# Inside PageRank

MONICA BIANCHINI, MARCO GORI, and FRANCO SCARSELLI
University of Siena

Although the interest of a Web page is strictly related to its content and to the subjective readers' cultural background, a measure of the page authority can be provided that only depends on the topological structure of the Web. PageRank is a noticeable way to attach a score to Web pages on the basis of the Web connectivity. In this article, we look inside PageRank to disclose its fundamental properties concerning stability, complexity of computational scheme, and critical role of parameters involved in the computation. Moreover, we introduce a circuit analysis that allows us to understand the distribution of the page score, the way different Web communities interact each other, the role of dangling pages (pages with no outlinks), and the secrets for promotion of Web pages.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*Computations on discrete structures* ; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Navigation*

General Terms: Algorithms

Additional Key Words and Phrases: Information retrieval, Markov chains, PageRank, search engines, searching the Web, Web page scoring

## 1. INTRODUCTION

Most commonly used scoring algorithms on the Web, directly derived from Information Retrieval (IR), employ similarity measures, based on a *flat*, vector-space model of each page. Unfortunately, those methods have some limitations and drawbacks when applied to searching the Web, since they only take into account the page content and neglect the graphical structure of the Web. Moreover, these approaches are prone to be cheated, so that pages can be highly ranked if they contain irrelevant but popular words, appropriately located in the page (e.g., in the title). This phenomenon is usually referred to as *search engine persuasion* or *Web spamming* [Marchiori 1997; Pringle et al. 1998].

This research was partially supported by the Italian Ministry of Education under grant COFIN 2000.

Authors' address: Dipartimento di Ingegneria dell'Informazione, University of Siena, Via Roma, 56 – 53100, Siena, Italy; email: {monica,marco,franco}@ing.unisi.it. Marco Gori is the corresponding author.

Some authors have recently concentrated their efforts on how to exploit the topological structure of large hypertextual systems (see, e.g., Brin and Page [1998], Kleinberg [1999], Cohn and Chang [2000], Borodin et al. [2001], and Henzinger [2001]). PageRank [Brin and Page 1998] relies on the "democratic nature of the Web" by using its topology as an indicator of the score to be attached to any page. The model underlying PageRank is tightly related to the citation indexes used in the scientific literature in order to evaluate the importance of publications. More generally, the same idea is found for the estimation of qualifications in self-evaluating groups [Bomze and Gutjahr 1994, 1995]. Here, each member of a group or an institution produces a judgment for all the other members of the same group.

PageRank is used by Google together with a number of different factors, including standard IR measures, proximity, and anchor text (text of links pointing to Web pages) in order to find most relevant answers to a given query. Unfortunately, neither the way these factors are computed nor how they are combined with PageRank are public domain.

In spite of its relevance, the theoretical properties of PageRank are only partially understood. In order to explain the computational properties of the algorithm, most of the authors [Ng et al. 2001b; Zhang and Dong 2000; Brin et al. 1999] cite the general theory of Markov chains [Motwani and Raghavan 1995; Seneta 1981]. However, such a theory can be applied to PageRank only under the assumption that the Web does not contain dangling pages.[1] A related popular ranking algorithm, proposed in Kleinberg [1999] and called HITS, computes two values for each page: the degree of authority and the degree of outdegree. Whereas the authority is a measure of the importance of the page, the outdegree is a measure of the usefulness of the page to act as a starting point for a surfer who wants to find important documents. PageRank and HITS belong to a large class of ranking algorithms, where the scores can be computed as a fixed point of a linear equation [Diligenti et al. 2002]. A completely different approach, rooted to statistics, was proposed in Cohn and Chang [2000] and Cohn and Hofmann [2001]. Starting from PageRank and HITS, some extensions have been proposed by hybrid solutions [Zhang and Dong 2000; Bharat and Henzinger 1998; Diligenti et al. 2002; Richardson and Domingos 2002; Haveliwala 2002; Borodin et al. 2001].

In this article, we look inside PageRank to disclose its fundamental properties concerning the score distribution in the Web and the critical role of parameters involved in the computation. The role of the graphical structure of the Web is thoroughly investigated and some theoretical results which highlight a number of interesting properties of PageRank are established. We introduce the notion of energy, which simply represents the sum of the PageRank for all the pages of a given community,[2] and propose a general circuit analysis which allows us to understand the distribution of PageRank. In addition, the derived energy balance equations make it possible to understand

[1] Dangling pages are pages that do not contain hyperlinks.

[2] A Web community is a subset of related pages. The relation among pages should be based on content similarity and/or on shared location.

the way different Web communities interact each other and to disclose some secrets for promotion of Web pages. In particular, it is pointed out that the energy of a given target community can be driven by a "promoting community" so as to grow linearly with the number of its pages. This property holds regardless of the structure of the promoting community, which makes it very hard its detection. In the next section, PageRank is briefly reviewed, while in Section 1.2 the most important results established in the article are summarized.

### 1.1 PageRank

The basic idea of PageRank is that of introducing a notion of page authority, which is independent of the page content. Such an authority measure only emerges from the topological structure of the Web. In PageRank, the authority reminds the notion of citation in the scientific literature. In particular, the authority of a page $p$ depends on the number of incoming hyperlinks (number of citations) and on the authority of the page $q$ which cites $p$ with a forward link. Moreover, selective citations from $q$ to $p$ are assumed to provide more contribution to the score of $p$ than uniform citations. Hence, PageRank $x_p$ of $p$ is computed by taking into account the set of pages $pa[p]$ pointing to $p$. According to Brin and Page [1998]:

$$x_p = d \sum_{q \in pa[p]} \frac{x_q}{h_q} + (1 - d)\,. \tag{1}$$

Here $d \in (0, 1)$ is a DUMPING FACTOR and $h_q$ is the OUTDEGREE of $q$, that is the number of hyperlinks outcoming from $q$. When stacking all the $x_p$ into a vector $\boldsymbol{x}$, we get

$$\boldsymbol{x} = d\boldsymbol{W}\boldsymbol{x} + (1 - d)\mathbb{I}_N, \tag{2}$$

where $\mathbb{I}_N = [1, \ldots, 1]'$ and $\boldsymbol{W} = \{w_{i,j}\}$—the TRANSITION MATRIX—is such that $w_{i,j} = 1/h_j$ if there is a hyperlink from $j$ to $i$ and $w_{i,j} = 0$, otherwise. Thus, $\boldsymbol{W}$ is a nonnull matrix, where each column either sums to 1 or to 0. More precisely, the $j$th column $\boldsymbol{W}_j$ is null if page $j$ does not contain hyperlinks. Otherwise, $\boldsymbol{W}_j$ can be constructed by the normalization of the $j$th row of the Web adjacency matrix.

Brin and Page [1998] report a simple iterative algorithm based on Eq. (1). They introduce the PageRank dynamics

$$\boldsymbol{x}(t) = d\boldsymbol{W}\boldsymbol{x}(t - 1) + (1 - d)\mathbb{I}_N. \tag{3}$$

It can easily be proven (see Section 2.2) that the system is stable and that the sequence $\{\boldsymbol{x}(t)\}$ always converges to the stationary solution of the linear system (2), provided that $d < 1$. Actually, the method used by Google and defined by Eq. (3) is just the Jacobi algorithm for solving linear systems (see Golub and Van Loan [1993, pp. 506–509]).

A slightly different approach to the computation of PageRank was proposed in Brin et al. [1998, 1999]. In that case, the following equation

$$\boldsymbol{x}(t) = d\boldsymbol{W}\boldsymbol{x}(t-1) + \frac{\alpha(t-1)}{N}\mathbb{1}_N, \tag{4}$$

is assumed, where, for each $t$, $\alpha(t-1) = \|\boldsymbol{x}(t-1)\| - \|d\boldsymbol{W}\boldsymbol{x}(t-1)\|$ in order to force the condition $\|x(t)\|_1 = 1$. System (4) produces a normalized version of PageRank, and converges to $\frac{\boldsymbol{x}^*}{\|\boldsymbol{x}^*\|_1}$, where $\boldsymbol{x}^*$ is the solution of Eq. (2). This results is proven in Section 2.2, where there is also a further discussion on related models.

## 1.2 Main Results

This article presents an in-depth analysis of PageRank by discussing important issues concerning the interaction amongst communities and their promotion. Moreover, the article addresses important complexity issues of the PageRank computation. The most significant results disclosed in the article can be summarized as follows:

(1) Web Communities and Their Interactions: Energy Balance. We define the notion of community with the associated energy. A community is any subgraph $\boldsymbol{G}_I$ of the Web and its energy is the sum $E_I = \sum_{p \in I} x_i^*$ of the PageRank of all its pages. A community could be a set of pages on a given topic, the researchers' home pages or a Website; the corresponding energy is a measure of its authority. A community is connected to the rest of the Web by a set of internal pages out($I$) that point to other communities/pages and by a set of external pages in($I$) that point to $\boldsymbol{G}_I$. In our analysis, an important role is also played by the set dp($I$), which collects the pages that do not contain hyperlinks (e.g., full text documents). Finally, a community isolated from the rest of the Web (i.e., out($I$) = in($I$) = ∅) is referred to as an island.

In Section 4, a theoretical result (Theorem 4.2) is given, which states that the total energy of $\boldsymbol{G}_I$ depends on four components, as follows:

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{dp}. \tag{5}$$

Here, $|I|$ denotes the number of pages of $\boldsymbol{G}_I$ and represents the "default energy" of the community. The component $E_I^{in}$ is the energy that comes from the other communities pointing to $\boldsymbol{G}_I$. The presence of $E_I^{in}$ in Eq. (5) is coherent with the fact that communities with many references have a high authority. The term $E_I^{out}$ is the energy spread over the Web by the pages in out($I$). The presence of $E_I^{out}$ suggests that having hyperlinks outside $\boldsymbol{G}_I$ leads to decrease the energy. Finally, $E_I^{dp}$ is the energy lost in the dangling pages. In fact, the presence of pages without hyperlinks yields a loss of energy. Section 4 provides details on the computation of the above energies, and on how they are involved in the interaction among communities. In

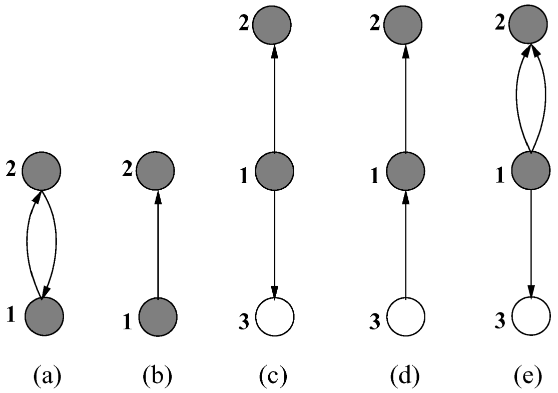

Fig. 1. Some communities and the effect on the energies $E_I^{in}$, $E_I^{out}$, and $E_I^{dp}$.

particular, Theorem 4.2 states that

$$E_I^{in} = \frac{d}{1-d} \sum_{i \in \mathrm{in}(I)} f_i x_i^*, \tag{6}$$

$$E_I^{out} = \frac{d}{1-d} \sum_{i \in \mathrm{out}(I)} (1 - f_i) x_i^*, \tag{7}$$

$$E_I^{dp} = \frac{d}{1-d} \sum_{i \in dp(I)} x_i^*, \tag{8}$$

where $f_i$ is the fraction of the hyperlinks of page $i$ that point to pages in $\boldsymbol{G_I}$ with respect to the total number of hyperlinks outgoing from $i$. Equations (5) through (8) provide useful information on the way the score migrates in the Web. From Eq. (5) we conclude that, in order to maximize the energy, one should not only pay attention to the references received from other communities, but also to the dangling pages, and to the external hyperlinks. Whereas the received references increase the energy, the dangling pages and the hyperlinks pointing outside $\boldsymbol{G_I}$ waste energy, thus reducing the score inside $\boldsymbol{G_I}$.

*Example* 1.1. Let us consider the simple community of two pages represented in Figure 1(a). It has no dangling page, no incoming hyperlink, and no outcoming hyperlink. Hence, the related energies are zero, that is, $E_{I_a}^{out} = 0, E_{I_a}^{in} = 0, E_{I_a}^{dp} = 0$. From Eq. (5) the energy of the community equals the number of pages, that is $E_{I_a} = |I|$. In fact, it can easily be verified that the PageRanks are $x_1^* = x_2^* = 1$ and $E_{I_a} = 2$. In Figure 1(b), a hyperlink was removed, thus transforming page 2 into a dangling page which, according to Eq. (5), causes a loss of energy. We have $x_1^* = 1 - d$ and $x_2^* = 1 - d^2$. Hence, from Eq. (8), it follows that $E_{I_b}^{dp} = d + d^2$ and, consequently,

$$E_{I_b} = 2 - d - d^2,$$

which expresses clearly the loss of energy. Moreover, if we extend the community with a hyperlink pointing to an external page, the energy becomes even smaller due to the loss $E_I^{out}$. In fact, in Figure 1(c), we have $x_1^* = 1 - d$

and $x_2^* = 1 - d + d(1-d)/2$, $f_1 = 1/2$. Hence, $E_{I_c}^{dp} = d + d^2/2$, $E_{I_c}^{out} = d/2$ and, as a consequence,

$$E_{I_c} = 2 - 3d/2 - d^2/2.$$

On the contrary, if we extend Figure 1(b) with a page that points to the community, as in Figure 1(d), the energy will grow because of the presence of the term $E_I^{in}$. In such a case, $x_1^* = 1 - d + d(1-d)$, $x_2^* = 1 - d + d[(1-d) + d(1-d)]$, and $x_3 = 1 - d$. Then $E_{I_d}^{in} = d$, $E_{I_d}^{dp} = d + d^2 + d^3$ and

$$E_{I_d} = 2 - d^2 - d^3.$$

Finally, we can easily check that

$$E_{I_a} > E_{I_d} > E_{I_b} > E_{I_c}.$$

Equation (7) shows that the loss of energy due to each page $i \in \text{out}(I)$ depends also on $f_i$. In particular, it turns out that in order to minimize $E_I^{out}$, the hyperlinks to outside $\boldsymbol{G}_I$ should be in pages with a small PageRank and that have many internal hyperlinks.

*Example* 1.2. Let us consider the community of Figure 1(e). It is the same as the one in Figure 1(c), but, there are two links from page 1 to page 2. The new hyperlink modifies the factor $f_i$, which becomes 2/3. Consequently, $E_{I_e}^{out}$ is smaller than $E_{I_c}^{out}$. In particular we find $x_1 = 1 - d$ and

$$E_{I_e}^{out} = d/3.$$

Likewise, similar considerations hold for $E_I^{dp}$. In fact, Corollary 4.1 proves that

$$E_I^{dp} = d\,|dp\,(I)| + \frac{d^2}{1-d} \sum_{i \in \text{ps}[I]} g_i x_i^*, \tag{9}$$

where $g_i$ is the fraction of the hyperlinks of page $i$ that point to pages in dp($I$), with respect to the total number of hyperlinks outgoing from $i$, and ps[$I$] is the set of the pages with at least a hyperlink to a dangling page.

From our discussion, it turns out that an appropriate organization of a community must avoid energy loss and give rise to a useful distribution of the available energy among the pages. However, Eq. (5) also clarifies that the energy of a community is bounded, that is

$$E_I \leq |I| + E_I^{in}. \tag{10}$$

This is a nice property that makes PageRank a robust measure of the page authority. Small communities with few references cannot have pages with high score.

Moreover, Eqs. (7)–(8) points out also that, when $d$ approaches 1 ($d = 0.85$ in Brin et al. [1999]), $E_I^{out}$ and $E_I^{dp}$ waste most of the available energy. The case $d \approx 1$ is discussed in Section 2.3. It is proved that the energy can even become null. For instance, in all the cases of Figure 1, the energy $E_I$ approaches 0 as $d$ approaches 1. This can easily be verified looking at the equations in the above examples.

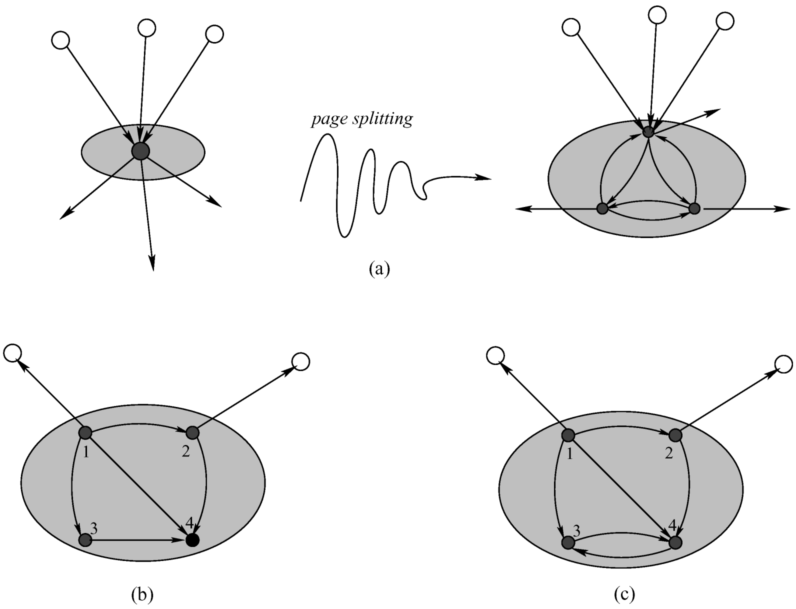

Fig. 2. A pictorial view of some of PageRank's golden rules: (a) Splitting the information into three nodes increases the PageRank; (b) Dangling pages should be avoided or carefully limited; (c) According to rule (c), the hyperlink departing from node "1" is preferable with respect to the hyperlink departing from node "2".

Finally, notice that Eqs. (5)–(10) are all derived in the hypothesis that a homogeneous default contribution is assigned to the rank of each page. The constant vector $\mathbb{I}_N$ can be replaced with any vector $\boldsymbol{E}_N$, which corresponds to attaching different "importance" to the pages. All the above results can be straightforwardly extended to this case.

(2) Page Promotion. Web visibility can be promoted working on both page content and pattern of connections. The circuit analysis carried out in this article provides some indications on the optimization of PageRank. In particular, for a given target community, the energy balance Eqs. (5)–(9) make it possible to derive the following rules which only take into account the topological structure of the community.

   (a) The same content divided into many small pages yields a higher score than the same content into a single large page. This is a straightforward consequence of energy balance in Eq. (5), which states that the community has a "default energy" $|\boldsymbol{I}|$. In Figure 2(a), the default energy goes from 1 to 3.

   (b) Dangling pages give rise to a loss of energy in the community they belong to (Figure 2(b)). The lost energy is small provided that the pages that point to dangling pages have a small score and many hyperlinks pointing to pages of the community. This can be seen from the energy balance in Eq. (5) and from Eq. (8).

   (c) Hyperlinks that point outside the community originate a loss of energy, which is high when the hyperlinks belong to pages with high PageRank.

The lost energy depends also on the fraction of all the links which point outside the community. Hence, this energy is small whenever the pages pointing outside have many hyperlinks to pages of the community (Figure 2(c)). This can be obtained again from the energy balance equation (5) and from Eq. (7), which makes it possible to determine the energy that is lost because of outlinks.

On the other hand, one can promote a target page (site) not only relying on the topological structure of the given community, but also by exploiting external links coming from another community. In this article, we give general indications on the distribution of PageRank and on its migration among communities. In particular, we prove that PageRank is highly affected by the degree of regularity of the pattern of connections. Let us consider a regular graph of degree $k$, that is, a graph where each node has exactly $k$ incoming and $k$ outgoing hyperlinks. If a community is an island and its connectivity is represented by a regular graph, then it is easy to see that the PageRank of all the pages is 1. On the other hand, when the connectivity becomes more and more irregular, then the score of some pages increases, whereas the score of others decreases. The limit case is depicted in Figure 7(a), where all the hyperlinks point to a single page. In such a situation, it is proved that the pointed page gets the maximum PageRank $1 + (N - 1)d$ for an island with $N$ pages (see Theorem 5.1). Of course, such pattern of connections can be maliciously used for the artificial promotion of a Web page. Even if a similar spamming technique could be easily detected, we prove (see Theorem 5.2) that the same growing mechanism of the energy for a given target page (community) can be obtained by using any "promoting community", regardless of its pattern of connections, provided that all its nodes have outlinks to the target page (community) (Figure 3).

Finally, we extend a previous result given in Ng et al. [2001a] by means of Theorem 5.3, stating that if $I$ is a subset of pages that are changed, and $\tilde{\boldsymbol{x}}^*$ is the PageRank after the changes, then

$$\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}^*\|_1 \leq 2\frac{d}{1-d}E_I. \tag{11}$$

Thus, the overall change on the whole Web is proportional to the energy of the modified pages in $I$. Such a result implies that PageRank is robust with respect to changes taking place in small communities of the Web.

(3) PAGERANK COMPUTATION. We give two results concerning the computation of PageRank:

(a) *PageRank Can be Computed on Graphs Changing Over Time.* The PageRank computation is normally based on a given fixed graph. In this article, we prove that we can also provide a more general interpretation of PageRank in the case in which the graph changes over time. This is of significant practical interest. Unlike the static scoring policy, one can calculate the page scoring dynamically, while the crawler is visiting the Web. In particular, the sequence $\{\boldsymbol{x}(t)\}$ remains bounded, even if the transition matrix $\boldsymbol{W} = \boldsymbol{W}(t)$ is updated during the computation of $\boldsymbol{x}(t)$.

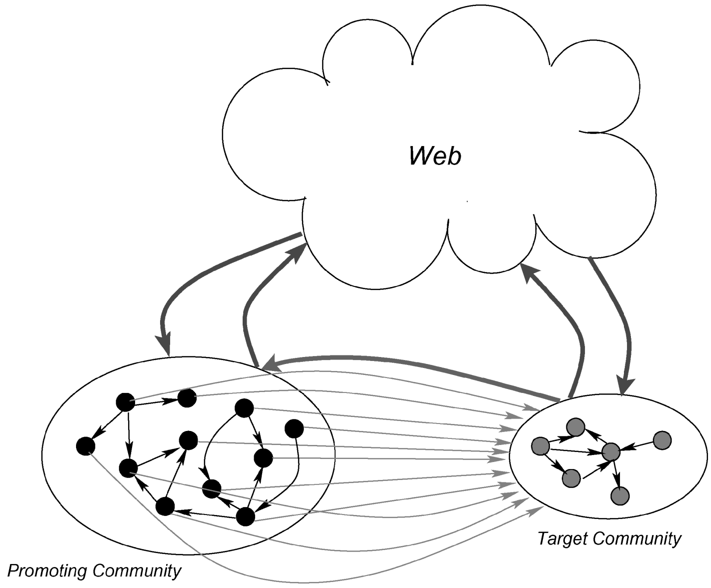

Fig. 3. The energy of the target community grows at least linearly with the number of pages of the promoting community, regardless of its pattern of connections. This makes it very hard to detect such a spamming method.

(b) *PageRank Can be Computed by an Optimal Algorithm.* Since today search engines operate on billions of pages, the scoring systems must face the problem of efficiency. In theoretical computer science, one usually regards the polynomial complexity as the barrier of tractability. In this case, the huge number of parameters involved changes the face of tractability.[3] Let $|H|$ be the number of hyperlinks in the Web. We prove that for a given degree of precision $\epsilon$, the solution can be found with $\mathcal{O}(|H| \log(1/\epsilon))$ floating-point operations. The convergence rate is neither dependent on the connectivity nor on the Web dimension, which offers an explanation of the experimental results reported in Brin et al. [1999].

The ideal computation of PageRank stops whenever the solution yields a sorted list of pages which does not change when continuing to refine the solution. This condition could be adopted in order to choose the degree of precision $\epsilon$. However, in this article, we argue that the above strict interpretation of the solution might not be meaningful, since a slight difference in PageRank is likely not to affect the whole process of page sorting for a given query. Moreover, there is another

[3] Interestingly enough, the issue of tractability arises in other problems involving a huge number of parameters. For instance the problem of computing the gradient of the error function attached to multilayer neural networks, having $w$ parameters, can be faced by using Backpropagation [Rumelhart et al. 1986], an algorithm which takes $\mathcal{O}(w)$ floating-point operations. The adoption of classical numerical algorithms, which neglect the neural network architecture, results in a $\mathcal{O}(w^2)$ complexity, that makes it completely unreasonable to attack problems involving up to a million of parameters.

strong argument to claim that the computation of PageRank does not require the above-mentioned strict constraint. We prove that the ordering induced by PageRank is significantly affected by the choice of the dumping factor $d$.

### 1.3 Organization of the Article

The remainder of this article is organized as follows. In the next section, we discuss some basic properties of PageRank, while in Section 3 we provide a new stochastic interpretation of the model described by Eq. (2). In Section 4, we discuss the interaction of communities by means of a circuit analysis based on the notion of energy, while, in Section 5 we analyze page (community) promotion. In Section 6, we deal with different aspects of the computation of PageRank, and, finally, some conclusions are drawn in Section 7.

## 2. BASIC PROPERTIES OF PAGERANK

In the following, the Web is represented by a graph $\boldsymbol{G}_W = (P, H)$, where each page $p \in P$ is a node and a hyperlink between two pages $h \in H$ is an edge between the corresponding nodes. The set of pages pointing to $p$ is denoted by $pa[p]$, and $ch[p]$ is the set of pages pointed by $p$, while "$'$" stands for the transpose operator on arrays. Moreover, $|\cdot|$ denotes the cardinality operator on sets and the module operator on reals, and $\|\boldsymbol{V}\|_1 = \sum |v_i|$ is the 1–norm of the array $\boldsymbol{V} = [v_1, \ldots, v_n]'$. Given a set of pages $I \subset P$, a community $\boldsymbol{G}_I$ is any subgraph $\boldsymbol{G}_I = (I, H_I)$ of the Web that contains all the hyperlinks between pages in $I$. If the pages belonging to $I$ are disconnected from the rest of the Web, then we say that $\boldsymbol{G}_I$ is an island. Let $\boldsymbol{x}^* = [x_1^*, \ldots, x_N^*]'$ be the PageRank defined by Eq. (2). The energy $E_I$ of community $\boldsymbol{G}_I$ is given by $E_I = \sum_{p \in I} |x_p^*|$, while the energy of the whole Web is denoted by $E_W$. Finally, $\text{out}(I) \subset I$ denotes the pages of the community that point to other pages not in $I$, and $\text{in}(I)$ represents the pages not belonging to $I$ and pointing to pages in $I$.

### 2.1 Removing Dangling Pages

Some of the elementary properties of matrix $\boldsymbol{W}$ will be particularly useful in order to discuss our results. First of all, notice that $\boldsymbol{W}$ is a stochastic matrix except for the null rows.[4] The pages that do not contain hyperlinks are called *dangling pages* and will have a special role in the following discussion. In fact, the presence of dangling pages prevents the direct application of the results from the theory of stochastic matrices (see Seneta [1981]). A simple trick to eliminate dangling pages consists of introducing a dummy page which has a link to itself and is pointed by every dangling page. Thus, the extended graph turns out to be $\overline{\boldsymbol{G}}_W = (\overline{P}, \overline{H})$, where $\overline{P} = P \cup \{N+1\}$ and $\overline{H} = H \cup \{(i, N+1)|$

[4] Stochastic matrices are nonnegative matrices having all columns that sum up to 1.

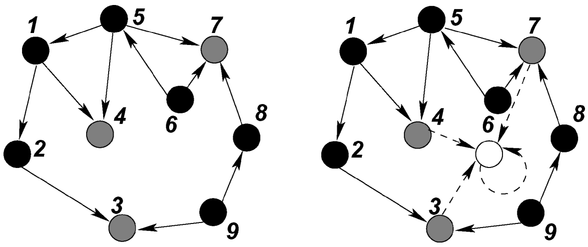

Fig. 4. A trick to eliminate dangling pages: A dummy node with a self-loop is added.

$\nexists j, (i, j) \in H\}$. The transition matrix $\overline{\boldsymbol{W}}$ that corresponds to $\overline{\boldsymbol{G}}_W$ is

$$\overline{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} \\ \boldsymbol{R} & 1 \end{pmatrix},$$

where $\boldsymbol{R} = [r_1, \ldots, r_N]$, and if $i$ is a dangling page then $r_i = 1$, else $r_i = 0$. After such a transformation, the Web has no dangling page (see Figure 4) and $\overline{\boldsymbol{W}}$ is a stochastic matrix.

For instance, referring to Figure 4, the transition matrix $\overline{\boldsymbol{W}}$ is

$$\overline{\boldsymbol{W}} = \left[\begin{array}{ccccccccc|c} 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 1/2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right].$$

The following proposition relates the PageRank of $\overline{\boldsymbol{G}}_W$ to the PageRank of $\boldsymbol{G}_W$.

Proposition 2.1. *Let us consider the dynamical system*

$$\overline{\boldsymbol{x}}(t+1) = d\overline{\boldsymbol{W}}\,\overline{\boldsymbol{x}}(t) + (1-d)\mathbb{1}_{N+1} \tag{12}$$

*related to the extended graph. Then, the following properties hold:*

(a) *Eq.* (12) *has a finite stable equilibrium point if and only if Eq.* (3) *has a finite stable equilibrium point.*

(b) *If* $\overline{\boldsymbol{x}}^*$ *is an equilibrium point of* (12) *and* $\boldsymbol{x}^*$ *is an equilibrium point of* (3), *then* $\overline{\boldsymbol{x}}^* = [\boldsymbol{x}^{*\prime}, 1 + \frac{d}{1-d}\boldsymbol{R}\boldsymbol{x}^*]'$.

(c) *If* $\boldsymbol{x}(0) = \mathbb{1}_N$ *and* $\overline{\boldsymbol{x}}(0) = \mathbb{1}_{N+1}$, *then* $\overline{\boldsymbol{x}}(t) = [\boldsymbol{x}(t)', 1 + \sum_{s=0}^{t-1} d^{t-s}\boldsymbol{R}\boldsymbol{x}(s)]'$.

Proof.

(a) Let $\overline{\boldsymbol{x}}^* = [\boldsymbol{Z}', y]'$ be a finite solution of (12), where $\boldsymbol{Z} \in I\!R^N$ and $y \in I\!R$. Then, with respect to the first $N$ rows of Eq. (12), $\boldsymbol{Z} = d\boldsymbol{W}\boldsymbol{Z} + (1-d)\mathbb{1}_N$, that is, $\boldsymbol{Z}$

is a solution of (3). Moreover, from the last row, we get $y = d\boldsymbol{RZ} + dy + (1-d)$, which yields $y = 1 + \frac{d}{1-d}\boldsymbol{RZ}$.

(b) Vice-versa, let $\boldsymbol{x}^*$ be a solution of (3). Then, by straightforward algebra, it follows that the vector $[\boldsymbol{x}^{*\prime}, 1 + \frac{d}{1-d}\boldsymbol{Rx}^*]^{\prime}$ is a solution of (12).

(c) The last statement can be proved by induction on $t$. For $t = 0$, $\overline{\boldsymbol{x}}(0) = [\boldsymbol{x}^{\prime}(0), 1]^{\prime}$ holds by hypothesis. Let $t > 0$ and assume by induction that assertion (c) holds for $t-1$. Then, the first $N$ components of $\overline{\boldsymbol{x}}(t) = d\overline{\boldsymbol{W}}\overline{\boldsymbol{x}}$ $(t-1) + (1-d)\mathbb{1}_{N+1}$ satisfy $\boldsymbol{x}(t) = d\boldsymbol{Wx}(t-1) + (1-d)\mathbb{1}_N$, while the last component can be calculated by

$$\begin{aligned} y(t) &= d\boldsymbol{Rx}(t-1) + dy(t-1) + 1 - d \\ &= d\boldsymbol{Rx}(t-1) + d\left(1 + \sum_{s=0}^{t-2} d^{t-1-s}\boldsymbol{Rx}(s)\right) + 1 - d \\ &= 1 + \sum_{s=0}^{t-1} d^{t-s}\boldsymbol{Rx}(s). \quad \square \end{aligned}$$

Other authors (e.g., Ng et al. [2001a] and Brin et al. [1999]) use a different trick in order to eliminate dangling pages. In fact, they assume that dangling pages are special nodes pointing all the pages of the Web. Hence, the PageRank equation becomes

$$\boldsymbol{x} = d(\boldsymbol{W} + \boldsymbol{V})\boldsymbol{x} + \frac{(1-d)}{N}\mathbb{1}_N \tag{13}$$

where $\boldsymbol{V} = \frac{1}{N}\mathbb{1}_N\boldsymbol{R}$ and $\boldsymbol{R}$ is the vector of Proposition 2.1. The approach is motivated by a stochastic interpretation of PageRank which will be clarified in Section 3.

Now we prove that Eqs. (12) and (13) describe related dynamical systems. Actually, Eqs. (2)–(3), Eq. (12), and Eq. (13) yield the same ranking scheme, modulo normalization. Therefore, Eq. (3) can be used instead of Eq. (13), taking into account the presence of dangling pages, but keeping a simpler mathematical formulation. Let us first introduce the following lemma.

LEMMA 2.1. *Let us consider the dynamical system* (3) *and its stationary point* $\boldsymbol{x}^*$. *Let* $S$ *be the set of dangling pages. Then,* $\boldsymbol{x}^* > 0$ *and*

$$\|\boldsymbol{x}^*\|_1 = N - \frac{d}{1-d}\sum_{i\in S} x_i^*. \tag{14}$$

*Moreover, if* $\boldsymbol{x}(0) = \mathbb{1}_N$, *then, for each* $t$, $\boldsymbol{x}(t) > 0$ *and*

$$\|\boldsymbol{x}(t)\|_1 = N - \sum_{i\in S}\sum_{k=1}^{t} d^k x_i(t-k)\,. \tag{15}$$

PROOF. The inequality $\boldsymbol{x}(t) \geq 0$ follows directly by induction on $t$, observing that $\boldsymbol{W}$ contains only nonnegative components, $\boldsymbol{x}(0) = \mathbb{1}_N \geq 0$, and $\boldsymbol{x}(t-1) = d\boldsymbol{Wx}(t) + (1-d)\mathbb{1}_N$. Moreover, $\boldsymbol{x}^* \geq 0$ since $\boldsymbol{x}^* = \lim_{t\to\infty}\boldsymbol{x}(t) \geq 0$. The proof of Eqs. (14) and (15) consists of two steps. In the former step, we assume $S = \emptyset$, that is, there is no dangling page. Then, the general case is considered.

*Case* $S = \emptyset$. Equation (15) is proven by induction on $t$.

(1) For $t = 0$, $\boldsymbol{x}(0) = \mathbb{1}_N$ and $\boldsymbol{x}(0) \geq \boldsymbol{0}$; hence Eq. (15) holds straightforwardly.

(2) Let $t > 0$ be and assume by induction that $\|\boldsymbol{x}(t)\|_1 = N$ holds. Then, based on the definition of stochastic matrices,

$$\begin{aligned}\|\boldsymbol{x}(t+1)\|_1 = \mathbb{1}'_N \boldsymbol{x}(t+1) &= d\,\mathbb{1}'_N \boldsymbol{W}\boldsymbol{x}(t) + (1-d)\mathbb{1}'_N \mathbb{1}_N \\ &= d\,\mathbb{1}'_N \boldsymbol{x}(t) + (1-d)N = N.\end{aligned}$$

Moreover,

$$\|\boldsymbol{x}^*\|_1 = \|\lim_{t\to\infty} \boldsymbol{x}(t+1)\|_1 = \lim_{t\to\infty} \|\boldsymbol{x}(t+1)\|_1 = N,$$

which proves (14).

*Case* $S \neq \emptyset$. Consider the graph $\overline{\boldsymbol{G}}_W$ extended from $\boldsymbol{G}_W$. Since $\overline{\boldsymbol{G}}_W$ has no dangling pages, we can use the result of case $S = \emptyset$ which yields, $\forall t$, $\|\overline{\boldsymbol{x}}(t)\|_1 = \|\overline{\boldsymbol{x}}^*\|_1 = N + 1$.

From Proposition 2.1,

$$N + 1 = \|\overline{\boldsymbol{x}}^*\|_1 = \|\boldsymbol{x}^*\|_1 + \frac{d}{1-d}\boldsymbol{R}\boldsymbol{x}^* + 1. \tag{16}$$

Hence, Eq. (14) follows by solving Eq. (16) with respect to $\|\boldsymbol{x}^*\|_1$ and observing that $\boldsymbol{R}\boldsymbol{x}^* = \sum_{i\in S} x_i^*$ holds by the definition of $\boldsymbol{R}$. Similarly, from Proposition 2.1,

$$N+1 = \|\overline{\boldsymbol{x}}(t)\|_1 = \|\boldsymbol{x}(t)\|_1 + 1 + \sum_{s=0}^{t-1} d^{t-s}\boldsymbol{R}\boldsymbol{x}(s) = \|\boldsymbol{x}(t)\|_1 + 1 + \sum_{i\in S}\sum_{r=1}^{t} d^r x_i(t-r),$$

which, in turn, yields Eq. (15). □

In fact, based on the above result, the solution $\tilde{\boldsymbol{x}}^*$ of Eq. (13) can be reconstructed by normalization from PageRank.

PROPOSITION 2.2. *The fixed points* $\boldsymbol{x}^*$ *and* $\tilde{\boldsymbol{x}}^*$ *of* (2) *and* (13) *are related by* $\tilde{\boldsymbol{x}}^* = \boldsymbol{x}^*/\|\boldsymbol{x}^*\|_1$.

PROOF. Since the spectral radius of both matrices $d\boldsymbol{W}$ and $d\boldsymbol{W}+d\boldsymbol{V}$ is smaller than 1, then Eqs. (2) and (13) have unique solutions. Thus, let $\boldsymbol{x}^*$ be a solution of (2). Then,

$$\begin{aligned}\frac{\boldsymbol{x}^*}{\|\boldsymbol{x}^*\|_1} &= \frac{d}{\|\boldsymbol{x}^*\|_1}\boldsymbol{W}\boldsymbol{x}^* + \frac{(1-d)}{\|\boldsymbol{x}^*\|_1}\mathbb{1}_N \\ &= \frac{d}{\|\boldsymbol{x}^*\|_1}\boldsymbol{W}\boldsymbol{x}^* + \frac{(1-d)}{N\|\boldsymbol{x}^*\|_1}\left(\frac{d}{1-d}\boldsymbol{R}\boldsymbol{x}^* + \|\boldsymbol{x}^*\|_1\right)\mathbb{1}_N \\ &= \frac{d}{\|\boldsymbol{x}^*\|_1}\boldsymbol{W}\boldsymbol{x}^* + \frac{d}{N\|\boldsymbol{x}^*\|_1}\mathbb{1}_N(\boldsymbol{R}\boldsymbol{x}^*) + \frac{(1-d)}{N}\mathbb{1}_N \\ &= d(\boldsymbol{W}+\boldsymbol{V})\frac{\boldsymbol{x}^*}{\|\boldsymbol{x}^*\|_1} + \frac{(1-d)}{N}\mathbb{1}_N,\end{aligned}$$

where $N \doteq \dfrac{d}{1-d}\boldsymbol{R}\boldsymbol{x}^* + \|\boldsymbol{x}^*\|_1$, because of Lemma 2.1. □

Thus, $\tilde{\boldsymbol{x}}^*$ gives the same ordering of pages as PageRank. However, in this article, we prefer to remove dangling pages by the method of Proposition 2.1, which allows us to use Eq. (3) directly. In fact, our analysis will show that interesting properties of PageRank are disclosed more easily when (3) is used instead of (13).

### 2.2 Stability and Dynamical Updating

The following proposition defines a property of $\boldsymbol{W}$ that is useful for our discussion. The proof of the proposition is a well-known result which we briefly resketch in the following for the sake of completeness.

PROPOSITION 2.3. *For any Web graph $\boldsymbol{G}_{\mathrm{W}}$, the spectral radius $\rho(\boldsymbol{W})$ fulfills $\rho(\boldsymbol{W}) \leq 1$. Moreover, if $\boldsymbol{G}_{\mathrm{W}}$ has no dangling page, then $\rho(\boldsymbol{W}) = 1$ and $\mathbb{1}_N$ is the left eigenvector associated with the largest eigenvalue $\lambda_{max} = 1$.*

PROOF. Based on the Gerschgorin Theorem [Golub and Van Loan 1993, pp. 341–342], the eigenvalues of matrix $\boldsymbol{W}$ belong to

$$S = \bigcup_i \left\{ c \in \mathbb{C},\ |c - w_{i,i}| \leq \sum_{j \neq i} |w_{j,i}| \right\}.$$

Since $\sum_{j=1}^{N} |w_{i,j}| \leq 1$ by definition, then the spectral radius of $\boldsymbol{W}$ fulfills $\rho(\boldsymbol{W}) \leq 1$. Moreover, if the given graph has no dangling page, then $\boldsymbol{W}$ is a stochastic matrix and every column sums up to 1. Thus, $\mathbb{1}'_N \boldsymbol{W} = \mathbb{1}'_N$ which also implies $\rho(\boldsymbol{W}) = 1$. □

An immediate consequence of Proposition 2.3 (see Golub and Van Loan [1993, p. 508]) guarantees that the PageRank scheme is well founded whenever $0 \leq d < 1$.

PROPOSITION 2.4. *Let $0 \leq d < 1$ hold. Equation (2) admits a unique solution $\boldsymbol{x}^* = (1-d)(\boldsymbol{I} - d\boldsymbol{W})^{-1}\mathbb{1}_N$. Moreover, the dynamics of Eq. (3) is such that $\lim_{t\to\infty} \boldsymbol{x}(t) = \boldsymbol{x}^*$ for any initial state $\boldsymbol{x}(0)$.*

Lemma 2.1 describes a fundamental property of PageRank: Regardless of the graph topology, the sum of the score over all the pages $\|\boldsymbol{x}(t)\|_1$, at each time step $t$, is always bounded by the number of pages $N$, provided that $\boldsymbol{x}(0) = \mathbb{1}_N$. Moreover, this result holds, in the limit, for $\|\boldsymbol{x}^*\|_1$.

An immediate corollary of Lemma 2.1 is derived in the case in which there is no dangling page.

COROLLARY 2.1. *If $\boldsymbol{G}_{\mathrm{W}}$ has no dangling page and $\boldsymbol{x}(0) = \mathbb{1}_N$, then*

$$\forall\, t \geq 0 : \ \|\boldsymbol{x}(t)\|_1 = \|\boldsymbol{x}^*\|_1 = N.$$

Lemma 2.1 also points out that, due to the presence of dangling pages, the Web loses part of its energy. The *energy loss* is represented by the negative term of Eq. (14), $E^{dp} = d/(1-d)\sum_{i\in S} x_i^*$. In the worst (ideal) case, when all the pages are dangling pages, $x_p = (1-d)$ for each $p$, and, therefore, the loss of energy is $dN$. If $d$ approaches 1 (e.g., $d = 0.85$ is the value suggested in Brin et al. [1999]), the loss can be an important percentage of the available energy.

Corollary 2.1 can be extended to the time–variant system

$$\boldsymbol{x}(t+1) = d\boldsymbol{W}(t)\boldsymbol{x}(t) + (1-d)\mathbb{1}_N.$$

Such a system corresponds to the case in which PageRank is computed online while the crawler of the search engine is visiting the Web.[5] In this case, we can prove that $\|\boldsymbol{x}(t)\|_1 = N$ ($\|\boldsymbol{x}(t)\|_1 \leq N$) holds, if $\boldsymbol{G}_W$ has no dangling pages (respectively, has dangling pages). In order to deal with the growth of the pages downloaded during the crawling (change of $N$), we can simply embed $\boldsymbol{x}$ into an infinite dimensional space ($\boldsymbol{x} \in \mathbb{R}^\infty$) so as to accommodate all incoming new pages. Moreover, we assume that the unknown pages have a single internal hyperlink and are not pointed to by other pages, so that $x_p(t) = 1$ before $p$ is visited. Thus, the most general version of Corollary 2.1 states that $\|\boldsymbol{x}_{V(t)}(t)\|_1 = |V(t)|$, where $V(t)$ is the set of pages visited at time $t$ and $\boldsymbol{x}_{V(t)}(t)$ is the related subvector of $\boldsymbol{x}(t)$.

Finally, let us consider Eq. (4), which has been suggested in Brin et al. [1998, 1999] in order to compute PageRank. The next theorem proves that system (4) is stochastic and produces a normalized version of PageRank. In fact, the dynamical system (4) converges to the linear system (13) and, therefore, PageRanks defined by (3) and (4) are equivalent.

THEOREM 2.1. *The following two systems*

$$\boldsymbol{x}(t) = d\boldsymbol{W}\boldsymbol{x}(t-1) + \frac{\alpha(t-1)}{N}\mathbb{1}_N, \tag{17}$$

$$\boldsymbol{x}(t) = \left[d(\boldsymbol{W}+\boldsymbol{V}) + \frac{(1-d)}{N}\mathbb{1}_N\mathbb{1}_N'\right]\boldsymbol{x}(t-1), \tag{18}$$

*produce the same sequence, provided that* $\|\boldsymbol{x}(0)\|_1 = 1$, $\boldsymbol{V} = \frac{1}{N}\mathbb{1}_N\boldsymbol{R}$, *and* $\alpha(t) = d\boldsymbol{R}\boldsymbol{x}(t) + (1-d)$, $t \geq 0$. *Moreover,* $d(\boldsymbol{W}+\boldsymbol{V}) + \frac{(1-d)}{N}\mathbb{1}_N\mathbb{1}_N'$ *is a stochastic matrix, the sequence* $\{\boldsymbol{x}(t)\}$ *satisfies* $\|x(t)\|_1 = 1$, $t \geq 0$, *and converges to* $\frac{\boldsymbol{x}^*}{\|\boldsymbol{x}^*\|_1}$.

PROOF. We will prove that both systems (17) and (18) are equivalent to

$$\boldsymbol{x}(t) = (\boldsymbol{W}+\boldsymbol{V})\boldsymbol{x}(t-1) + \frac{(1-d)}{N}\mathbb{1}_N \tag{19}$$

which, according to Theorem 2.2, converges to $\frac{\boldsymbol{x}^*}{\|\boldsymbol{x}^*\|_1}$ In fact, Eq. (17) is the same as Eq. (19) by definition of $\alpha(t)$ and $\boldsymbol{V}$. Moreover, since $\boldsymbol{W}+\boldsymbol{V}$ represents the transition matrix of a graph with no dangling page (see the discussion about Eq. (13)), it follows that $\|x(t)\|_1 = \mathbb{1}_N' x(t) = 1$, $t \geq 0$. Then, system (19) is equivalent to system (18). Furthermore, by straightforward algebra, it can be seen that all the columns of the matrix $d(\boldsymbol{W}+\boldsymbol{V}) + \frac{(1-d)}{N}\mathbb{1}_N\mathbb{1}_N'$ sum up to 1. Finally, according to Proposition 2.2, system (19) converges to $\frac{\boldsymbol{x}^*}{\|\boldsymbol{x}^*\|_1}$. □

[5] Google search engine completely rebuilds its indexes, on average, after a few weeks, instead of performing dynamic updating.

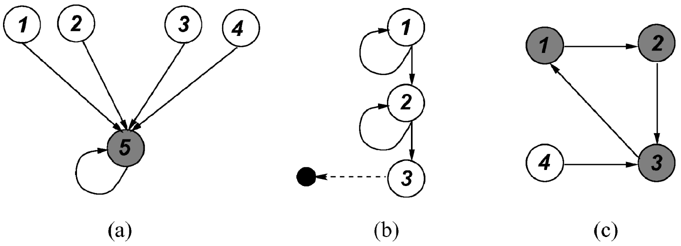

Fig. 5. Examples of essential (grey circles) and inessential nodes (white circles). The black node (connected with a dashed line in (b)) is the dummy node of the extended Web.

## 2.3 The Dumping Factor Boundary Values

When $d$ is set to the boundary values 0 or 1, the ranking system ceases to operate properly. In fact, if $d = 0$, all the PageRanks equals 1. On the other hand, if $d = 1$, the sequence $\{\boldsymbol{x}(t)\}$ might not converge (even if it is still bounded). Interestingly, if $d = 1$, many pages would have a zero PageRank. This can be seen by applying the theory of Markov chains[6] to the sequence $\{\overline{\boldsymbol{x}}(t)\}$ generated by Eq. (12) on the extended graph. The results we address are based on the concept of ESSENTIAL and INESSENTIAL nodes. Intuitively, a page is essential if it belongs to a subgraph where a surfer can be entrapped, that is, a group of connected pages without any way of escaping. On the other hand, an inessential page $p$ has some escaping path: when the surfer follows that path, it cannot come back to $p$. Figure 5 shows some examples: the grey circles represent the essential nodes, while the white circles represent the inessential ones. Notice that since the theory of Markov chains can be applied only to stochastic matrices, in order to define the inessential pages, we use the extended Web, which has no dangling pages. For example, the dangling pages are inessential pages (see Figure 5(b)), since they are connected to the dummy node in the extended Web, where any surfer gets trapped. Formally, we have the following definition.

*Definition* 2.1. A node $p$ is said to be INESSENTIAL if there exists a node $p'$ such that $p \to p'$ ($p'$ can be reached by a path from $p$) and $p' \not\to p$ (there is no path from $p'$ to $p$).

According to the theory of Markov chains, if $i$ is an inessential node, then $x_i^* = 0$ (see Seneta [1981]). Thus, the scoring of inessential pages approaches 0 when $d \to 1$. In other words, the communities without out–links have an increasing advantage as $d \approx 1$. In theory, the term $d$ could be used to control this mechanism and to move part of the energy from inessential to essential nodes or vice-versa.

*Example* 2.1. The concept of essential and inessential nodes has a particular meaning in the context of PageRank, since those nodes that are inessential gain a low PageRank, while essential nodes are those where most of the energy is concentrated. Let us consider Figure 5 again. In particular, referring to Figure 5(a), $\boldsymbol{x}^* = [1-d, 1-d, 1-d, 1-d, 4d+1]'$ which, in accordance with

[6] If $d = 1$, then Eq. (12) is the stochastic system $\overline{\boldsymbol{x}}(t+1) = \overline{\boldsymbol{W}}\overline{\boldsymbol{x}}(t)$.

Eq. (38), guarantees that $x_i^* \to 0, i = 1, \dots, 4$, as $d \to 1$, while $x_5^* \to |I|$, being $|I|$ the cardinality of $I$. For the case of Figure 5(b),

$$\boldsymbol{x}^* = \left[\frac{2(1-d)}{2-d}, \frac{4(1-d)}{(2-d)^2}, \frac{(d^2-2d+4)(1-d)}{(2-d)^2}\right]'$$

holds. Moreover, $\boldsymbol{x}^* \to \boldsymbol{0}$ for $d \to 1$.

Finally, with respect to Figure 5(c),

$$\boldsymbol{x}^* = \left[\frac{1+d+2d^2}{1+d+d^2}, \frac{1+d+d^2+d^3}{1+d+d^2}, \frac{1+2d+d^2}{1+d+d^2}, 1-d\right]',$$

and

$$\boldsymbol{x}^* \to \left[\frac{4}{3}, \frac{4}{3}, \frac{4}{3}, 0\right]'$$

for $d \to 1$, that is, at the steady state, the total energy of the community, $|I|$, is distributed among the three essential nodes.

## 3. STOCHASTIC INTERPRETATION

The random walk theory has been widely used to compute the authority of a page in the Web [Ng et al. 2001b; Lempel and Moran 2000; Brin et al. 1999]. A common assumption is that the rank of a page $p$ is modeled by the probability $\mathcal{P}(p)$ of reaching that page during a random walk in the Web. The motivation underlying the stochastic interpretation of PageRank is that a random surfer spends a lot of time in important pages: in fact, $\mathcal{P}(p)$ describes how often a page will be visited. However, the random walk theory holds for the extended models (see Eqs. (12) and (13)) which remove the dangling pages.

In this section, we provide a stochastic interpretation of PageRank computed according to Eq. (3), which is appropriate to handle also the problem of dangling pages.

*Definition* 3.1. Let us denote $\mathcal{S}_i(t) = (p_i(t), a_i(t))$, $p_i(t) \in P \cup \{$"*idle*"$\}$, $a_i(t) \in \mathcal{A} = \{l, s\}$, and $i = (t_{i(0)}, p_{i(0)})$. $\mathcal{S}_i$ defines a random Markovian *Web surfer,* which starts its navigation from page $p_{i(0)}$ at time $t_{i(0)}$. Function $\mathcal{S}_i$ describes the position on $\boldsymbol{G}_W$ at time $t \geq t_{i(0)}$ and the action which will be performed next. In particular, $\mathcal{S}_i$ can perform two actions $a_i(t) \in \mathcal{A}$:

—follow a hyperlink (action $l$);
—stop surfing and become "*idle*" (action $s$).

Therefore, we define:

—$\mathcal{P}(a_i(t) = a \mid p_i(t) = p)$ as the probability of performing $a \in \mathcal{A}$, provided that $\mathcal{S}_i$ stays in $p$ at time $t$;
—$\mathcal{P}(p_i(t+1) = p | p_i(t) = q, a_i(t) = l)$ as the probability of moving from page $q$ to page $p$, provided that $\mathcal{S}_i$ is located in $q$ at time $t$ and it will follow a hyperlink;

—$\mathcal{P}(p_i(t+1)=p|p_i(t)=q) = \sum_{a\in\mathcal{A}} \mathcal{P}(p_i(t+1) = p \mid p_i(t) = q, a_i(t) = a)\mathcal{P}(a_i(t) = a \mid p_i(t) = q)$ as the probability of moving to page $p$, provided that $\mathcal{S}_i$ is located in $q$ at time $t$ and the performed action is unknown;

—$\mathcal{P}(p_i(t) = p)$ as the probability that $\mathcal{S}_i$ is located in $p$ at time $t$.

Since the surfers are Markovian, the future behavior of $\mathcal{S}_i$ depends on the current page $p \in P$. In fact, when visiting an interesting page, $\mathcal{S}_i$ will likely follow a hyperlink towards a neighboring page. If the current page is not significant, the surfer will get "bored" and stop surfing. Hence, at time $t+1$, $\mathcal{S}_i$ is located in $p$ with probability

$$\begin{aligned}\mathcal{P}(p_i(t+1)=p) &= \sum_{q\in P}\mathcal{P}(p_i(t+1)=p \mid p_i(t)=q)\mathcal{P}(p_i(t)=q)\\ &= \sum_{q\in P}\left(\sum_{a\in\mathcal{A}}\mathcal{P}(p_i(t+1)=p \mid p_i(t)=q, a_i(t)=a)\mathcal{P}(a_i(t)=a \mid p_i(t)=q)\right)\mathcal{P}(p_i(t)=q).\end{aligned}$$

Let $M^\alpha$ be a multiset operator that replicates $\alpha$–times the set $M$ and let $Q$ be defined as $Q = \bigcup_{t_{i(0)}\geq 0,\ p_{i(0)}\in P}\{(t_{i(0)}, p_{i(0)})\}$. The PageRank scheme models the behavior of the set of independent surfers $U_\alpha = \bigcup_{i\in Q}\{\mathcal{S}_i\}^\alpha$. Thus, for each time $t_{i(0)}$ and each page $p_{i(0)}$, $U_\alpha$ contains $\alpha$ surfers $\mathcal{S}_i$ which start navigating from $p_{i(0)}$ at time $t_{i(0)}$. Hence, the expected number of surfers $E(n_p(t+1))$ which lie in $p$ at time $t+1$ can be calculated as

$$\begin{aligned}E(n_p(t+1)) &= \sum_{i\in Q}\mathcal{P}(p_i(t+1)=p)\\ &= \alpha + \sum_{i\in Q}\sum_{q\in P}\mathcal{P}(p_i(t+1)=p \mid p_i(t)=q)E(n_q(t)).\end{aligned}\tag{20}$$

Equation (20) becomes Eq. (1) when, for each $p, q \in P$, we make the following assumptions:

(1) $\alpha = 1-d$, that is, $1-d$ surfers start navigating from each page, at each time step;

(2) if $p$ is not a dangling page, $\mathcal{P}(a_i(t) = l \mid p_i(t) = p) = d$, that is, every $\mathcal{S}_i$ has the same constant probability $d$ of following a link in each page (and a constant probability $1-d$ of stopping the navigation);

(3) $\mathcal{P}(p_i(t+1) = p \mid p_i(t) = q, a_i(t) = l) = 1/h_q$ if there exists a link $(q, p)$, and 0 otherwise. Here $h_q$ is the outdegree of page $q$, that is, the number of outcoming links from $q$. This assumption makes the surfer "random", since all the outcoming links from a page are followed with the same probability;

(4) if $p$ is a dangling page, $\mathcal{P}(a_i(t) = s \mid p_i(t) = p) = 1$, that is, the surfers stop the navigation on dangling pages.

The above result can be formalized as follows.

THEOREM 3.1. *If the assumptions* 1, 2, 3, *and* 4 *hold and* $x_p(0) = E(n_p(0))$ *then the PageRank* $x_p(t)$*, computed according to Eq.* (1)*, of page* $p$ *at time* $t$ *counts the expected number of surfers* $E(n_p(t))$ *that lie in* $p$ *at* $t$.

The above stochastic interpretation of PageRank is slightly different from those presented in Brin and Page [1998], Ng et al. [2001a], and Lempel and Moran [2000]. In fact, PageRank is usually modeled using the classical random walk theory on graphs, with the following assumptions:

—there is only one Web surfer and $x_p(t)/N$ represents the probability that the surfer lies in $p$ at time $t$;
—the surfer starts navigating at time 0 and the initial page is set according to a distribution which is proportional to $\boldsymbol{x}(0)$;
—the surfer never stops navigating; at each time instant $t$, it still may become "bored" with probability $1-d$ and *jumps* to another Web page with a uniform probability.

Such an interpretation to PageRank works only when there is no dangling page. In Ng et al. [2001a] and Lempel and Moran [2000], each dangling page is assumed to be a special node pointing to all the pages (see Section (2.1)), which means that classical random walk is used to model (18) instead of Eq. (3): matrix $d\boldsymbol{W}$ collects the probabilities that the surfer moves from a page following a hyperlink; matrix $(1-d)\mathbb{1}_{N\times N}$ defines the probabilities of jumping out of pages.

According to Theorem (2.1), the two stochastic models are equivalent. However, the interpretation of PageRank introduced in this article provides a direct model for Eq. (3) and simplifies the comprehension of some properties of PageRank. For example, according to our interpretation, surfers are forced to stop the navigation in dangling pages, which clearly explains why dangling pages produce a loss of energy.

## 4. COMMUNITIES AND ENERGY BALANCE

This section is devoted to a simple analysis of the mechanisms that are involved in the interaction among communities. Let $\mathrm{dp}(I)$ be the set of dangling pages in $\boldsymbol{G}_I$ and let $R$ and $C$ be subsets of pages in $\boldsymbol{G}_I$. Moreover, $\boldsymbol{x}_R$ stands for a vector that contains only the components $R$ of $\boldsymbol{x}$ and $\boldsymbol{W}_{R,C}$ is the submatrix of $\boldsymbol{W}$, which contains only the rows in $R$ and the columns in $C$.

*Definition* 4.1. The *vectorial scores*

$$\boldsymbol{x}_I^i = (1-d)(\boldsymbol{I}-d\boldsymbol{W}_{I,I})^{-1}\mathbb{1}_{|I|} \tag{21}$$

$$\boldsymbol{x}_I^e = d(\boldsymbol{I}-d\boldsymbol{W}_{I,I})^{-1}\boldsymbol{W}_{I,\mathrm{in}(I)}\boldsymbol{x}_{\mathrm{in}(I)} \tag{22}$$

are referred to as the INTERNAL and the EXTERNAL PAGERANK of the community $\boldsymbol{G}_I$, respectively.

The external PageRank $\boldsymbol{x}_I^e$ depends linearly on $\boldsymbol{x}_{\mathrm{in}(I)}$, a vector that includes only the external nodes that have hyperlinks to the community. A similar comment holds for the internal PageRank. Due to the linearity of Eq. (2), PageRank meets the decomposition property.

THEOREM 4.1. *For any community $\boldsymbol{G}_I$, the page score can be decomposed as follows*

$$\boldsymbol{x}_I^* = \boldsymbol{x}_I^i + \boldsymbol{x}_I^e = \sum_{p \in \boldsymbol{G}_I} \boldsymbol{\Phi}_p + \sum_{p \in \mathrm{in}(I)} \boldsymbol{\Psi}_p x_p^*,$$

*where $\boldsymbol{\Psi}_p$ is the pth column of $d(\boldsymbol{I} - d\boldsymbol{W}_{I,I})^{-1}\boldsymbol{W}_{I,\mathrm{in}(I)}$ and $\boldsymbol{\Phi}_p$ is the pth column of $(1-d)(\boldsymbol{I} - d\boldsymbol{W}_{I,I})^{-1}$.*

PROOF. By an appropriate reordering of the pages, $\boldsymbol{x}^{*'} = [\boldsymbol{x}_I^{*'}, \boldsymbol{x}_{\mathrm{in}(I)}^{*'}, \boldsymbol{x}_{P\setminus(I\cup\mathrm{in}(I))}^{*'}]$, Eq. (2) becomes

$$\boldsymbol{x}^{*'} = d \begin{pmatrix} \boldsymbol{W}_{I,I} & \boldsymbol{W}_{I,\mathrm{in}(I)} & \boldsymbol{0} \\ \cdots & \cdots & \cdots \end{pmatrix} \boldsymbol{x}^{*'} + (1-d)\mathbb{1}_N,$$

where, for the sake of simplicity, only the rows of $\boldsymbol{W}$ which play an active role in the calculation are displayed. Solving the first $|I|$ equations with respect to $\boldsymbol{x}_I^*$, we get

$$\boldsymbol{x}_I^* = (1-d)(I - d\boldsymbol{W}_{I,I})^{-1}\mathbb{1}_{|I|} + d(I - d\boldsymbol{W}_{I,I})^{-1}\boldsymbol{W}_{I,\mathrm{in}(I)}\boldsymbol{x}_{\mathrm{in}(I)}^*$$

The hypothesis of the theorem follows by the definition of $\boldsymbol{x}_I^i, \boldsymbol{x}_I^e, \boldsymbol{\Phi}_p$, and $\boldsymbol{\Psi}_p$. □

Notice that, regardless of the amount of injected energy, $\boldsymbol{\Psi}_p$, which depends only on the topology of the community, defines how the injected energy is distributed within the community.

Furthermore, like the energy of an island, also the energy of a community depends on dangling pages. However, in the case of communities, the equation must take into account the energy, $E_I^{in}$, that comes from outside and the energy, $E_I^{out}$, that is spread outside.

THEOREM 4.2. *Given a community $\boldsymbol{G}_I$, let $f_p$ be the fraction of the hyperlinks of page p that point to pages in $\boldsymbol{G}_I$ with respect to the total number of hyperlinks outgoing from p. Let $E_I^{in}$, $E_I^{out}$, and $E_I^{dp}$ be defined by*

$$E_I^{in} = \frac{d}{1-d}\sum_{i\in\mathrm{in}(I)} f_i x_i^*,\ E_I^{out} = \frac{d}{1-d}\sum_{i\in\mathrm{out}(I)}(1-f_i)x_i^*,\ E_I^{dp} = \frac{d}{1-d}\sum_{i\in dp(I)} x_i^*.$$

*Then, PageRank $\boldsymbol{x}_I^*$ of $\boldsymbol{G}_I$ satisfies*

$$E_I = |I| - E_I^{dp} + E_I^{in} - E_I^{out}. \tag{23}$$

PROOF. Without loss of generality, let us assume that pages in $P$ are ordered such that we can write consistently $\boldsymbol{x}^* = [\boldsymbol{x}_{\mathrm{out}(I)}^{*'}, \boldsymbol{x}_{dp(I)}^{*'}, \boldsymbol{x}_{o(I)}^{*'}, \boldsymbol{x}_{P\setminus I}^{*'}]'$, with $o(I) = I \setminus (\mathrm{out}(I) \bigcup \mathrm{dp}(I))$. First, we consider the following system

$$\boldsymbol{Y} = d\boldsymbol{Q}\boldsymbol{Y} + \boldsymbol{U}, \tag{24}$$

with

$$\begin{aligned}
\boldsymbol{Q} &= \begin{pmatrix} \boldsymbol{W}_{I,\mathrm{out}(I)} & \boldsymbol{0} & \boldsymbol{W}_{I,o(I)} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{F} & \boldsymbol{0} & \boldsymbol{0} & 1 & 0 \\ \boldsymbol{0} & \mathbb{1}'_{|dp(I)|} & \boldsymbol{0} & 0 & 1 \end{pmatrix}, \\
\boldsymbol{F} &= [1-f_1, \ldots, 1-f_{|\mathrm{out}(I)|}], \\
\boldsymbol{U} &= v[\boldsymbol{U}'_I, 0, (1-d)\big(N - E_I^{in} + E_I^{out}\big)]', \\
\boldsymbol{U}_I &= d\boldsymbol{W}_{I,\mathrm{in}(I)}\boldsymbol{x}^*_{\mathrm{in}(I)} + (1-d)\mathbb{1}_{|I|}.
\end{aligned}$$

If $\boldsymbol{x}^*$ is a solution of (2), then

$$\boldsymbol{Y}^* = \big[\boldsymbol{x}^*_I, E_I^{out}, E_I^{dp} + \big(N - E_I^{in} + E_I^{out}\big)\big] \tag{25}$$

is a solution of (24). In fact, by definition, $E_I^{out} = d/(1-d)\boldsymbol{F}\boldsymbol{x}^*_{out}$ and $E_I^{dp} = d/(1-d)\mathbb{1}'_{|dp(I)|}\boldsymbol{x}^*_{dp(I)}$, so that

$$\begin{aligned}
\boldsymbol{Y}_I &= d\boldsymbol{Q}_{I,I}\boldsymbol{Y}^*_I + \boldsymbol{U}_I \\
&= d\boldsymbol{W}_{I,\mathrm{out}(I)}\boldsymbol{x}^*_{\mathrm{out}(I)} + d\boldsymbol{W}_{I,o(I)}\boldsymbol{x}^*_{o(I)} + d\boldsymbol{W}_{I,\mathrm{in}(I)}\boldsymbol{x}^*_{\mathrm{in}(I)} + (1-d)\mathbb{1}_{|I|} = \boldsymbol{x}^*_I, \\
y^*_{|I|+1} &= d\boldsymbol{F}\boldsymbol{x}^*_{\mathrm{out}(I)} + [d^2/(1-d)]\boldsymbol{F}\boldsymbol{x}^*_{\mathrm{out}(I)} = [d/(1-d)]\boldsymbol{F}\boldsymbol{x}^*_{\mathrm{out}(I)} = E_I^{out}, \\
y^*_{|I|+2} &= d\,\mathbb{1}'_{|dp(I)|}\boldsymbol{x}^*_{dp(I)} + [d^2/(1-d)]\mathbb{1}'_{|dp(I)|}\boldsymbol{x}^*_{dp(I)} + d\big(N - E_I^{in} + E_I^{out}\big) \\
&\quad + (1-d)\big(N - E_I^{in} + E_I^{out}\big) = E_I^{dp} + \big(N - E_I^{in} + E_I^{out}\big)
\end{aligned}$$

follows. Moreover, $\boldsymbol{Q}$ is a stochastic matrix. In fact, let $\boldsymbol{Q}_p$ and $\boldsymbol{W}_p$ be the $p$th columns of $\boldsymbol{Q}$ and $\boldsymbol{W}$, respectively. The equality $\|\boldsymbol{Q}_p\|_1 = 1$ must be proved for $p \in dp(I)$, $p \in o(I)$, and $p \in \mathrm{out}(I)$. If $p \in dp(I)$, then $\|\boldsymbol{Q}_p\|_1 = \|[0, \ldots, 0, 1]\|_1 = 1$ by definition of $\boldsymbol{Q}_p$. Assume $p \in o(I)$. Then, $\|\boldsymbol{W}_p\|_1 = 1$ holds since $p$ is not a dangling page, $\boldsymbol{W}_p = [0, \ldots, 0, \boldsymbol{Q}'_p, 0, \ldots, 0]'$ holds by definition, and $\|\boldsymbol{Q}_p\|_1 = 1$ is an immediate consequence. On the other hand, let $p \in \mathrm{out}(I)$. Then $\|\boldsymbol{Q}_p\|_1 = \sum_{q \in ch[p] \cap I} 1/h_p + 1 - f_p = 1$.

Since $\boldsymbol{Q}$ is a stochastic matrix, the solution of (24) fulfills $\|\boldsymbol{Y}^*\|_1 = d\,\|\boldsymbol{Q}\|_1\|\boldsymbol{Y}^*\|_1 + \|\boldsymbol{U}\|_1$ and, as a consequence,

$$\|\boldsymbol{Y}^*\|_1 = \|\boldsymbol{U}\|_1/(1-d). \tag{26}$$

Combining (26) with the definition of $\boldsymbol{U}$

$$\begin{aligned}
\|\boldsymbol{Y}^*\|_1 &= \frac{1}{1-d}\|d\boldsymbol{W}_{I,\mathrm{in}(I)}\boldsymbol{x}^*_{\mathrm{in}(I)} + (1-d)\mathbb{1}_{|I|}\|_1 + (1-d)\big(N - E_I^{in} + E_I^{out}\big) \\
&= E_I^{in} + |I| + \big(N - E_I^{in} + E_I^{out}\big).
\end{aligned} \tag{27}$$

Finally, by the definition of $\boldsymbol{Y}^*$,

$$\|\boldsymbol{Y}^*\|_1 = \|\boldsymbol{x}^*_I\|_1 + E_I^{out} + E_I^{dp} + \big(N - E_I^{in} + E_I^{out}\big), \tag{28}$$

and, matching Eqs. (27) and (28),

$$E_I^{in} + |I| + \big(N - E_I^{in} + E_I^{out}\big) = \|\boldsymbol{x}^*_I\|_1 + E_I^{out} + E_I^{dp} + \big(N - E_I^{in} + E_I^{out}\big) \tag{29}$$

which, in turn, yields

$$\|\boldsymbol{x}^*_I\|_1 = |I| + E_I^{in} - E_I^{out} - E_I^{dp}\,. \quad \square$$

The following corollary gives some details about $E_I^{dp}$, which represents the energy lost by dangling pages.

COROLLARY 4.1. *Given a community $\boldsymbol{G}_I$, let $g_p$ be the fraction of the hyperlinks of page $p$ that point to pages in $dp(I)$ with respect to the total number of hyperlinks outgoing from $p$ and let* ps$[I]$ *be the set of pages with at least a hyperlink to a dangling page. Then*

$$E_I^{dp} = d\,|dp(I)| + \frac{d^2}{1-d} \sum_{i \in \mathrm{ps}[I]} g_i x_i^*$$

PROOF. Let us rewrite $E_I^{dp}$ with respect to the PageRanks of the pages in ps$[I]$ as follows.

$$\begin{aligned} E_I^{dp} &= \frac{d}{1-d} \sum_{i \in dp(I)} x_i^* = \frac{d}{1-d} \sum_{i \in dp(I)} \left( \sum_{j \in pa[i]} d \frac{x_j^*}{h_j} + 1 - d \right) \\ &= d\,|dp(I)| + \frac{d^2}{1-d} \sum_{j \in \mathrm{ps}[I]} \sum_{i \in ch[j] \cap I} \frac{x_j^*}{h_j} = d\,|dp(I)| + \frac{d^2}{1-d} \sum_{j \in \mathrm{ps}[I]} \frac{s_j x_j^*}{h_j}, \end{aligned}$$

where $s_j$ is the number of hyperlinks of page $j$ which point to a dangling page. Hence, the thesis follows straightforwardly by observing that $g_j = \frac{s_j}{h_j}$. □

Due to the decomposition property, the energy that comes from outside $E_I^{in}$ can be analyzed in more details, page by page, separating the contribution $c_I^p$, of each page $p \in \mathrm{in}(I) \cup I$, into $e_I^p$, $s_I^p$, and $o_I^p$, which describe the energy which is entrapped in the community, lost in dangling pages, and spread onto the Web, respectively.

THEOREM 4.3. *Given a community $\boldsymbol{G}_I$, let $\psi_{i,p}$ be the ith component of $\boldsymbol{\Psi}_p$, that is, the element in position $(i, p)$ in the matrix $d(\boldsymbol{I} - d\boldsymbol{W}_{I,I})^{-1}\boldsymbol{W}_{I,\mathrm{in}(I)}$, and let*

$$c_I^p = \frac{d}{1-d} f_p x_p^*, \qquad o_I^p = \frac{d}{1-d} \left( \sum_{i \in \mathrm{out}(I)} (1 - f_i) h_{i,p}, \right) x_p^*,$$

$$s_I^p = \frac{d}{1-d} \left( \sum_{i \in dp(I)} h_{i,p} \right) x_p^*, \quad e_I^p = \left( \sum_{i \in I} h_{i,p} \right) x_p^*.$$

*Then, $e_I^p = c_I^p - o_I^p - s_I^p$ and $E_I^{in} = \sum_{p \in \mathrm{in}(I)} c_I^p$ hold.*

PROOF. The proof that $e_I^p = c_I^p - o_I^p - s_I^p$ holds can be carried out by the same analysis as Theorem 4.2, where $E_I^{in}$, $E_I^{out}$, $E_I^{dp}$, and $\|\boldsymbol{x}_I\|_1$ are replaced by $c_I^p, o_I^p, s_I^p$, and $e_I^p$, respectively. Moreover, $E_I^{in} = \sum_{p \in \mathrm{in}(I)} c_I^p$ follows straightforwardly from the definition of $E_I^{in}$. □

The quantities $c_I^p$, $e_I^p$, $s_I^p$, $o_I^p$ cannot be computed exactly, unless the whole matrix $\boldsymbol{W}$ is known, since the PageRank of $p$ may depend on the whole Web. On the other hand, one can compute $c_I^p/x_p^*$, $e_I^p/x_p^*$, $s_I^p/x_p^*$ that depend only on the connectivity of $\boldsymbol{G}_I$ and $p$. Those quantities provide an estimation on how the community spends the energy given by page $p$.

Theorem 4.3 can be also extended to the pages of the community. When $p \in I$, the default energy of $p$ must be considered, such that $c_I^p = 1$ and $e_I^p, s_I^p, o_I^p$ denote the parts of the default energy which are entrapped in the community, lost in dangling pages, and spread onto the Web, respectively.

Moreover, notice that $c_I^p/e_I^p$ represents the portion of the energy, provided by page $p$, which is entrapped in the community. Theorem 4.3 clarifies that if a community has no dangling page and no external hyperlink, then it entraps all the input energy, that is, $c_I^p/e_I^p = 1$. In order to study how the connectivity of a community affects the entrapped energy in more general cases, we introduce the following two lemmas.

LEMMA 4.1. *Let $l$ be a positive integer, $p \in \mathrm{in}(I)$, $q \in I$ be pages, and $\psi_{q,p}^l$ be the element in position $(q, p)$ in matrix*

$$d(d\boldsymbol{W}_{I,I})^{l-1}\boldsymbol{W}_{I,\mathrm{in}(I)} .$$

*Moreover, suppose that $\Theta_l$ is the subset of $I$ containing the pages that can be reached from $p$ using a path that consists of exactly $l$ arcs and contains only pages in $I$. Then,*

$$\sum_{r\in\Theta_l} \psi_{r,p}^l = d^l f_p , \tag{30}$$

*provided that $1 \le l \le L_p$, where $L_p(I) = \max\{n|\ \Theta_n \cap (dp(I) \cup \mathrm{out}(I)) = \emptyset\}$.*

PROOF. The proof is carried out by induction on $l$.
**Base**: $l = 1$.
The thesis follows by the definition of $\psi_{q,p}^l$, since

$$\sum_{r\in\Theta_1} \psi_{r,p}^1 = d \sum_{r\in\Theta_1} \frac{1}{h_p} = df_p .$$

**Induction**: Let $1 < l < L_p(I)$ and assume, by induction on $l$, that $\sum_{r\in\Theta_l} \psi_{r,p}^l = d^l f_p$. Since $d(d\boldsymbol{W}_{I,I})^{l+1}\boldsymbol{W}_{I,\mathrm{in}(I)} = (d\boldsymbol{W}_{I,I})[d(d\boldsymbol{W}_{I,I})^l\boldsymbol{W}_{I,\mathrm{in}(I)}]$, then

$$\psi_{r,p}^{l+1} = d \sum_{q\in pa[r]\cap I} \frac{\psi_{q,p}^l}{h_q}. \tag{31}$$

Moreover,

$$\sum_{q\in ch[r]} \frac{1}{h_q} = 1 \tag{32}$$

holds for any $r \in \Theta_l$ since, by hypothesis, $\Theta_l$ does not contain dangling pages or pages pointing outside $I$. Finally, using Eqs. (31) and (32),

$$\sum_{r\in\Theta_{l+1}} \psi_{r,p}^{l+1} = d \sum_{r\in\Theta_{l+1}} \sum_{q\in pa[r]\cap I} \frac{\psi_{q,p}^l}{h_q} = d \sum_{q\in\Theta_l} \sum_{r\in ch[q]} \frac{\psi_{q,p}^l}{h_q} = d \sum_{q\in\Theta_l} \psi_{q,p}^l = d^{l+1} f_p$$

follows. □

LEMMA 4.2. *Let $\psi^l_{q,p}$, $L_p(I)$ be defined as in Lemma* 4.1 *and let $\psi_{q,p}$ be the qth element of $\mathbf{\Psi}_p$. Then,*

$$\psi_{r,p} = \sum_{k=1}^{\infty} \psi^k_{r,p} \tag{33}$$

*and*

$$\sum_{r\in I} \psi_{r,p} \geq \frac{1-d^{L_p}}{1-d} df_p\,. \tag{34}$$

PROOF. Equation (33) follows from the definition of $\psi_{q,p}$ and $\psi^l_{q,p}$ and by

$$\begin{aligned} d(\boldsymbol{I} - d\boldsymbol{W}_{I,I})^{-1}\boldsymbol{W}_{I,\text{in}(I)} &= d\left(\sum_{k=1}^{\infty}(d\boldsymbol{W}_{I,I})^{k-1}\right)\boldsymbol{W}_{I,\text{in}(I)} \\ &= \sum_{k=1}^{\infty} d(d\boldsymbol{W}_{I,I})^{k-1}\boldsymbol{W}_{I,\text{in}(I)}\,. \end{aligned}$$

In order to prove Eq. (34), notice that

$$\sum_{r\in I}\psi_{r,p} = \sum_{r\in I}\sum_{k=1}^{\infty}\psi^k_{r,p} = \sum_{k=1}^{\infty}\sum_{r\in I}\psi^k_{r,p} \geq \sum_{k=1}^{\infty}\sum_{r\in\Theta_k}\psi^k_{r,p} \geq \sum_{k=1}^{L_p(I)}\sum_{r\in\Theta_k}\psi^k_{r,p}\,.$$

Thus, by Lemma 4.1,

$$\sum_{r\in I}\psi_{r,p} \geq \sum_{k=1}^{L_p(I)} d^k f_p = \frac{1-d^{L_p}}{1-d} df_p\,. \quad \square$$

In the following, we prove that the entrapped energy $e^p_I$ is proportional to the input energy $c^p_I$. The factor of proportionality depends upon the length of the paths from $p$ to the pages in $dp(I) \cup \text{out}(I)$. As the input pages become far from the dangling pages and the output pages, the entrapped energy becomes larger and larger, approaching $c^p_I$.

THEOREM 4.4. *Let $L_p(I)$ be defined as in Lemma* 4.1 *and let $L(I) = \max_{p\in\text{in}(I)} L_p(I)$. Then,*

$$e^p_I \geq (1-d^{L_p}(I))c^p_I\,, \tag{35}$$

$$E_I \geq (1-d^{L(I)})E^{in}_I\,. \tag{36}$$

PROOF. Equation (35) follows from the definition of $e^p_I$, $c^p_I$ and from Lemma 4.2:

$$e^p_I = \left(\sum_{r\in I}\psi_{r,p}\right)x^*_p \geq \left(1-d^{L_p(I)}\right)\frac{d}{1-d}f_p x^*_p = \left(1-d^{L_p(I)}\right)c^p_I.$$

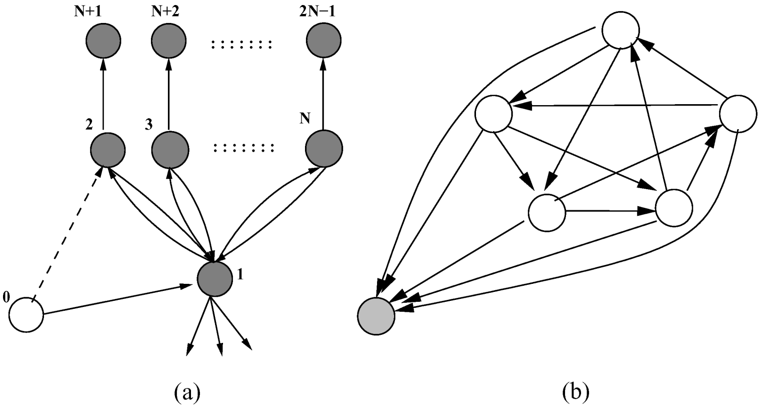

Fig. 6. Examples of islands extended by dangling pages.

Moreover, by Theorem 4.1,

$$E_I = \|\boldsymbol{x}_I^*\|_1 \geq \left\| \sum_{p \in \text{in}(I)} \boldsymbol{\Psi}_p x_p^* \right\|_1 ,$$

which implies

$$\begin{aligned} E_I &= \sum_{p \in \text{in}(I)} \sum_{r \in I} \psi_{r,p} x_p^* = \sum_{p \in \text{in}(I)} e_I^p \geq \sum_{p \in \text{in}(I)} \left(1 - d^{L_p(I)}\right) c_p \\ &\geq \sum_{p \in \text{in}(I)} \left(1 - d^{L(I)}\right) c_p = \left(1 - d^{L(I)}\right) E_I^{in} . \quad \square \end{aligned}$$

More generally, the entrapped energy depends not only on $L_p(I)$ but also on the proportion between the number of paths that connect $p$ to output and dangling pages with respect to the paths that connect $p$ to the other pages of the community. This fact can be easily explained considering the stochastic interpretation of Section 3. Random surfers remain entrapped in a community having a large number of internal paths, since they have a small chance to select hyperlinks to the outside.

Notice that, some Web designers have already exploited this fact in practice, even if the theoretical motivations were probably still unknown. In fact, they have built communities with a large number of internal hyperlinks in order to improve the PageRank.

*Example* 4.1. Let us consider a regular graph of degree $k$ with $N$ pages, where each page has been extended by a hyperlink to a dangling page (see Figure 6(b)).[7] Using straightforward algebra, we find that the PageRank of the dangling page is

$$1 - d + \frac{Nd(1-d)}{k - d(k-1)}$$

[7] In practice, such an island could be a small personal page and the dangling page might be a text document.

and the energy loss is

$$E^{dp} = d + \frac{Nd^2}{k - d(k-1)}.$$

Thus, $E^{dp}$ increases when the connectivity $k$ becomes smaller. Such a behavior is sketched by with rule (2) (see Section 1.2, item 2). Moreover, $E^{dp}$ increases also when $d$ approaches 1, which exemplifies the property discussed in Section 2.3. Notice that the energy loss can be a considerable part of the available energy $N+1$. In the limit case $d = 1$, all the energy is lost, since $E^{dp} = N + 1$. On the other hand, if $k = 0$, $E^{dp} = d(N+1)$, which is most of the available energy for $d = 0.85$.

*Example* 4.2. Let us consider the example depicted in Figure 6(a), which reminds us the structure of a Web directory where there is also a node "0" which injects energy inside the community pointing to the root. Using Theorem 4.3 and straightforward algebra, we derive

$$\psi_{i,0} = \begin{cases} \dfrac{2d(N-1+o)}{(N-1)(2-d^2)+2o} & \text{if } i = 1, \\[2ex] \dfrac{2d^2}{(N-1)(2-d^2)+2o} & \text{if } 2 \le i \le N, \\[2ex] \dfrac{2d^3}{(N-1)(2-d^2)+2o} & \text{if } i > N, \end{cases}$$

such that

$$\begin{aligned} \frac{e_I^0}{x_0^*} &= 2d\frac{(N-1)(1+d+d^2)+o}{(N-1)(2-d^2)+2o}, \\ \frac{o_I^0}{x_0^*} &= \frac{2d^2 o}{(1-d)[(N-1)(2-d^2)+2o]}, \\ \frac{s_I^0}{x_0^*} &= \frac{2d^4(N-1)}{(1-d)[(N-1)(2-d^2)+2o]}. \end{aligned}$$

The above equations suggest some interesting remarks, which illustrate how Theorem 4.3 can be exploited in practice. For example, when $N$ increases, the energy lost in dangling pages $s_I^0$ (and provided by 0) becomes larger and larger. However, such a lost is completely balanced by having less energy spread to the Web. In fact, the entrapped energy $e_I^0$ is always an increasing function of $N$.

One may wonder whether it is preferable that 0 points to the root of the directory (see previous analysis) or to another page, (e.g., page 2). Let us denote by $\bar{e}_I^0$ the entrapped energy when 0 points to node 2 instead of 0. We have

$$\frac{\bar{e}_I^0}{x_0^*} = d^2\frac{(N-1)(1+d+d^2)+o}{(N-1)(2-d^2)+2o} + d(1+d).$$

Then, it can be easily shown that the inequality $e_I^0 \ge \bar{e}_I^0$ always holds. As a result, the use of the incoming energy is more efficient when 0 points to the root.

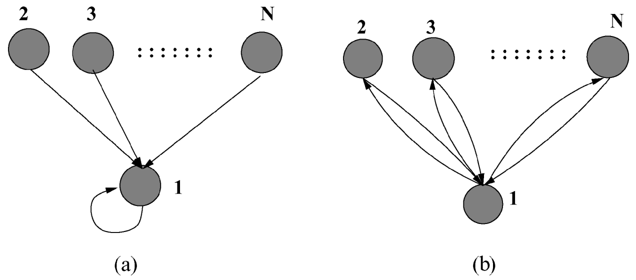

Fig. 7. The islands which yield the maximum PageRank onto a target page in the case in which self–hyperlinks are (a) and are not (b) considered.

Finally, let us consider the island in which we remove node 0. In this case

$$E^{dp} = (N-1)d\left[1 + \frac{d^2}{(N-1)(2-d^2)} + \frac{d}{2-d^2}\right]$$

is the energy loss. Notice that if $d = 0.85$, $E^{dp} > 1.4(N-1)$ and, as a consequence, $\|\boldsymbol{x}^*\|_1 < N$. Thus, even if this island contains more pages than the island of Figure 7(b), its energy is smaller. Moreover, notice that when $d \approx 1$, we observe the same behavior of the previous example and $E^{dp} \approx 2N-1$.

## 5. PAGERANK DISTRIBUTION AND COMMUNITY INTERACTION

The examples given in the previous section provides a clear picture on the effect of dangling pages and external connections to the energy of communities. This provides evidence of the rules described in Section 1.2 concerning the promotion of Web pages on the base of the topological properties of communities. In this section, we discuss some issues about distribution of PageRank and the way communities can interact with each other. In particular, we show the effect of external communities that can be purposely designed to promote the energy of a given target community.

### 5.1 Master/Slave Islands

Given a fixed number of nodes $N$, we want to find the island that contains exactly $N$ pages and has a page with the highest PageRank. Intuitively, in order to increase $x_p$, $p$ must be pointed by many other pages, and the pages that point to $p$ should have few outgoing links. Moreover, dangling pages must be avoided to limit the loss of energy. Formally, the question is different depending on whether or not self-hyperlinks, that is, hyperlinks pointing to the same page where they are originated, are admissible. Figure 7 shows the two islands with the largest PageRanks in the case of presence (absence) of self-hyperlinks. This property is stated formally as follows:

THEOREM 5.1. *Let $x_1^{(a)}$ be the PageRank of node 1 in Figure 7(a) and $x_1^{(b)}$ be the PageRank of node 1 in Figure 7(b). Then,*

$$x_1^{(a)} = 1 + (N-1)d, \tag{37}$$

$$x_1^{(b)} = \frac{d(N-1)+1}{d+1}. \tag{38}$$

*Moreover, the islands of Figure* 7(a) *(Figure* 7(b)) *yield the maximum PageRank that can be accumulated onto a given target page in the case in which self-hyperlinks are considered (are not considered, respectively).*

Proof. Let us consider the graph of Figure 7(a), that is, the case in which there are self-hyperlinks. Nodes from 2 to $N$ get the minimal PageRank, that is, $x_i^{(a)} = 1-d,\ \forall i,\ 2 \le i \le N$. From Corollary 2.1, $\sum_{i=1}^N x_i^{(a)} = N$ holds. Hence, $x_1^{(a)} = N - \sum_{i=2}^N x_i^{(a)} = 1 + (N-1)d$, which is the maximum PageRank.

Now, let us consider the graph of Figure 7(b). From Eq. (1), for each $i$, $2 \le i \le N$,

$$x_1^{(b)} = (N-1)dx_i^{(b)} + 1 - d, \tag{39}$$

$$x_i^{(b)} = \frac{dx_1^{(b)}}{N-1} + 1 - d, \tag{40}$$

from which we derive (38).

Now we prove that there are no different patterns of connectivity with higher PageRank than islands of Figure 7. Let $\boldsymbol{G}_W$ be any graph with $N$ pages and no self-hyperlinks,[8] and let $\boldsymbol{x}^*$ be the PageRank computed on $\boldsymbol{G}_W$. Without loss of generality, we assume that the largest component of $\boldsymbol{x}^*$ is $x_1^*$.

Let us split $\boldsymbol{x}^*$ and the transition matrix $\boldsymbol{W}$, associated with $\boldsymbol{G}_W$, into blocks

$$\boldsymbol{x}^* = [x_1^*, \boldsymbol{Z}']',$$
$$\boldsymbol{W} = \begin{pmatrix} 0 & \boldsymbol{R} \\ \boldsymbol{B} & \boldsymbol{G} \end{pmatrix},$$

where $\boldsymbol{Z} \in I\!R^{N-1}$, $\boldsymbol{R} \in I\!R^{1,N-1}$, $\boldsymbol{B} \in I\!R^{N-1,1}$, and $\boldsymbol{G} \in I\!R^{N-1,N-1}$, and define $\alpha$, $\beta$, $\gamma$ as follows:

$$\alpha = \frac{\boldsymbol{R}\boldsymbol{x}^*}{\mathbb{1}'_{N-1}\boldsymbol{x}^*},$$
$$\beta = \frac{\mathbb{1}'_{N-1}\boldsymbol{G}\boldsymbol{x}^*}{\mathbb{1}'_{N-1}\boldsymbol{x}^*},$$
$$\gamma = \mathbb{1}'_{N-1}\boldsymbol{B}.$$

It is easy to verify that $\alpha, \beta, \gamma$ must satisfy the following constraints:

$$\alpha, \beta, \gamma \ge 0,\ \ \alpha + \beta \le 1,\ \ \gamma \le 1. \tag{41}$$

With this notation, Eq. (2) becomes

$$\begin{cases} x_1^* = d\boldsymbol{R}\boldsymbol{Z} + 1 - d \\ \boldsymbol{Z} = d\boldsymbol{B}x_i^* + d\boldsymbol{G}\boldsymbol{Z} + (1-d)\mathbb{1}_{N-1} \end{cases},$$

and multiplying the latter equation by $\mathbb{1}'_{N-1}$,

$$\begin{cases} x_1^* = d\alpha\mathbb{1}'_{N-1}\boldsymbol{Z} + 1 - d \\ \mathbb{1}'_{N-1}\boldsymbol{Z} = d\gamma x_1^* + d\beta\mathbb{1}'_{N-1}\boldsymbol{Z} + (1-d)(N-1) \end{cases},$$

[8] For the sake of simplicity, in the following we use $x_1$ instead of $x_1^{(a)}$.

which yields

$$x_1^* = (1-d)\frac{1-d\beta+d\alpha(N-1)}{1-d\beta-d^2\alpha\gamma}. \tag{42}$$

Therefore, $x_1^*$ can be written as a function of $\alpha$, $\beta$, $\gamma$. We are interested in finding the maximum of $x_1^*$. Without loss of generality we assume $\gamma = 1$, since $x_1^*$ is a monotonically nondecreasing function of $\gamma$. Moreover, notice that $x_1^*$ does not admit a maximum in the interior of the polytope defined by (41), since

$$\frac{\partial x_1^*}{\partial \beta} = (1-d)\frac{d^2(N-1-d\gamma)\alpha}{(1-d\beta-d^2\alpha)^2}$$

is null only when $\alpha = 0$. Thus, $x_1^*$ reaches the maximum either when $\beta = 0$, or $\alpha = 0$, or $\alpha + \beta = 1$. However, if $\alpha = 0$, then $x_1^* = 1 - d$, which is not maximal. If $\beta = 0$, then $x_1^* = (1-d)[1+d\alpha(N-1)]/(1-d^2\alpha)$, which reaches the maximum for $\alpha = 1$. If $\alpha + \beta = 1$, by replacing $\beta$ by $1-\alpha$ in (42), it follows that $x_1^* = (1-d)(1-d+dN\alpha)/(1-d+d(1-d)\alpha)$, which is maximal when $\alpha = 1$. Summing up, $x_1^*$ is maximal when $\alpha = 1$ and $\beta = 0$ and we have $x_1^* = [1+d(N-1)]/(1+d)$.[9] □

Of course, the maximum score which can be transferred to a target page depends on $d$ and takes the supremum ($x_1^{(a)} \to N,\ x_1^{(b)} = N/2$) as $d \to 1$. Notice that when self-hyperlinks are not considered, the island depicted in Figure 7(b) is not the only one having a page with the largest PageRank. In fact, one can remove any proper subset of the hyperlinks of page 1, without changing $x_1^{(b)}$. Intuitively, this follows from the fact that removing the hyperlink $(1, i)$, $x_i^{(b)}$ decreases, whereas the PageRanks of the pages that are still referenced by 1 increase of the same amount. Therefore, the PageRank $x_i^{(b)}$ remains unchanged.

### 5.2 Linear Growth of PageRank

A promoting technique based on the results given in Theorem 5.1 can be easily detected. Instead, in the following, we will prove that an analogous increase in the energy of a target community can be obtained by any promoting community, regardless of its connectivity pattern, with the only constraint that all its nodes have outlinks to the target community.

THEOREM 5.2. *Let us consider two communities C and D such that every node of C is connected to at least one of D, and let F be defined as* $F = \max_{p\in C} h_p$. *Then,*

$$E_D \geq \frac{d(1-d)}{F}|C|. \tag{43}$$

PROOF. By Theorem 4.4, we get

$$E_D \geq (1-d^{L(I)})E_D^{in}.$$

[9] In fact, it is easily seen that in the case of Figure 7(b), we have $\alpha = 1$, $\beta = 0$ and $\gamma = 0$, since $\boldsymbol{G} = \boldsymbol{0}$, and $\boldsymbol{R}' = \boldsymbol{B} = \mathbb{1}_N$.

Since all the paths from $p$ to a page in $I$ contains at least an arc, then $L(I) \geq 1$, and it follows that

$$E_D \geq (1-d)E_D^{in}\,. \tag{44}$$

Moreover, using $x_I^* \geq 1-d$ and by straightforward algebra,

$$E_D^{in} = \sum_{p\in \mathrm{in}(I)} \frac{d}{1-d} f_p x_p^* \geq \sum_{p\in \mathrm{in}(I)} d f_p \geq d \sum_{p\in C} \frac{1}{F} = \frac{d}{F}|C|\,. \tag{45}$$

The thesis follows putting together Eqs. (44) and (45). □

The above results provide only a static picture of the effects that a page can produce onto a community, since they analyze the energy flows under the assumption that the Web remains fixed. On the other hand, even a simple change to the Web connectivity, like the introduction or the removal of a hyperlink, may cause a modification of all the PageRanks $\boldsymbol{x}^*$. When a community is altered, the PageRank $\boldsymbol{x}_I^*$ is affected both because the community propagates the external PageRank in a different way and because the energy spread onto the Web, $E_I^{out}$, is changed, modifying recursively the external energy $E_I^{in}$. However, Web changes actually cause only a redistribution of the energy associated with the altered pages, such that the distance between the old and the new rank is bounded accordingly. The following theorem proves that this intuitive idea is correct.

THEOREM 5.3. *Suppose that $C$ is a set of pages where we change the outlinks and denote by $\tilde{\boldsymbol{x}}^*$ the PageRank after the changes were carried out. Then,*

$$\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}^*\|_1 \leq \frac{d}{1-d}\sum_{p\in C}\delta_p x_p^* \leq \frac{2d}{1-d}E_c, \tag{46}$$

*where $\delta_p \leq 2$, $\forall p$. More precisely,*

$$\delta_p = \begin{cases} \left|\dfrac{1}{\bar{h}_p} - \dfrac{1}{h_p}\right| u_p + \dfrac{n_p}{\bar{h}_p} + \dfrac{r_p}{h_p}, & \text{if } h_p \neq 0,\ \bar{h}_p \neq 0, \\ 1, & \text{if } h_p = 0 \text{ or } \bar{h}_p = 0, \end{cases}$$

*being $h_p$ and $\bar{h}_p$ the number of hyperlinks in $p$ before and after the change, respectively, $n_p$ the number of new hyperlinks, $r_p$ the number of removed hyperlinks, and $u_p$ the number of unchanged hyperlinks.*

PROOF. Let $\boldsymbol{W}_2$ be the transition matrix of the changed Web and let us define $\boldsymbol{D} = \boldsymbol{W}_2 - \boldsymbol{W}$. Then, by Eq. (2)

$$\boldsymbol{x}_2^* - \boldsymbol{x}^* = d\boldsymbol{W}_2\boldsymbol{x}_2^* - d\boldsymbol{W}\boldsymbol{x}^* = d\boldsymbol{W}_2(\boldsymbol{x}^* + \boldsymbol{x}_2^* - \boldsymbol{x}^*) - d\boldsymbol{W}\boldsymbol{x}^* = d\boldsymbol{W}_2(\boldsymbol{x}_2^* - \boldsymbol{x}^*) + d\boldsymbol{D}\boldsymbol{x}^*.$$

Considering the 1-norm,

$$\|\boldsymbol{x}_2^* - \boldsymbol{x}^*\|_1 \leq d\,\|\boldsymbol{W}_2\|_1 \|\boldsymbol{x}_2^* - \boldsymbol{x}^*\|_1 + d\,\|\boldsymbol{D}\boldsymbol{x}\|_1$$

follows, which implies

$$\|\boldsymbol{x}_2^* - \boldsymbol{x}^*\|_1 \leq \frac{d}{1-d}\|\boldsymbol{D}\boldsymbol{x}\|_1 = \frac{d}{1-d}\sum_{p=1}^{N}\boldsymbol{D}_p x_p^*,$$

where $\boldsymbol{D}_p$ is the $p$th column of $\boldsymbol{D}$.

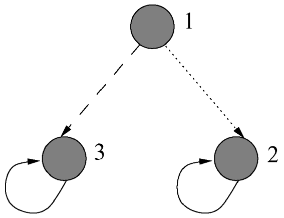

Fig. 8. A case in which the upper bound of Theorem 5.3 is exactly achieved.

By the definition of $\boldsymbol{D}$, if page $p$ is unchanged then $\boldsymbol{D}_p = 0$. Otherwise, the $i$th component $d_{i,p}$ of $\boldsymbol{D}_p$ is $1/\bar{h}_p$ if the hyperlink $(p, i)$ was introduced, $d_{i,p} = -1/h_p$ if $(p, i)$ was removed, $d_{i,p} = 1/\bar{h}_p - 1/h_p$ if $(p, i)$ was unchanged, and $d_{i,p} = 0$ if the page has never contained the hyperlink $(p, i)$. By straigthforward algebra, $\|\boldsymbol{D}_p\|_1 = \delta_p$.

Finally, assuming $r_p \geq n_p$ and $h_p \neq 0$, $\bar{h}_p \neq 0$, we get $\delta_p = 2r_p/h_p \leq 2$. A similar analysis can be carried out if $r_p < n_p$. □

Theorem 5.3 extends a similar result given in Ng et al. [2001a], where the authors prove that $\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}^*\|_1 \leq 2E_c/(1-d)$. In fact, Theorem 5.3 gives a tighter bound—it replaces the factor $2/(1-d)$ with $2d/(1-d)$—and yields a more detailed analysis—it introduces the $\delta_i$s. Moreover, Theorem 5.3 can be applied both to the PageRank defined by Eq. (2) and Eq. (13), whereas the result in Ng et al. [2001a] can be used only for the PageRank of Eq. (13).

Notice that there are cases when the equality holds in (46). For example, consider the graph of Figure 8 where the change described in Theorem 5.3 consists of removing the dotted link and of introducing the dashed link is introduced. Before the change, $\boldsymbol{x}^* = [1-d, 1, 1+d]'$ whereas, after the change, $\tilde{\boldsymbol{x}}^* = [1-d, 1+d, 1]'$, and $\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}^*\|_1 = 2d$, which is exactly the bound found in Theorem 5.3.

Theorem 5.3 highlight a nice property of PageRank, namely that that a community can only make a very limited change to the overall PageRank of the Web. Thus, regardless the way they change, nonauthoritative communities cannot affect significantly the global PageRank.[10]

The bound of Theorem 5.3 gives also indications on the frequency of PageRank updating. Assuming that every day the Web dynamics alters $\alpha$ pages and that the affected pages are random, then $2d\alpha t/(1-d)$ is an approximate measure of $\|\boldsymbol{x}_0 - \boldsymbol{x}_t\|_1$, the difference between the PageRanks computed at days 0 and $t$. Thus, for example, if we want that the error remains bounded by $\varepsilon$, then PageRank must be computed approximately every $((1-d)/2d\alpha)\,\epsilon$ days.

## 6. PAGERANK COMPUTATION

The Jacobi algorithm, described by Eq. (3), is an efficient solution, especially when compared to noniterative algorithms for solving linear systems, such as the Gaussian elimination. In fact, the Gauss method would require $\mathcal{O}(N^3)$ flops

[10] In Ng et al. [2001a], the authors analyze HITS [Kleinberg 1999] proving that Kleinberg's method does not share this robustness property.

to solve Eq. (2), which is nowadays prohibitive due to the dimension of the Web. On the other hand, each iterative computation based on Eq. (3) requires $\mathcal{O}(m \cdot |H|)$ flops, where $m$ is the number of iterations and $|H|$ is the total number of hyperlinks in the Web.[11] Using experimental arguments, in Brin et al. [1999] and Haveliwala [1999] the system (3) is proven to converge in a reasonably limited number of steps (i.e., $m = 52$ when Google's PageRank was applied to a database of 322 millions of pages). Therefore, since $m << N$ (and $|H| << N^2$), the Jacobi method is much faster than the Gaussian elimination.

The convergence rate of PageRank is also an interesting issue. Brin et al. [1999] claim that the rapid convergence rate of Google's PageRank depends on the particular connectivity of the Web-graph and that the Web should be an *expander-like graph* [Motwani and Raghavan 1995]. However, as pointed out in the following theorem, the exponential reduction of the error of Eq. (3) yields even without the hypothesis of expander-like graphs.

THEOREM 6.1. *Let* $|e_{rel}(t)|_1 = \frac{\|\boldsymbol{x}^* - \boldsymbol{x}(t)\|_1}{\|\boldsymbol{x}^*\|_1}$ *be the* 1*-norm of the relative error made in the computation of PageRank at time* $t$*. Then,*

$$|e_{rel}(t)|_1 \le d^t |e_{rel}(0)|_1.$$

*Moreover, if there is no dangling page, then there exists* $v$*, where* $v \ge 0$ *and* $v = \boldsymbol{W}v$*, such that the equality holds.*

PROOF. From Eq. (3), which has $\boldsymbol{x}^*$ as the stable equilibrium point,

$$\begin{aligned} |e_{rel}(t)|_1 &= \frac{\|\boldsymbol{x}^* - \boldsymbol{x}(t)\|_1}{\|\boldsymbol{x}^*\|_1} \\ &= \frac{\left\|d\boldsymbol{W}\boldsymbol{x}^* + (1-d)\mathbb{1}_N - d\boldsymbol{W}\boldsymbol{x}(t-1) - (1-d)\mathbb{1}_N\right\|_1}{\|\boldsymbol{x}^*\|_1} \\ &= \frac{\|d\boldsymbol{W}[\boldsymbol{x}^* - \boldsymbol{x}(t-1)]\|_1}{\|\boldsymbol{x}^*\|_1} \\ &= \frac{\|d^t\boldsymbol{W}^t[\boldsymbol{x}^* - \boldsymbol{x}(0)]\|_1}{\|\boldsymbol{x}^*\|_1} \\ &\le d^t\|\boldsymbol{W}\|_1^t \frac{\|\boldsymbol{x}^* - \boldsymbol{x}(0)\|_1}{\|\boldsymbol{x}^*\|_1} \le d^t|e_{rel}(0)|_1, \end{aligned}$$

since, from Proposition 2.3, $\|\boldsymbol{W}\|_1 \le 1$. On the other hand, if the graph has no dangling page, then $\|\boldsymbol{W}\|_1 = 1$ and $\boldsymbol{W}$ is a stochastic matrix. Therefore, according to Frobenius' Theorem, the maximal eigenvalue of $\boldsymbol{W}$ is 1. Thus, let $\boldsymbol{v} \ge 0$ be the corresponding real eigenvector, then $\boldsymbol{W}^t(\boldsymbol{x}^* - \boldsymbol{x}(0)) = \boldsymbol{W}^t\boldsymbol{v} = \boldsymbol{v}$, from which $|e_{rel}(t)|_1 = d^t|e_{rel}(0)|_1$ follows. □

*Remark* 6.1. Let us consider the computation of PageRank over all the Web. The dynamical system (3) reaches the solution in a number of steps that is logarithmic in $d$. More precisely, since $\|\boldsymbol{x}^* - \boldsymbol{x}(0)\|_1 \le N$ and $\|\boldsymbol{x}^*\|_1 \ge N(1-d)$, then $|e_{rel}(t)|_1 \le d^t/(1-d)$. Therefore, in order to gain a relative error which

[11] In fact, $\boldsymbol{Wx}$ can be calculated in $O(|H|)$ steps, due to the sparsity of $\boldsymbol{x}$.

is under a desired threshold $\varepsilon$, we must impose $d^t/(1-d) \leq \varepsilon$, from which $t \geq \log((1-d)\varepsilon)/\log d \approx \log(1/\varepsilon)$.[12]

For example, when choosing $d = 0.5$, after 50 steps we have $|e_{rel}|_1 \leq 3.5e-15$. Vice-versa, the number of steps needed to reach a precision up to 8 digits, that is, $\varepsilon = 10^{-8}$, is 28, which is congruent with the experimental results found in Brin et al. [1999].

It is worth mentioning that PageRank can be also evaluated by solving the linear system (2) with any of the available algorithms. For instance, the Gauss-Seidel method (see Golub and Van Loan [1993, pp. 506–511]) guarantees a faster convergence rate with respect to the Jacobi method, which is the algorithm represented by (3). More precisely, consider a linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$, where $\boldsymbol{A} = \{a_{i,j}\}$. By the Stein–Rosenberg Theorem [Varga 1962], the Gauss-Seidel method is always faster than the Jacobi procedure, when $a_{i,i} \neq 0$ for each $i$, $a_{i,j} \leq 0$ for each $i, j$, and $\rho(A) < 1$; this holds in our case, since $\boldsymbol{A} = \boldsymbol{I} - d\boldsymbol{W}$ and $w_{i,j} < 1$ hold. The Gauss-Seidel is a modified Jacobi method where, during each iteration, the components of $\boldsymbol{x}(t)$ are updated one at a time, instead of in parallel. Formally, we have $x_p(t) = d\sum_{q\in pa[p],\, q<p} x_p(t) + d\sum_{q\in pa[p],\, q\geq p} x_p(t-1) + 1 - d$. Notice that this procedure has also the advantage of using less memory, since it uses only a copy of $\boldsymbol{x}$, instead of the two copies (i.e., $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t-1)$) required by Eq. (3).

*Remark* 6.2. Several implementation issues may decisively support the use of the Jacobi algorithm. In fact, the efficiency of the computation of PageRank mainly depends on the data structures used to store the huge matrix $\boldsymbol{W}$. At each time step $t$, all the elements of $\boldsymbol{W}$ must be accessed in order to compute $\boldsymbol{x}(t+1)$. Since the matrix cannot fit into the main memory, but must be stored onto disks, a sequential access is the best way for reading $\boldsymbol{W}$ (see Haveliwala [1999]). However, a crawler naturally produces a sequence $Seq = c_1; c_2; \ldots; c_N$, where $c_i = p^i_1, \ldots, p^i_{|\mathrm{ch}[i]|}$ and $p^i_j$ is the $j$th page referenced by $i$. Thus, the subsequence $c_i$ is a compact representation of the $i$th column of $\boldsymbol{W}$ and $Seq$ is a data structure where $\boldsymbol{W}$ is stored by columns. Such a data structure can be used straightforwardly by the Jacobi algorithm. On the contrary, this data structure cannot be used directly by Gauss-Seidel algorithm, which needs the matrix $\boldsymbol{W}$ be stored by rows. Producing the new data structure used by Gauss-Seidel method can be very expensive from a computational point of view and definitely suggests that the Jacobi algorithm is better suited for computing PageRank.

Notice that, due to the particular *"pseudo-stochastic"* nature of matrix $\boldsymbol{I} - d\boldsymbol{W}$, an upper bound can be established for its condition number, $k(\boldsymbol{I} - d\boldsymbol{W})$. In particular, $k(\boldsymbol{I} - d\boldsymbol{W}) \leq (d+1)/(d-1)$, which is independent of the matrix dimension.[13] Such a property guarantees the solution of the linear system (2)

[12] In the special case of a graph with no dangling page, then $\|\boldsymbol{x}^*\|_1 = N$ and $t \geq \log(\varepsilon/2)/\log d \approx log(1/\varepsilon)$.

[13] By its definition, using the 2-norm, the condition number of $\boldsymbol{A}$ is $k(\boldsymbol{A}) = \lambda_{max}/\lambda_{min}$, being $\lambda_{max}$ and $\lambda_{min}$ the maximum and the minimum eigenvalue of $\boldsymbol{A}$, respectively. Moreover, again by definition, $\lambda \leq \sup_{x\neq 0} \|\boldsymbol{A}\mathbf{x}\|_p/\|\mathbf{x}\|_p$, for all eigenvalues $\lambda$ and for each $p$–norm. Therefore, in this case, $\lambda \leq \|\boldsymbol{I} - d\boldsymbol{W}\|_1 \leq \|\boldsymbol{I}\|_1 + d\|\boldsymbol{W}\|_1 \leq 1 + d$, $\forall\lambda$, and if $\mathbf{x}$ is an eigenvector, then $\lambda\|\mathbf{x}\|_1 = \|\boldsymbol{I}\mathbf{x} - d\boldsymbol{W}\mathbf{x}\|_1 \geq$

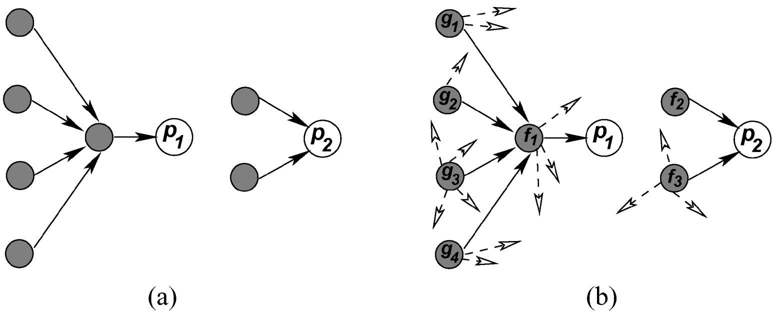

Fig. 9. (a) An example of how $d$ can effect the ordering of pages in PageRank. Page $p_1$ has many far relatives, whereas $p_2$ has a few near ancestors. (b) The effect of the outdegree in PageRank extinguishes the effect of $d$.

using optimal iterative solvers (see, e.g., Björck [1996] and Bianchini et al. [2001]) in a number of steps which is independent of the Web dimension, thus confirming the result stated by Theorem 6.1.

The proposed complexity analysis shows that the number of steps is only dependent on the precision term $\log(1/\varepsilon)$. Since PageRank contributes to the sorting of the pages returned on a given query, as a matter of fact, the precision requirements might be dictated by the constraint of keeping the ordering unchanged. However, the above strict interpretation is not meaningful in practical applications. In fact, PageRank does not model exactly the concept of "page authority" as it is conceived by users. Perturbations on the PageRank may result in different orderings of the pages, but if the perturbation is small, it is unlikely that users can perceive differences in the quality the service provided by the search engine. Moreover, we now show that the precision requirement are not a crucial issue since we prove that PageRank is strongly affected by the choice of the dumping factor $d$.

PageRank deeply depends on parameter $d$. When $d$ is small, only the nearest relatives give an actual contribution to $x_p$, whereas, as $d \to 1$, far ancestors are also significant for the evaluation of $x_p$.

*Example* 6.1. Let $p_1$ and $p_2$ be connected as in Figure 9(a). Suppose that $p_1$ has more ancestors than $p_2$, which are far from $p_1$, whereas $p_2$ has less but closer relatives. When $d \approx 1$, $x_{p_1}$ will be larger than $x_{p_2}$, but when $d \approx 0$ the converse holds. In particular,

$$\begin{aligned} x_{p_1} &= (1-d) + d(1-d) + 4d^2(1-d), \\ x_{p_2} &= (1-d) + 2d(1-d), \end{aligned}$$

and $x_{p_1} > x_{p_2}$ if $d > \dfrac{1}{4}$.

*Example* 6.2. Let us consider trees $t_p$ and $t_q$ which, by construction, have $p$ and $q$ as their roots and are constructed by following back the hyperlinks

$\|\boldsymbol{I}\mathbf{x}\|_1 - d\,\|\boldsymbol{W}\mathbf{x}\|_1$, from which $\lambda \geq 1 - d$ follows. Finally, $k(\boldsymbol{I} - d\boldsymbol{W}) \leq (d + 1)/(d - 1)$ holds.

driving to $p$ and $q$, respectively. Let the outdegree be constant for each node. If $t_p$ and $t_q$ have the same number of nodes in levels $0, 1, \ldots, i-1$, but $t_p$ has fewer nodes than $t_q$ at level $i$, and a greater number of nodes at level $i+1$, then $x_p < x_q$ as $d$ approaches 0. When the hypothesis on the outdegree is relaxed, the score of each page is deeply influenced by the importance of the hyperlinks pointing to it.

*Example* 6.3. Let us consider again the page score of $p_1$ and $p_2$ with the connectivity described in Figure 9(a), but taking into account parent nodes with distinct outdegree for both pages (see Figure 9(b)). In this case $x_{p_2} \geq x_{p_1}$ $\forall d \in [0, 1)$. In particular, we can notice that pages $f_1$ and $f_3$ give the same contribution to the PageRank of $p_1$ and $p_2$, respectively. Nevertheless, when $d \approx 1$, the PageRanks due to pages $g_i$, $i = 1, 4$, sum up to approximately 1/3, which is far away from the unitary score given by $f_2$.

## 7. CONCLUSIONS

In this article, we have presented an in-depth analysis of PageRank to disclose its fundamental properties concerning stability, complexity of the computational scheme, and the critical role of parameters involved in the computation. In particular, we have pointed out that the inherent structure of the Markovian matrices associated with the Web makes it to possible to perform an optimal computation of PageRank, a property of crucial importance for the actual scaling up to the Web. It is shown that, as a matter of fact, the effectiveness of the computational scheme also depends on the limited precision requirements imposed by the sensitivity of PageRank from the dumping factor. Some nice properties concerning PageRank robustness have been derived, which extend previous results in the literature. In addition, we have shown the technical soundness of a dynamic computational scheme taking place while changing the structure of the Web.

We have introduced the notion of energy and a circuital analysis to understand the evolution of PageRank inside Web communities. The derived energy balance equations make it possible to understand the way different Web communities interact each other and to disclose some secrets for promotion of Web pages. In particular, it is pointed out that the energy of a given target community can be driven by a "promoting community" so as to grow linearly with the number its pages. This property holds regardless of the structure of the promoting community, which makes it very hard its detection.

Finally, it worth mentioning that PageRank is only one of the parameters involved in Google's ranking of the answers to a given query. The aim of the article are limited to the investigation of PageRank and do not allow to make conclusions on the actual ranking attached by Google to the Web pages.

### ACKNOWLEDGMENTS

We thank Michelangelo Diligenti, Marco Maggini, Edmondo Trentin (University of Siena), Paolo Frasconi (University of Florence), Ah Chung Tsoi (Faculty of Computer Science, University of Wollongong), Prabhakar Raghavan (Verity, Inc.) for fruitful discussions which contributed to shed lights on some topics covered in this paper. Barbara Hammer (University of Osnabrück), Immanuel Bomze (TU University of Wien), and Monika Henzinger (Google, Inc.) read an early version of the article and provided corrections and/or thoughtful comments. Finally, we are also indebted with Nicola Baldini, Duncan Wilcox, Michele Bini, and Sandro Tolaini (focuseek.com) for having brought to our attention some crucial aspects of PageRank, that provided strong motivations for the research carried out in this article.

## REFERENCES

Bharat, K. and Henzinger, M. R. 1998. Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 104–111.

Bianchini, M., Fanelli, S., and Gori, M. 2001. Optimal algorithms for well-conditioned nonlinear systems of equations. *IEEE Trans. Comput. 50*, 7, 689–698.

Björck, A. 1996. *Numerical Methods for Least Squares Problems*. Society for Industrial and Applied Mathematics.

Bomze, I. and Gutjahr, W. 1994. The dinamics of self–evaluation. *Appl. Math. Comput. 64*, 47–63.

Bomze, I. and Gutjahr, W. 1995. Estimating qualifications in a self-evaluating group. *Qual. Quant.* 29, 241–250.

Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P. 2001. Finding authorities and hubs from link structures on the world wide web. In *Proceedings of the 10th International World Wide Web Conference*.

Brin, S., Motwani, R., Page, L., and Winograd, T. 1998. What can you do with a web in your pocket? *IEEE Bulle. Techn. Comm. Data Eng., IEEE Comput. Soc. 21*, 2, 37–47.

Brin, S. and Page, L. 1998. The anatomy of a large–scale hypertextual Web search engine. In *Proceedings of the 7th World Wide Web Conference (WWW7)*.

Brin, S., Page, L., Motwani, R., and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the Web. Tech. Rep. 1999-66, Stanford University. Available on the Internet at http://dbpubs.stanford.edu:8090/pub/1999-66.

Cohn, D. and Chang, H. 2000. Learning to probabilistically identify authoritative documents. In *Proceedings of 17th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, Calif., 167–174.

Cohn, D. and Hofmann, T. 2001. The missing link—A probabilistic model of document content and hypertext connectivity. In *Neural Inf. Proc. Syst. 13*.

Diligenti, M., Gori, M., and Maggini, M. 2002. Web page scoring systems for horizontal and vertical search. In *Proceedings of the 11th World Wide Web Conference (WWW11)*.

Golub, G. H. and Van Loan, C. F. 1993. *Matrix computation*. The Johns Hopkins University Press.

Haveliwala, T. H. 1999. Efficient computation of pagerank. Tech. Rep. 1999-66, Stanford University. Available on the Internet at http://dbpubs.stanford.edu:8090/pub/1999-66.

Haveliwala, T. H. 2002. Topic sensitive pagerank. In *Proceedings of the 11th World Wide Web Conference (WWW11)*. Available on the Internet at http://dbpubs.stanford.edu:8090/pub/2002-6.

Henzinger, M. 2001. Hyperlink analysis for the Web. *IEEE Internet Computing 5*, 1, 45–50.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *J. ACM 46*, 5, 604–632.

Lempel, R. and Moran, S. 2000. The stochatic approach for link–structure analysis (SALSA) and the TKC effect. In *Proceedings of the 9th World Wide Web Conference (WWW9)*. Elsevier Science, 387–401.

Marchiori, M. 1997. The quest for correct information on the Web: Hyper search engines. *Computer Networks and ISDN Systems 29*, 1225–1235.

Motwani, R. and Raghavan, P. 1995. *Randomized algorithms*. Cambridge University Press.

Ng, A. Y., Zheng, A. X., and Jordan, M. I. 2001a. Link analysis, eigenvectors and stability. In *Proceedings of International Conference on Research and Development in Information Retrieval (SIGIR 2001)*. ACM, New York.

Ng, A. Y., Zheng, A. X., and Jordan, M. I. 2001b. Stable algorithms for link analysis. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'2001)*.

PRINGLE, G., ALLISON, L., AND DOWE, D. L. 1998. What is tall poppy among the Web pages? *Comput. Netwo. ISDN Syst. 30*, 369–377.

RICHARDSON, M. AND DOMINGOS, P. 2002. The intellingent surfer: probabilistic combination of link and content information in pagerank. In *Advances in Neural Information Processing Systems, 14*. MIT Press, Cambridge, Mass.

RUMELHART, D., HINTON, G., AND WILLIAMS, R. 1986. Learning representations by back-propagating errors. *Nature 323*, 533–536.

SENETA, E. 1981. *Non-negative matrices and Markov chains*. Springer-Verlag, New York, Chap. 4, pp. 112–158.

VARGA, R. S. 1962. *Matrix Iterative Analysis*. Prentice–Hall, Englewood Cliffs, N.J.

ZHANG, D. AND DONG, Y. 2000. An efficient algorithm to rank web resources. In *Proceedings of the 9th International World Wide Web Conference (WWW9)*. Elsevier Science, Amsterdam, The Netherlands.

Received February 2003; revised June 2003; accepted August 2003