcal{I}$ pointing to an external node). And the example in Figure 4 exposes one case where all nodes in the clique have external parents. There the optimal linkage strategy do not lead to point to an external node giving the best mean return time to the clique, that is an external node in $\mathcal{L}^*$.

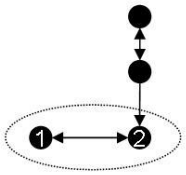

Fig. 3. Let $\mathbf{z_1} = .\mathbf{7}$ and $z_i = .1$ for $i = 2, 3, 4$. The optimal linkage for the clique $\mathcal{I} = \{1, 2\}$ is $A_{\mathrm{out}(\mathcal{I})} = \mathbf{e_i e_j^T}$ with $i = 1$ (the leak is in the node with no external parent) when $(1 - c) = .15$, but with $i = 2$ when the zapping increases $(1 - c) = .3$.

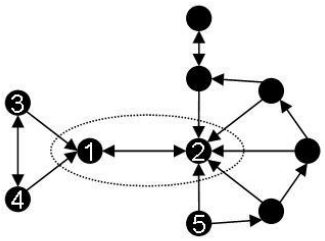

Fig. 4. When all nodes in the clique $\mathcal{I} = \{1, 2\}$ have at least one external parent, the optimal for the sum of the PageRank of the clique may be $A_{\mathrm{out}(\mathcal{I})} = \mathbf{e_i e_j^T}$ with $j \notin \mathcal{L}^*$. Here $\mathcal{L}^* = \{3, 4\}$, however the optimal linkage strategy is $i = 1$ and $j = 5 \notin \mathcal{L}^*$.

3 Conclusion

Let us remind that we focus on the matrix $A_{\mathrm{out}(\mathcal{I})}$ being given $A_{\mathcal{I}}$, $A_{\mathrm{in}(\mathcal{I})}$ and $A_{\bar{\mathcal{I}}}$. Clearly, other questions arise if we consider other combinations. For example, we could fix the matrices $A_{\mathrm{out}(\mathcal{I})}$, $A_{\mathrm{in}(\mathcal{I})}$, $A_{\bar{\mathcal{I}}}$ and look for the optimal internal structure of $\mathcal{I}$, that is determined $A_{\mathcal{I}}$ to maximize $\pi_{\mathcal{I}}^T \mathbf{1}$. The answer is not trivial and the optimal structure is not a clique. We have seen that finding $A_{\mathrm{out}(\mathcal{I})}$ when $A_{\mathcal{I}}$ is a clique is not direct. Surely that the problem becomes even more complex for other $A_{\mathcal{I}}$. However, we make the following conjecture:

For any fixed $A_{\mathcal{I}}$, the sum of PageRanks of $\mathcal{I}$ is maximal only if $A_{\mathrm{out}(\mathcal{I})} = \mathbf{e_i e_j^T}$ for some $i, j \in \mathcal{N}$.

Let us remark that this is no more true when we want to maximize a linear combination of the PageRanks of $\mathcal{I}$. That is maximizing the more general expression $\pi_{\mathcal{I}}^T \mathbf{v}$ instead of $\pi_{\mathcal{I}}^T \mathbf{1}$, with $\mathbf{v}$ any vector of $n_{\mathcal{I}}$ entries. For instance, Figure 5 shows with $\mathbf{v} = (\mathbf{1000})$ that the optimal $A_{\mathrm{out}(\mathcal{I})}$ has not one nonzero entry and is even not of rank one.

When $\mathcal{I} = \{1\}$, the optimal linkage strategy does not depend on the damping factor c, nor the personalization vector $\mathbf{z}$ contrarily to the case of several pages in $\mathcal{I}$. The best consists to point to one of its own parents if it exists. However, the result is different if node 1 wants to keep outlinks to a set of pages and add outlinks to maximize its PageRank. Then node 1 needs to add iteratively

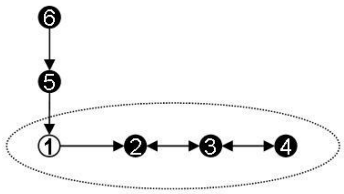

Fig. 5. The set $\mathcal{I} = \{1, 2, 3, 4\}$ wants to maximize $\pi_{\mathcal{I}}^T \mathbf{v}$ where $\mathbf{v} = (\mathbf{1\ 0\ 0\ 0})$ and the damping factor is $c = .85$. Then $A_{\mathrm{out}(\mathcal{I})}$ won't be of rank one. The optimal is achieved for nodes $1, 2, 4$ pointing to node 5 and node 3 has to point to nodes 5 and 6.

an outlink to the page having the minimum mean return time to itself. This process stops once all remaining pages have a mean return time to node 1 higher than the average of the mean return times to node 1 of the pages pointed by node 1. That problem of having a fixed set of outlinks for $\mathcal{I}$ and maximizing the sum of PageRanks of the clique $\mathcal{I}$ is similar to the one node case required that we have the assumption of Proposition 14.

Acknowledgements

This paper presents research supported by the Concerted Research Action (ARC) "Large Graphs and Networks" of the French Community of Belgium, and by the Belgian Programme on Inter-university Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with the authors.

References

[1] Konstantin Avrachenkov and Nelly Litvak, *The effect of new links on Google pagerank*, Stoch. Models **22** (2006), no. 2, 319–331.

[2] C. Castillo et al. Baeza-Yates, R., *Pagerank increase under different collusion topologies.*, Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[3] Monica Bianchini, Marco Gori, and Franco Scarselli, *Pagerank: A circuital analysis*, Proceedings of WWW2002 (Honolulu, Hawaii (USA)), May 2002.

[4] ———, *Inside PageRank*, ACM Trans. Inter. Tech. **5** (2005), no. 1, 92–128.

[5] Zoltan Gyongyi and Hector Garcia-Molina, *Link spam alliances*, VLDB '05: Proceedings of the 31st international conference on Very large data bases, VLDB Endowment, 2005, pp. 517–528.

[6] I.C.F. Ipse and R.S. Wills, *Mathematical properties and analysis of Google's PageRank*, Bol. Soc. Esp. Mat. Apl. **34** (2006), 191–196.

[7] Amy N. Langville and Carl D. Meyer, *Deeper inside PageRank*, Internet Math. **1** (2004), no. 3, 335–380.

[8] Carl D. Meyer, *Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems*, SIAM Rev. **31** (1989), no. 2, 240–272.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, *The pagerank citation ranking: Bringing order to the web*, Tech. report, Stanford Digital Library Technologies Project, 1998.

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

[11] M. Sydow, *Can one out-link change your pagerank?*, Lecture Notes in Computer Science **3528** (2005), 408–414.