

Generic Damping Functions for Propagating Importance in Link-Based Ranking

Ricardo Baeza-Yates, Paolo Boldi, and Carlos Castillo

Abstract. This paper introduces a family of link-based ranking algorithms that propagate page importance through links. The algorithms include a damping function that decreases with distance, thus a direct link implies greater endorsement than a link via a longer path. PageRank is the most widely known ranking function of this family.

The main objective of this paper is to determine whether this family of ranking techniques is of some interest *per se* and how different choices for the damping function affect rank quality and convergence speed. Even though our results suggest that PageRank can be approximated with other, more simple forms of rankings that may be computed more efficiently, our focus is more speculative in nature, given that it aims at separating the kernel of PageRank, that is, link-based importance propagation, from the way propagation decays over paths.

We focus on three damping functions that have, respectively, linear, exponential, and hyperbolic decay on the lengths of the paths. The exponential decay corresponds to PageRank, and the other functions are new. The work we carry includes algorithms, analysis, comparisons, and experiments that study their behavior under different parameters in real web graph data.

Amongst other results, we show how to calculate a linear approximation that induces a page ordering that is almost identical to PageRank’s using a fixed number of iterations. Comparisons were made using Kendall’s τ on large domain datasets.

1. Introduction

While traditional information retrieval (IR) methods are used by web search engines to some extent, the web is much more extensive, more dynamic, and less coherent than traditional text collections [Arasu et al. 01]. The web is an

open medium in which everyone can publish information; this has been key to its success but, at the same time, acts as a major source of problems for information retrieval researchers.

Fortunately, the web provides an extra source of information that is not present in traditional text repositories: there are hyperlinks among pages, and these hyperlinks convey information—they are not placed at random. For instance, a pair of pages linked together would be much more likely to belong to the same topic than two pages taken at random [Davison 00].

1.1. Link Analysis

In the web, we can identify three levels of link analysis:

- The *microscopic level* of link analysis is related to the statistical properties of links of individual nodes.
- The *macroscopic level* of link analysis is related to the structure of the web at large.
- The *mesoscopic level* of link analysis is related to the properties of areas or regions of the web.

The macroscopic level of description of the web started with a seminal paper by Broder et al., in which a global structure was described based on the presence of a large strongly connected component [Broder et al. 00]. This is called the *bow-tie* structure of the web, presented in Figure 1. Further refinements of this model identified areas inside the CORE component [Donato et al. 05, Baeza-Yates et al. 04].

A related macroscopic description is the *Jellyfish structure* (described in [Tauro et al. 01]) for autonomous systems in the Internet topology. According to this

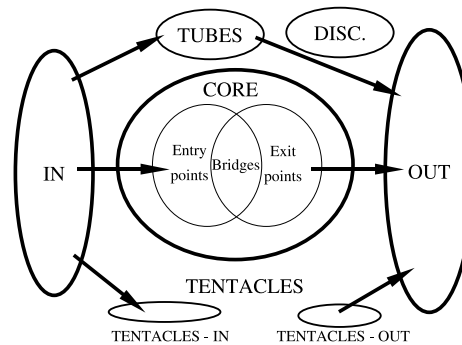


Figure 1. Schematic depiction of the macroscopic “bow-tie” structure of the web [Broder et al. 00].

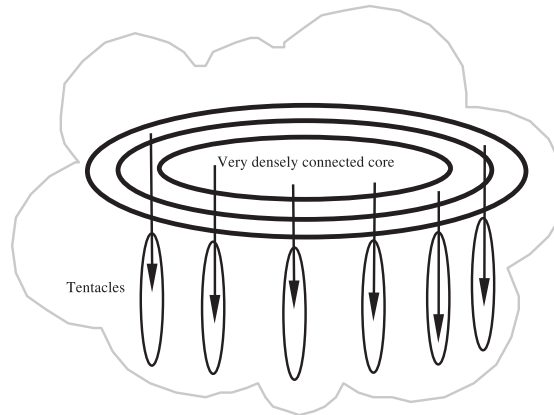


Figure 2. Schematic depiction of the macroscopic “jellyfish” structure of the Internet [Tauro et al. 01].

view, depicted in Figure 2, we can identify a core portion, surrounded by areas of decreasing link density and with many nodes forming long, loosely-connected chains or *tentacles*.

The microscopic level of description on the web has been done by several authors (e.g., [Huberman 01, Barabási 02]) and is based on the observation that the distribution of degrees on the web is very skewed, not showing the typical Poisson distribution observed in classical random graphs [Erdős and Rényi 60]. In scale-free networks, such as the web, the distribution of the number of links of a page p follows a power law:

$$Pr(\text{page } p \text{ has } k \text{ links}) \propto k^{-\theta}. \quad (1.1)$$

Scale-free networks have a few highly-connected links that act as “hubs” connecting many other nodes to the network. The connectivity of scale-free networks is resistant to random removal of edges [Callaway et al. 00] and can be explained in part by a “preferential attachment” process [Barabási and Albert 99], also called a *rich-get-richer* phenomenon or Yule process.

Mesoscopic link analysis is related to the properties of the neighborhood of a node, the context in which most of the link-based ranking functions work. A way of describing the neighborhood of a node is known as the “hop-plot”: a plot of the number of different neighbors at different distances, such as the one depicted in Figure 3.

The class of functional rankings that we describe in this paper, including PageRank, belong to this level of analysis, given that most of the ranking of a node comes from its short-range connections. This will be clearer later on in this article, in particular in Section 4.

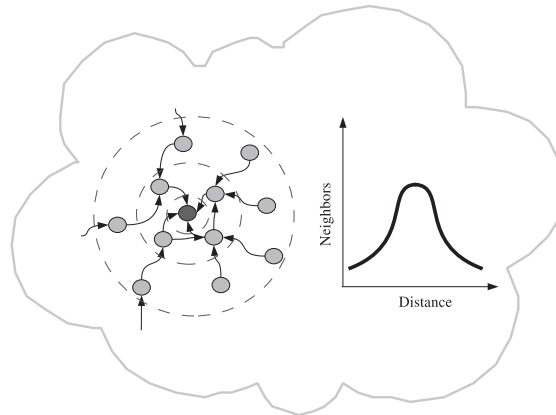


Figure 3. Schematic depiction of the “hop-plot”: a plot of the number of neighbors at different distances.

The mesoscopic level is also the level of description at which local structures, such as communities or clusters of nodes, can be observed.

Figure 4 shows a visual summary of the levels of link-based analysis that we have described.

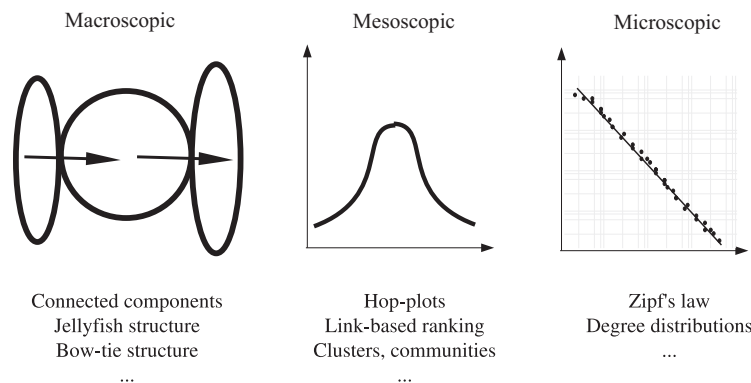


Figure 4. Levels of link-based analysis.

1.2. Ranking through Links

The fact that there might be thousands, or even millions, of pages available for any given topic makes the problem of *ranking* these pages into a short list one the main problems of web IR, thus requiring a method of estimating relevance. Links can be used for this task, if we accept that links frequently imply an endorsement or “vote” [Lifantsev 00], that goes from the author of the source document to the

author of the target document. The efficiency of link-based ranking algorithms indicates that this hypothesis is true in general, but of course, links are placed on webpages for a wide variety of reasons [Haas and Grams 98], and not all those reasons can be interpreted as a positive recommendation.

One of the measures of importance of a scientific paper is the number of citations that the article receives. Following this idea, several authors proposed to use link counts for ranking webpages [Marchiori 97, Carrière and Kazman 97, Joo and Myaeng 98, Li 98]. However, it quickly became clear that just counting the links was not a very reliable measure of authority (it was not in scientific citations either), because it is very easy to manipulate in the context of the web, where creating a page costs almost nothing.

An extensive survey of link ranking methods is presented in [Borodin et al. 05]. We describe here three of the main algorithms: PageRank, HITS, and Salsa.

The PageRank technique, introduced by Page et al., actually tries to mend this problem by looking at the importance of a page in a recursive manner: “a page with high PageRank is a page referenced by many pages with high PageRank” [Page et al. 98]. The algorithm not only counts the direct links to a page but also includes indirect links. The same is valid for scientific and bibliographic citations in general.

PageRank is a simple, robust, and reliable way to measure the importance of webpages, has a clear interpretation as a Markovian process, and can be computed in a very efficient way. For these reasons, most of today’s commercial search engines are believed to use it as a part of their ranking function.

The HITS algorithm [Kleinberg 99] is another method for ranking webpages. It starts by building a set of pages related to a topic by querying a search engine, and then it expands this set by using incoming and outgoing links, by crawling the web or by querying a search engine again. Next, two scores per page are computed: a *hub score* and an *authority score*. A page with a high hub score is a page that links to many pages with a high authority score. A page with a high authority score is a page linked by many pages with high hub score. Despite the circularity of this definition, both hub and authority scores can be computed efficiently by an iterative computation.

In a variation of the HITS algorithm [Bharat and Henzinger 98], the edges are weighted considering the websites to which the source and target nodes of the edges belong. If both edges are in the same site, the weight is zero, and if both edges are in different sites, then the weight of the edge is inversely proportional to the number of inter-site links between them. This makes the algorithm more resistant to manipulation and provides better ranking in practice.

SALSA [Lempel and Moran 01] is related to the idea of hubs and authorities. In SALSA, each page generates two nodes: a hub node and an authority node. All the out-links of the page are attached to the hub node, and all the in-links of the page are attached to the authority node. Next, two random walks are carried, one starting from the “hubs side” and one starting from the “authorities

side.” The probabilities of each state (node) in the stationary states in these two walks are interpreted as two scores for each page.

Given the large size of the graphs arising in web information retrieval applications, scalability is an issue. Most algorithms for large web graphs are limited to a semi-streaming model [Haveliwala 99]. In this model in the case of graph algorithms, the graph is stored on secondary memory as an adjacency list (occupying space in the order of the number of edges), and there is an amount of main memory available that is in the order of the number of nodes. The computation takes place by making a limited number of sequential scans of the graph.

1.3. Our Contribution

In this paper we describe general ranking functions that depend on incoming paths of varying length, and we show that PageRank belongs to this class of functions. We also provide stream algorithms for computing these ranking functions that use memory in the order of the number of nodes and disk space in the order of the number of edges.

Next, we question how these functions relate to each other (i.e., if they produce similar rankings), and finally we test one of the ranking functions for an information retrieval task (ranking a set of pages).

The rest of this paper is organized as follows: Section 2 describes the datasets and experimental framework that we use in the rest of the paper. Section 3 introduces the notion of functional ranking, and Section 4 describes several damping functions. Section 5 compares the ranking functions analytically and experimentally. Finally, Section 6 tests the precision of one of the damping functions, and Section 7 presents our conclusions.

This paper extends the results presented in preliminary form in [Baeza-Yates et al. 06].

2. Datasets and Experimental Framework

In the following sections we experiment with several web datasets. We use several snapshots from the web obtained by the Laboratory of Web Algorithmics, *Dipartimento di Scienze dell’Informazione, Università degli studi di Milano*. These data sets are available at <http://law.dsi.unimi.it/>. In particular, we used the `uk-2002`, `it-2004`, and `eu-int-2005` web graphs. They correspond to a 18-million-page crawl of the `.uk` domain in 2002, a 40-million-page crawl of the `.it` domain in 2004, and a 860,000-page crawl from the `.eu.int` domain in 2005.

In addition to real Web data, we also considered a synthetic scale-free network produced according to the evolving model described by [Kumar et al. 00] (a combination of preferential attachment and random links) with parameters

suggested by Pandurangan et al. [Pandurangan et al. 02]. In the generated graph the exponents for the power law in the center part of the distributions are -2.1 for in-degree and PageRank, and -2.7 for out-degree. We generated a 100,000-node graph without disconnected nodes.

To compare ranking orders among different ranking functions, we used Kendall’s τ [Kendall and Gibbons 90]: this is one of the most widely used and intuitive nonparametric correlation indices, which has recently received much attention within the web community for its possible applications to rank aggregation [Fagin et al. 03b, Fagin et al. 03a, Dwork et al. 01] and for determining the convergence speed in the computation of PageRank [Kamvar et al. 03b]. Kendall’s τ is usually defined as the normalized difference between the number of concordances (i.e., pairs on which the two orders agree) and the number of discordances (i.e., pairs on which the two orders disagree). There are some variants of this measure, which differ on the way ties are treated. Kendall’s τ is always in the range $[-1, 1]$: $\tau = 1$ happens if the two total orders induced by the ranks are the same, whereas $\tau = -1$ happens when the two total orders are opposite of each other; $\tau = 0$ can be interpreted as lack of correlation.

3. Propagating Rank through Links

In this section, we introduce the notion of *functional ranking*, a general family of ranking functions that includes PageRank. To describe PageRank formally, we consider a web graph of N pages. Let $\mathbf{A}_{N \times N}$ be the adjacency matrix in this graph, with $a_{i,j} = 1$ if and only if there is a link from page i to page j . This link matrix is hardly ever used as it is, mainly because it is not normalized and it has “dangling nodes.”

3.1. Normalization

In the web, creating an out-link is free, so there is an incentive for webpage authors to create pages with many out-links; this is the reason why a metaphor of “voting” is enforced [Lifantsev 00] in which each page has only one “vote” that has to be split among its linked pages. This is typically done in link-based ranking by normalizing \mathbf{A} row-wise: the normalization process means that every webpage can only decide how to divide its own score among the pages to which it leads, but it cannot distribute more score than the score it has received. Another way to look at normalization is that the matrix is turned into the transition matrix of a stochastic process.

The normalization does not need to give each out-link the same value, due to the evidence that web links have different purposes such as navigating in a multi-page set, expanding the contents of the current page, pointing to another resource, etc. [Haas and Grams 98]. Also, links within the same site can be

considered self-links and as such do not confer as much authority as a link between different sites; indeed, there are ranking methods like BHITS [Bharat and Henzinger 98] that treat them differently. Other characteristics of links, such as the exploration level at which they appear in websites [Liu and Ma 05], or if they are at the beginning or the bottom of individual pages, or inside a certain HTML element, can also be used for non-uniform normalization [Baeza-Yates and Davis 04].

To simplify our treatment, we will assume uniform normalization, so if a page has d out-links, each of those links has a weight of $1/d$, but the results of this paper can be applied to other forms of normalization.

3.2. Dangling Nodes

Special attention should be paid to the possible presence of nodes with no outgoing arcs (known as “sinks” in graph theory): in fact, dangling nodes fail to produce a row-stochastic matrix, because the rows of dangling nodes are filled with zeroes. Dangling nodes can be dealt with by adding an extra node that is linked to and from all other nodes or by introducing new arcs from each dangling node to every node in the graph [Eiron et al. 04]. In our analysis, we shall assume that all dangling nodes have been eliminated already in some way, so that we do not have to worry about their presence. All the algorithms we will present can be modified so that dangling nodes can be dealt with explicitly and with virtually no additional cost.

Let \mathbf{P} be the row-normalized link matrix of the graph with N nodes. PageRank $\mathbf{r}(\alpha)$ is defined as the stationary distribution of the Markov chain with state transitions given by the matrix

$$\alpha\mathbf{P} + (1 - \alpha)\mathbf{1}^T\mathbf{v},$$

where $\alpha \in [0, 1)$ is a parameter called *damping factor* (sometimes also called dampening factor) and \mathbf{v} is a fixed *preference vector* that may represent the interests of a particular user or another ranking vector that is used for weighting pages. Note that the above matrix is ergodic (at least, if every entry of \mathbf{v} is strictly positive), so it has exactly one stationary distribution. Even though most of our results can be easily restated with a non-uniform preference vector \mathbf{v} , for the sake of clarity we shall only consider the uniform preference $\mathbf{1}/N$ in the rest of the paper.

As previously observed, the PageRank vector $\mathbf{r}(\alpha)$ can be written as

$$\mathbf{r}(\alpha) = (1 - \alpha) \sum_{t=0}^{\infty} \alpha^t \frac{1}{N} \mathbf{1} \mathbf{P}^t$$

or in matrix form

$$\mathbf{r}(\alpha) = (1 - \alpha) \frac{1}{N} \mathbf{1} (\mathbf{I} - \alpha \mathbf{P})^{-1}, \quad \|\alpha \mathbf{P}\| < 1$$

[Fogaras 03, Boldi et al. 05]. There is, in fact, an equivalent, very intriguing way of rewriting this formula [Newman et al. 01] that leads to a conclusion similar to those of Brinkmeier [Brinkmeier 06]: given a path, that is, a sequence of edges in the graph $p = \langle x_1, x_2, \dots, x_k \rangle$, such that node x_i is connected to node x_{i+1} , we define its *branching contribution* as follows:

$$\text{branching}(p) = \frac{1}{d_1 d_2 \cdots d_{k-1}},$$

where d_j is the out-degree, this is, the number of outgoing arcs, of node x_j .

Then, the ranking of node i according to PageRank is

$$r_i(\alpha) = \sum_{p \in \text{Path}(-, i)} \frac{(1 - \alpha)\alpha^{|p|}}{N} \text{branching}(p),$$

where $\text{Path}(-, i)$ is the set of all paths into node i and $|p|$ is the length of path p ; this is because $(\mathbf{P}^t)_{ij}$ contains the sum of the branching contributions of all paths of length t from i to j , as one can easily show by induction on t (a path of length 0 and branching 1 is also included in the summation). This way of expressing the PageRank of a node is interesting, because it highlights the fact that the rank of a node is essentially obtained as a weighted sum of contributions coming from every path entering into the node, with weights that decay exponentially in the length of the path.

A natural generalization of this idea consists in taking into consideration a ranking \mathbf{R} of the general form:

$$\mathbf{R} = \sum_{t=0}^{\infty} \text{damping}(t) \frac{1}{N} \mathbf{1} \cdot \mathbf{P}^t$$

or, equivalently,

$$R_i = \sum_{p \in \text{Path}(-, i)} \text{damping}(|p|) \frac{1}{N} \text{branching}(p),$$

where the damping function is a suitable choice of weights.

We call this form of ranking a *functional ranking*, as it is parametrized by a damping function. This generalizes Lifantsev’s model in which the damping factor is a matrix of *voting trust* that is fixed during the computation [Lifantsev 00], whereas in our case this explicitly depends on the iterations. Our damping function could be even more general by using $\mathbf{D}(t)$, a damping matrix instead of $\text{damping}(t) \frac{1}{N} \mathbf{1}$; in this paper we analyze only the latter form. Fogaras proposed using decreasing link weights depending on path lengths in the reverse link graph and used exponentially decreasing weights as in PageRank for finding good web browsing “starting points” in the web graph [Fogaras 03]. Another, yet unexplored, possible direction would be to consider damping functions that depend

on other properties of the paths (e.g., whether the path passes through some node out of a certain set) rather than on their length.

As we have seen, generic PageRank is a functional ranking where the damping function

$$\text{damping}(t) = (1 - \alpha)\alpha^t$$

decays exponentially fast (something similar was first considered in citation analysis back in 1953! [Katz 53]).

3.3. Characteristic Path Lengths

In scale-free networks, the distances between pairs of nodes follow a Gaussian distribution [Albert et al. 99]. Analytic estimations for the average distance of a graph of scale-free network of n nodes include the following:

- $O(\log(n))$ [Watts and Strogatz 98];
- $O(\log(n)/\log(np))$ in sparse graphs with p links [Chung and Lu 01];
- $1 + \log(n/z_1)/\log(z_2/z_1)$ where z_1 is the average in-degree, and z_2 is the average number of nodes at distance 2 [Newman et al. 01]; and
- $O(\log(n)/\log(\log(n)))$ [Bollobás and Riordan 04].

The above results apply to different static scale-free networks, not to the evolution of a particular scale-free network over time. Empirical observations in several different domains demonstrate that given a specific graph, its diameter may shrink over time, even if its number of nodes n is increasing [Leskovec et al. 05].

In the static graphs we have (.eu.int, .uk, and .it), we did the following experiment: starting from a node picked at random, we followed the links backwards and counted the number of nodes at different distances. Figure 5 plots the

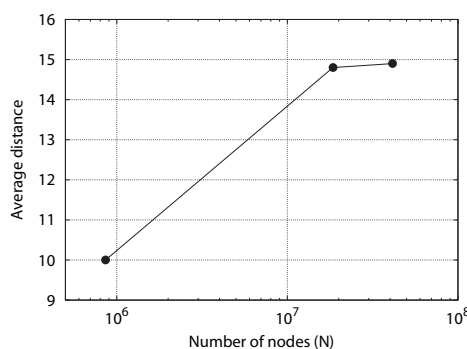


Figure 5. Average distances versus number of nodes in three web graphs.

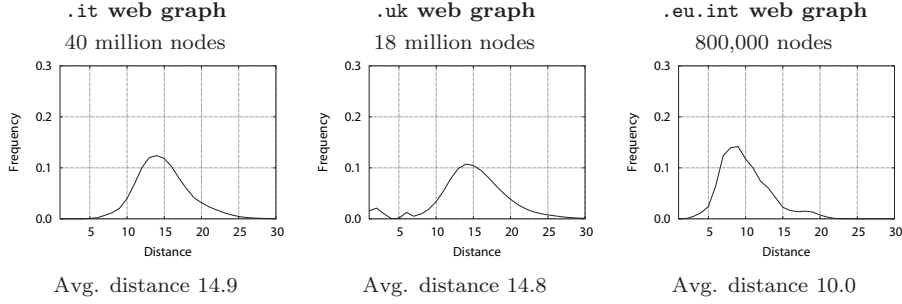


Figure 6. Distribution of the average number of nodes at a certain distance from a given node, in three web samples.

average distances found, which appear to be related (sub)logarithmically with the size of the graph. Figure 6 shows the distribution obtained in these samples. For this experiment, we are not counting the pages without in-links.

The act of linking a page represents human endorsement and should not be affected by the size of the graph. Nor should the act of following a link, in terms of a random surfer, be affected. However, an algorithm for *propagating* this endorsement through links for computing a ranking function needs to account for the typical distances involved. This requirement is typical in a situation where local properties have a global impact: for example, the addition of a single arc could drastically reduce the diameter of a graph.

In most cases, researchers have used exponential damping with base 0.85 or 0.90 in graphs that are much smaller than the full web (concept graphs, social networks, email graphs, etc.), meaning that a potentially much larger fraction of the nodes contributed towards link ranking. We consider that, in a smaller graph, the damping function should decay faster.

If graphs of different sizes show different path lengths, what is the effect of this in the ranking calculation? Let's suppose that for a graph with N_1 nodes it is found, by experimental or analytic means, that a good parameter for PageRank is α_1^* . Now, we would like to have a good parameter α_2^* for a graph with the same properties, except that the size of the new graph is $N_2 < N_1$.

One possible approach, remaining consistent with what we have done so far, is to view the sum of the damping functions up to the average path lengths of the graphs (L_1, L_2) as having to be similar in order for both rankings to behave in a similar way. If we take this approach, the solution is

$$\begin{aligned} \sum_{t=0}^{L_1} (1-\alpha)\alpha^t &= \sum_{t=0}^{L_2} (1-\alpha)\alpha^t, \\ 1 - (\alpha_1^*)^{L_1+1} &= 1 - (\alpha_2^*)^{L_2+1}, \end{aligned}$$

$$\begin{aligned}\alpha_2^* &= (\alpha_1^*)^{\frac{L_1+1}{L_2+1}}, \\ \alpha_2^* &\approx (\alpha_1^*)^{\frac{\log(N_1)}{\log(N_2)}}.\end{aligned}$$

An example that can be put into practice is the following: let’s consider a web graph with $N_1 = 11.5 \times 10^9$ pages (the size of the full web estimated by [Gulli and Signorini 05]) and another graph with only $N_2 = 50 \times 10^6$ pages (the size of the web of a large country); the second graph is roughly three orders of magnitude smaller.

If it is shown empirically that $\alpha_1^* = 0.85$ is a good value for the PageRank parameter for the whole web, then $\alpha_2^* = 0.81$ should have a similar behavior in the 50-million-page set, which is natural as the path lengths are shorter. If the subset of webpages were even smaller, for instance, $N_2 = 10^6$ pages (the size of the web of a large organization), then $\alpha_2^* = 0.76$, and for smaller graphs of $N_2 = 10^5$ nodes, $\alpha_2^* = 0.72$. We recommend using these values for graphs that are not comparable in size to the full web graph.

4. Damping Functions

First, we show which class of damping functions generates well-defined functional rankings. For every pair of nodes i and j , and for every length t [Brinkmeier 06, Corollary 2.4],

$$\sum_{p \in \text{Path}(i,j), |p|=t} \text{branching}(p) \leq 1.$$

A more general property holds:

Theorem 4.1. *For every node i and every length t ,*

$$\sum_{p \in \text{Path}(i,-), |p|=t} \text{branching}(p) = 1.$$

Proof. We proceed by induction on t . For $t = 0$, there is only one path from i of length 0, and its branching is 1. For the inductive step the above expression can be rewritten by observing that, if i has out-degree d_i , every path from i of length $t + 1$ is the concatenation of i with a path of length t from an out-neighbor of i :

$$\begin{aligned}\sum_{p \in \text{Path}(i,-), |p|=t+1} \text{branching}(p) &= \sum_{j: i \rightarrow j} \frac{1}{d_i} \sum_{p \in \text{Path}(j,-), |p|=t} \text{branching}(p) \\ &= \sum_{j: i \rightarrow j} \frac{1}{d_i} = 1.\end{aligned}$$

□

As a consequence, to guarantee that the functional ranking is well-defined and normalized (i.e., that rank values sum to 1), we need

$$\sum_{i=1}^N \sum_{p \in \text{Path}(-, i)} \text{damping}(|p|) \frac{1}{N} \text{branching}(p) = 1;$$

that is,

$$\sum_{t=0}^{\infty} \text{damping}(t) \frac{1}{N} \sum_{p \in \text{Path}(-, -), |p|=t} \text{branching}(p) = 1.$$

Using Theorem 4.1, $\sum_{p \in \text{Path}(-, -), |p|=t} \text{branching}(p) = N$, so the latter equality is equivalent to

$$\sum_{t=0}^{\infty} \text{damping}(t) = 1.$$

Hence, every choice of the damping function such that the sum of dampings is 1 yields a well-defined normalized functional ranking. However, not all choices are equivalent, so we have to find out which functions generate better rankings. Since a direct link should be more valuable as a source of evidence than a distant link, we focus on damping functions that are decreasing on t , the length of the paths. We also focus on normalized ranking functions, as they are easier to combine with other signals to produce a combined ranking for an object.

For calculating functional rankings, we use the general algorithm shown in Figure 7. Informally, the general process is to initialize a set of vectors of size proportional to the number of nodes and then perform passes over the graph until a stop condition is met. At each pass, an accumulator is being increased for each node. The next sections provide details on the initialization, stop condition, and iteration steps for each calculation.

4.1. Linear Damping

Let's start by considering a simple damping function such as

$$\text{damping}(t) = \begin{cases} \frac{2(L-t)}{L(L+1)}, & t < L \\ 0, & t \geq L; \end{cases}$$

that is, a damping function that decreases linearly with distance and that reaches zero at distance L . The trivial case $L = 1$ gives a uniform ranking, and $L = 2$ is ranking by in-degree; as in the latter case, all paths of length ≥ 2 are not considered.

From the definition,

$$\begin{aligned}\mathbf{R} &= \sum_{t=0}^{\infty} \text{damping}(t) \mathbf{v} \mathbf{P}^t = \sum_{t=0}^L \frac{2(L-t)}{L(L+1)} \mathbf{v} \mathbf{P}^t \\ &= \frac{2}{L(L+1)} \mathbf{v} \sum_{t=0}^{L-1} (L-t) \mathbf{P}^t \\ &= \frac{2}{L(L+1)} \mathbf{v} (L(\mathbf{I} - \mathbf{P}) - \mathbf{P}(\mathbf{I} - \mathbf{P}^L)) ((\mathbf{I} - \mathbf{P})^2)^{-1},\end{aligned}$$

provided that $(\mathbf{I} - \mathbf{P})^2$ is not singular.

An advantage of this type of ranking is that only the first few levels are taken into consideration, so the number of iterations is fixed. The rationale for this is that after a certain distance the information given by links can be disregarded.

For computing this functional ranking, we can define the following sequence:

$$\begin{aligned}\mathbf{R}^{(0)} &= \frac{2}{L+1} \mathbf{v}, \\ \mathbf{R}^{(k+1)} &= \frac{(L-k-1)}{(L-k)} \mathbf{R}^{(k)} \mathbf{P}.\end{aligned}$$

The functional ranking with linear damping is $\sum_{k=0}^{L-1} \mathbf{R}^{(k)}$. For computing this ranking, the generic algorithm shown in Figure 7 can be used, with

$$\begin{aligned}\text{START} &: 2v[i]/(L+1), \\ \text{STOP} &: k = L, \\ \text{DAMP}(k) &: (L-k)/(L-(k-1)).\end{aligned}$$

4.2. Exponential Damping: PageRank

As we already noted, PageRank can be seen as a functional ranking where the damping function decays exponentially:

$$\text{damping}(t) = (1 - \alpha) \alpha^t.$$

Given that longer paths are of lower importance in the calculation of PageRank, it could be approximated by using only a few levels of links. It has been shown that by using only the nodes at distance 1 from a target node (equivalent to linear damping with $L = 2$), PageRank values can be approximated with 30% of average error [Chen et al. 04]. Using nodes at distance 2, the average error drops to 20%, and at distance 3 to 10%. After that, there are no significant improvements by adding a few more levels, and the cost (the number of nodes to be explored) is much higher.

Since PageRank is the principal eigenvector of the modified graph matrix, it can be easily approximated by the iterative power method algorithm, as suggested by [Page et al. 98] in their original paper; this iterative algorithm gives good approximations (both in norm and with respect to the induced node order) in few iterations, even though convergence speed and numerical stability decay when α gets close to 1 [Haveliwala and Kamvar 03b, Haveliwala and Kamvar 03a]. Other methods to compute PageRank have been proposed, some of them using techniques for the solution of systems of linear equations, some others concentrating on some specific features of the web as a graph that determine forms of locality in the computation of PageRank (see, for example, [Page et al. 98, Haveliwala 99, Golub and Greif 04, Lee et al. 04, Kamvar et al. 03b, Kamvar et al. 03a]).

Of course, the generic algorithm shown in Figure 7 can be used, with

START : $(1 - \alpha)v[i]$,
 STOP : convergence,
 DAMP(k) : α .

Require: N: number of nodes, \mathbf{v} : preference vector

```

1: for  $i : 1 \dots N$  do {Initialization}
2:    $S[i] \leftarrow R[i] \leftarrow \text{START}$ 
3: end for
4: for  $k : 1 \dots \infty$  do {Iteration step}
5:   if STOP then
6:     break
7:   end if
8:    $\text{Aux} \leftarrow \mathbf{0}$ 
9:   for  $i : 1 \dots N$  do {Follow links in the graph}
10:    for all  $j$  such that there is a link from  $i$  to  $j$  do
11:       $\text{Aux}[j] \leftarrow \text{Aux}[j] + R[i]/\text{out-degree}(i)$ 
12:    end for
13:  end for
14:  for  $i : 1 \dots N$  do {Add to ranking value}
15:     $R[i] \leftarrow \text{Aux}[i] \times \text{DAMP}(k)$ 
16:     $S[i] \leftarrow S[i] + R[i]$ 
17:  end for
18: end for
19: return S

```

Figure 7. Template algorithm for computing a functional damping. START, STOP, and DAMP(k) differ for each functional ranking.

4.3. Quadratic Hyperbolic Damping: TotalRank

Recently, a ranking method called TotalRank has been proposed [Boldi 05]. The method aims at eliminating the necessity for an arbitrary parameter by integrating PageRank over the entire range of α . If $\mathbf{r}(\alpha)$ is the vector of PageRank, then TotalRank is defined as

$$\mathbf{T} = \int_0^1 \mathbf{r}(\alpha) d\alpha .$$

\mathbf{T} can be written as

$$\begin{aligned} \int_0^1 \mathbf{r}(\alpha) d\alpha &= \frac{1}{N} \sum_{t=0}^{\infty} \int_0^1 (1-\alpha) \alpha^t \mathbf{1} \cdot \mathbf{P}^t d\alpha \\ &= \frac{1}{N} \sum_{t=0}^{\infty} \frac{1}{(t+1)(t+2)} \mathbf{1} \cdot \mathbf{P}^t, \end{aligned}$$

where the first equality is obtained by applying Theorem 1.27 of [Rudin 86].

By using the definition of the logarithm of a matrix,

$$\ln(\mathbf{I} - \mathbf{P}) = - \sum_{k=1}^{\infty} \frac{\mathbf{P}^k}{k},$$

we can write TotalRank as

$$\mathbf{T} = \mathbf{P}^{-1}(\mathbf{I} + (\mathbf{I} - \mathbf{P}^{-1}) \ln(\mathbf{I} - \mathbf{P}))$$

provided that \mathbf{P} is not singular and $\mathbf{P} \neq \mathbf{I}$.

TotalRank is a weighted sum of the scores associated with paths of varying lengths, in which the weights are hyperbolically decreasing on the lengths of the paths. In other words, TotalRank is a functional ranking with damping function

$$\text{damping}(t) = \frac{1}{(t+1)(t+2)} = \frac{1}{t+1} - \frac{1}{t+2},$$

and it is well defined since $\sum_{t=0}^{\infty} \text{damping}(t) = 1$.

It is known that the cost of calculating TotalRank is the same as the cost of calculating PageRank via the power method [Boldi et al. 05], even though some more iterations are required to obtain the same precision.

4.4. General Hyperbolic Damping: HyperRank

TotalRank is part of a more general family of weighting schemes for paths of different lengths that can be approximated using

$$\mathbf{s}(\beta) = \frac{1}{N\zeta(\beta)} \sum_{t=0}^{\infty} \frac{1}{(t+1)^\beta} \mathbf{1} \cdot \mathbf{P}^t .$$

Again, this way of ranking follows the general scheme, with damping function chosen as

$$\text{damping}(t) = \frac{1}{\zeta(\beta)(t+1)^\beta}.$$

Here, we are using Riemann’s zeta function, $\zeta(\beta) = \sum_{t=1}^{\infty} t^{-\beta}$ for normalization, and we need $\beta > 1$ for it to converge. Note that when $\beta = 2$ we get weights similar to those of TotalRank, in which the t th coefficient is $1/((t+1)(t+2))$ whereas here it is $1/(\zeta(2)(t+1)^2)$.

A meaningful choice for β should be made considering the distribution of paths of different lengths in a scale-free graph. A large α in PageRank, or a small β in HyperRank, means increasing the effect of longer paths in the score.

Let us define a vector sequence $\mathbf{R}^{(t)}$ as follows:

$$\begin{aligned}\mathbf{R}^{(0)} &= \frac{1}{N\zeta(\beta)}, \\ \mathbf{R}^{(k+1)} &= \left(\frac{k+1}{k+2}\right)^\beta \mathbf{R}^{(k)} \mathbf{P}.\end{aligned}$$

It is easy to see that $\sum_{t=0}^{\infty} \mathbf{R}^{(k)} = \mathbf{s}(\beta)$, because $\mathbf{R}^{(k)} = 1/(N \cdot \zeta(\beta)(k+1)^\beta) \mathbf{1} \cdot \mathbf{P}^k$; this observation allows us to use the generic algorithm of Figure 7 with the following parameters:

$$\begin{aligned}\text{START} &: v[i]/\zeta(\beta), \\ \text{STOP} &: \text{convergence}, \\ \text{DAMP}(k) &: (k/(k+1))^\beta.\end{aligned}$$

Note that convergence speed is much slower than ordinary PageRank, especially when β is close to 1, the norm of the k th summand being bound by $1/(1+1/k)^\beta$. Interestingly enough, though, convergence speed is reasonable if β is sufficiently large.

4.5. An Empirical Damping

An empirical damping function would consider how much the value of an endorsement decreases by following longer paths in the real web graph. This cannot be known exactly, but we can attempt to measure it indirectly. Pages that are linked to each other share a greater degree of similarity than pages chosen at random [Davison 00]; evidence from topical crawlers [Srinivasan et al. 05] shows that when doing breadth-first exploring, the topic “drifts” as the distance increases. On the same line of thought, we propose to use the decrease of text similarity as an approximation to an “empirical” damping function. It has been shown that text similarity and link distance are anti-correlated up to 4–5 links [Menczer 04].

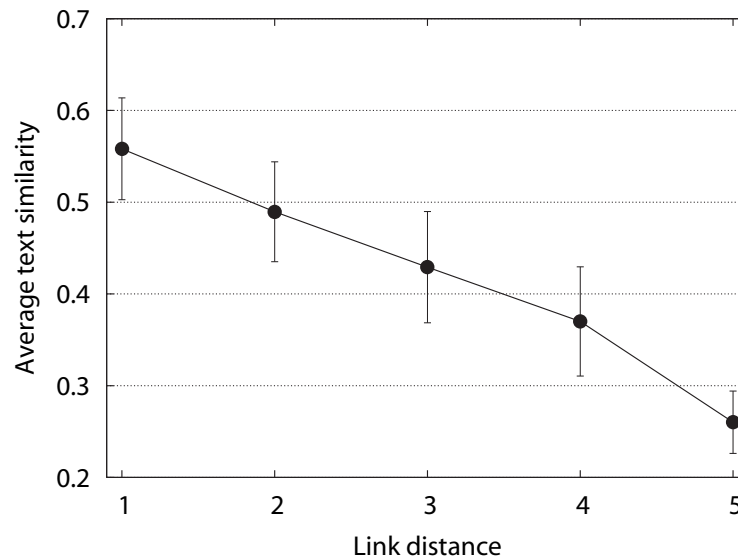


Figure 8. Link distance versus average text similarity in a sample of 18 million pages from the .uk domain. A link distance of one means direct linking. The text similarity appears to decrease linearly in the first few levels.

In order to assess the correlation between link distance and similarity, we performed the following experiment: We considered a web graph corresponding to a partial snapshot of the .uk domain with 18 million pages and sampled 200 nodes at random. For each sampled node, we followed links backwards to obtain nodes at a minimum distance of 1, 2, 3, 4, or 5 links. Then, we sampled 12,000 pairs at each minimum distance at random and computed their similarities with the original nodes. Similarity was measured using the normalization of TF.IDF [Baeza-Yates and Ribeiro-Neto 99], without stemming or stop-word removal.

The resulting averages are shown in Figure 8, with standard deviation error bars. Text similarity clearly decreases with distance, and in some applications the empirical distribution of text similarity versus distance could be used as an “empirical” damping function. Different measures of text similarity can yield different distributions; for instance [Wu et al. 04] used the number of repeated words and phrases between pages and obtained a faster decrease in similarity. Our results show that in our data set a linear damping with $L = 8$ or $L = 9$ approximates better the decrease of text similarity with distance than an exponential damping (as suggested in [Menczer 04]). text similarity does not seem to decrease exponentially fast, so there is no *a priori* reason to prefer exponential damping (pagerank) over other functional rankings. An observation by [Menczer 04] is that for different communities, the link structure could

be different. For instance, academic webpages might be better connected than commercial pages, so an empirical damping function should measure first which is the correlation of link distance to text similarity in the specific collection that we want to rank.

5. Comparing Damping Functions

A comparison of the damping functions described in the previous section is shown in Figure 9: of course, hyperbolic damping functions decay asymptotically more slowly than exponential damping, but note that for short paths the latter may dominate the former in many cases.

We can empirically observe that the ranking orders produced by different functional rankings are different. Nevertheless, in this section, we show that the ranking order produced by one functional ranking can be approximated with great precision by carefully choosing the parameters. This approximation has to be done mostly by considering the weight of the first few levels of links.

The possibility of approximating the order of one functional ranking with another is interesting, for instance, to approximate PageRank using LinearRank (given that the latter uses a fixed number of iterations), or by another functional ranking in the future. In this section we also include approximations of PageRank with TotalRank and HyperRank for completeness.

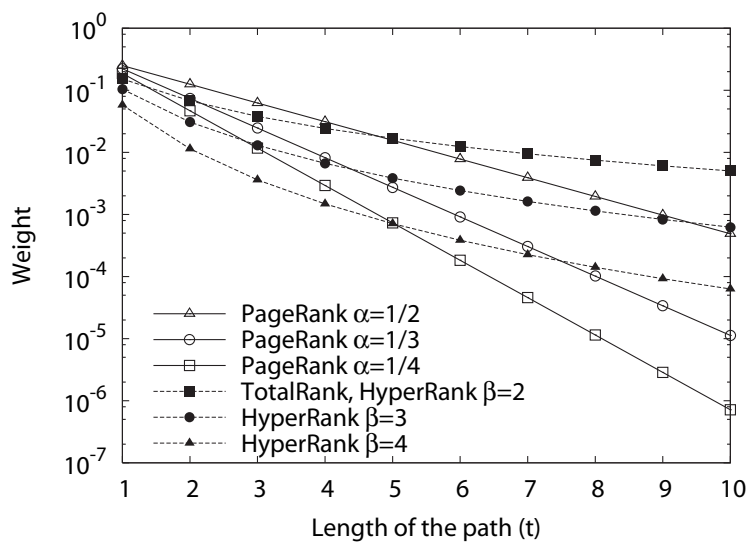


Figure 9. Weights given by the different damping functions for values of α and β .

5.1. Approximating PageRank with TotalRank

It has been observed experimentally that the rank correlation (Kendall’s τ) between TotalRank and PageRank is maximal when $\alpha \approx 0.7$ [Boldi 05]; the maximum value for τ is over 0.95, so for that specific choice of α , PageRank and TotalRank induce almost equivalent ranking orders.

We now want to approach the same problem in an analytic fashion; to be more exact, we aim to study the difference between TotalRank and PageRank by calculating the difference between their respective damping functions:

$$\begin{aligned} \text{damping}_{\text{TotalRank}}(t) &= \frac{1}{(t+1)(t+2)}, \\ \text{damping}_{\text{PageRank}(\alpha)}(t) &= (1-\alpha)\alpha^t. \end{aligned}$$

As they are normalized, both damping functions have the same summation over the entire range of t . Our approach is to consider the summation of their differences up to a maximum length for a path. As the two functions are decreasing, the difference in the first levels makes most of the difference in the rankings. If ℓ is the maximum path length in which we are interested, we aim at minimizing this sum:

$$\sum_{t=0}^{\ell} \left(\frac{1}{(t+1)(t+2)} - (1-\alpha)\alpha^t \right) = \alpha^{\ell+1} - \frac{1}{\ell+2}.$$

The minimum absolute value is 0, and it is obtained when α is equal to

$$\alpha^*(\ell) = \frac{1}{\sqrt[\ell+1]{\ell+2}} = 1 - \frac{\log \ell}{\ell} + O\left(\frac{\log^2 \ell}{\ell^2}\right).$$

Figure 10 shows $\alpha^*(\ell)$ as a function of ℓ . Recall that for the World Wide Web graph, the average length of a path between two nodes, when a path exists,

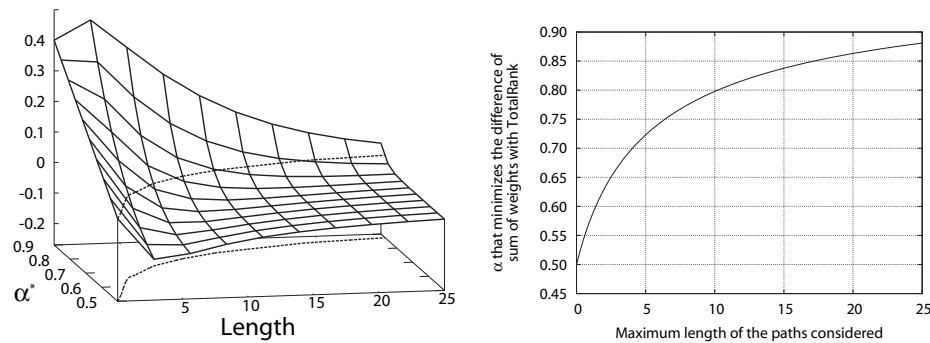


Figure 10. The difference of the sum of the weights for various combinations of α and ℓ (left), and $\alpha^*(\ell)$ for minimizing the difference of the sum of weights between PageRank and TotalRank (right).

has been estimated in about 16 [Broder et al. 00] or 19 [Albert et al. 99], but clearly today it is over 20. Now, in the range of path lengths between 15 and 20, the value of $\alpha^*(\ell)$ parameters that minimizes the difference between the exponentially decaying weights of PageRank and the hyperbolically decaying weights of TotalRank is roughly 0.85. Note that 0.85 is also the most typically used value for the damping factor, so this merits further study.

5.2. Approximating PageRank with HyperRank

Now we want to approximate the values of the damping in

$$\mathbf{r}(\alpha) = \frac{1-\alpha}{N} \sum_{t=0}^{\infty} \alpha^t \mathbf{P}^t,$$

using the values in

$$\mathbf{s}(\beta) = \frac{1}{N\zeta(\beta)} \sum_{t=0}^{\infty} \frac{1}{(t+1)^\beta} \mathbf{P}^t.$$

We proceed again by considering paths up to a certain length:

$$\sum_{t=0}^{\ell} \left(\frac{1}{\zeta(\beta)(t+1)^\beta} - (1-\alpha)\alpha^t \right).$$

The minimum can be zero, and it is attained at

$$\alpha^*(\ell, \beta) = \sqrt[\ell]{1 - \frac{1}{\zeta(\beta)} \sum_{t=0}^{\ell} \frac{1}{(t+1)^\beta}}.$$

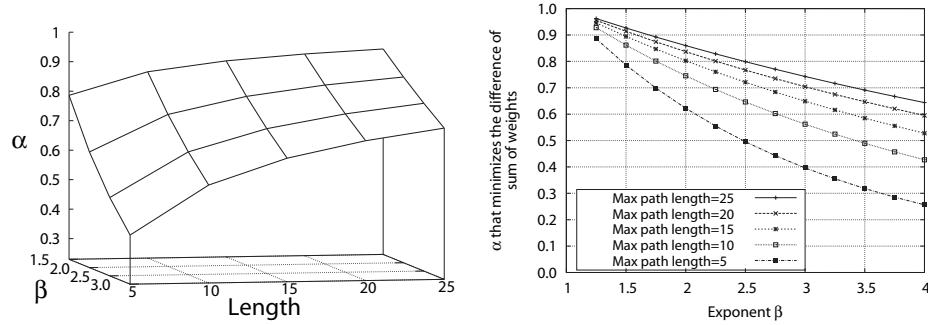


Figure 11. The best α for minimizing the difference of the sum of weights between PageRank and HyperRank, for various parameter combinations (left), and a two-dimensional view of the same plot (right).

The α that minimizes the difference of summation of damping functions for different values of β and of the maximum path lengths ℓ is shown in Figure 11. In the case of $\beta = 2$, for instance, for path lengths up to 10 to 20, the best α is between 0.75 and 0.85.

5.3. Approximating PageRank with LinearRank

For approximating the damping function of PageRank with the damping function of LinearRank, we consider the summation of the differences up to a certain path length. If $\ell \leq L$,

$$\sum_{t=0}^{\ell} \left((1-\alpha)\alpha^t - \frac{2(L-t)}{L(L+1)} \right),$$

and if $\ell > L$,

$$\sum_{t=0}^{L-1} \left((1-\alpha)\alpha^t - \frac{2(L-t)}{L(L+1)} \right) + \sum_{t=L}^{\ell} (1-\alpha)\alpha^t.$$

We will assume that $\ell \leq L$, so the evaluation of the difference between the two rankings is done in an area where both rankings have non-zero values. The L that minimizes the difference for a given combination of α and ℓ is

$$\begin{aligned} L^*(\alpha, \ell) &= \ell + \frac{(2\ell+1)\alpha^{\ell+1} + 1 + \sqrt{(1+\alpha^{\ell+1})^2 + 4\ell(\ell+2)\alpha^{\ell+1}}}{2(1-\alpha^{\ell+1})} \\ &= \ell + 1 + O\left(\ell\alpha^{(\ell+1)/2}\right), \end{aligned}$$

and we have plotted it for different values of α and ℓ in Figure 12.

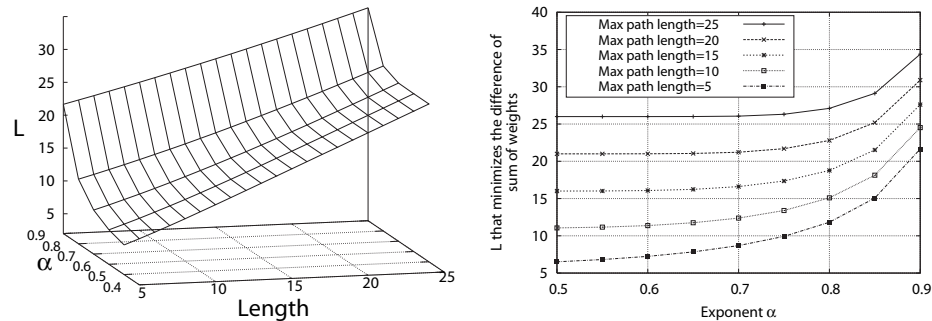


Figure 12. Best L for minimizing the difference of the sum of weights between LinearRank and PageRank, for various parameter combinations (left), and a two-dimensional view of the same plot (right).

5.4. Experimental Comparison of Ranking Orders

In this section, we present experimental results about the similarity between the ranking orders induced by functional rankings discussed in the previous sections. To perform the experiments, we used data from the U.K. web graph.

Figure 13 shows how PageRank compares with HyperRank for various pairs of α and β . In the limit $\alpha, \beta \rightarrow 1$ both rankings are equivalent, and they remain similar in a large region of the parameter space. In this figure, we can see that the rankings obtained with HyperRank and PageRank can be almost equivalent (Kendall's $\tau \geq 0.95$). Furthermore, the analysis shown in Section 5.2, which only considers paths of lengths less than 5, provides a very good approximation for the optimal combination of parameters. This means that, in fact, the difference in the damping functions in the first few levels is crucial.

The exponents β required for giving a good approximation of PageRank are very small when $\alpha \geq 0.7$, limiting the practical applicability of HyperRank, as it does not converge more quickly than PageRank.

This comparison was corroborated by an analogous series of experiments where we used another (dis)similarity measure [Fagin et al. 03b]: this measure, called *intersection metric*, is essentially an averaged normalized measure of the symmetric difference between the two top- k sets according to two given rankings; the intersection metric evaluates to 1 when the top lists are disjoint. To allow comparison with Kendall's τ , we choose to graphically represent in Figure 14 one minus the intersection metric; the choice of k is of course relevant, but the results are uniform for sufficiently large k .

As far as LinearRank and PageRank are concerned, long paths and large α should be considered to obtain a sufficiently similar ranking, as shown in Figure 15. In the range of $\alpha = 0.8 \dots 0.9$, paths of roughly 10 to 20 links should be considered to obtain rankings that are almost equivalent.

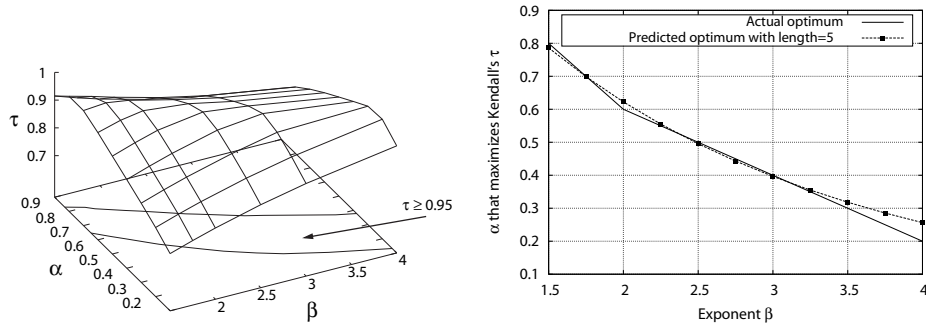


Figure 13. Comparison (using Kendall's τ) between PageRank and HyperRank, with various damping parameters in the U.K. web graph (left). The optimum predicted in the analysis with $\ell = 5$ is very close to the real one (right).

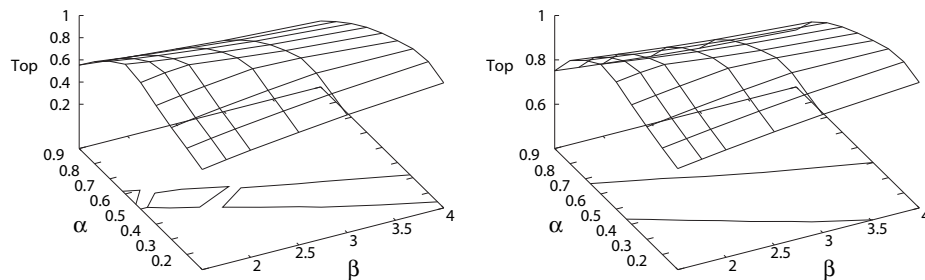


Figure 14. Comparison (using the top- k intersection metric) between PageRank and HyperRank, with various damping parameters in the U.K. web graph, for $k = 1,000$ (left) and $k = 100,000$ (right).

The predicted optimum given in Section 5.3 with $\ell = 5$ (i.e., considering only the summation of the differences between both damping functions up to paths of length 5) is very close to what was obtained in practice. For $\alpha = 0.8$, calculating LinearRank with $L = 10$ (which means the same number of iterations) gives $\tau \geq 0.98$; for $\alpha = 0.9$, calculating LinearRank with $L = 15$ also gives $\tau \geq 0.98$. In both cases, the ranking order of PageRank is approximated by the ranking order of LinearRank with very high precision. (A similar comparison was performed using the intersection metric instead, obtaining quite similar results.)

As a final remark, observe that (as shown in Figure 16) even though LinearRank is a good approximation to PageRank, stopping PageRank computation after ℓ iterations usually gives a better (in the sense of Kendall's τ) approximation to real PageRank than LinearRank parametrized by ℓ , especially for small α , where convergence is fast. Whether this observation could be extended to the

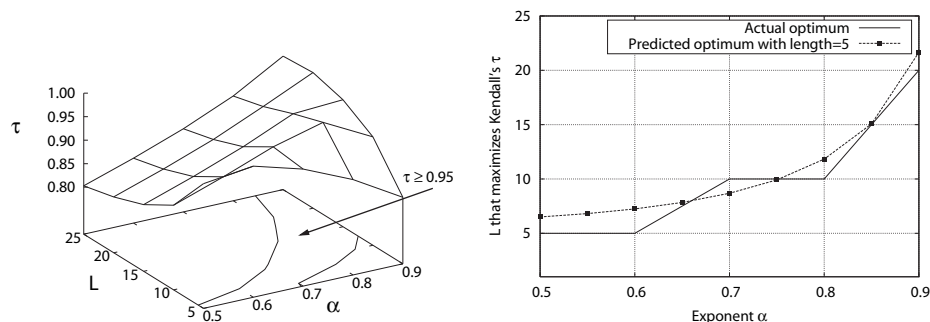


Figure 15. Comparison (using Kendall's τ) between PageRank and LinearRank in the U.K. web graph, with various damping parameters (left). Again, the predicted optimum with $\ell = 5$ is very close to the actual optimum (right).

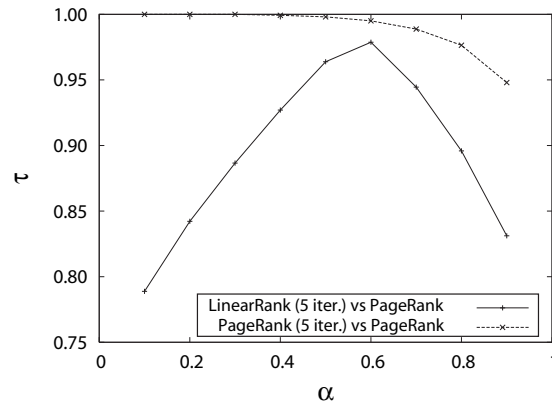


Figure 16. Comparison (using Kendall’s τ) between PageRank stopped after five iterations and LinearRank with $\ell = 5$ in the U.K. web graph, with various damping parameters.

precision of both ranking functions for information retrieval tasks is a problem that merits further experiments and investigation.

5.5. Comparison with In-Degree

In this section, we study the behavior of the ranking functions for different values of their parameters. We are using data from the .uk web graph and a 8,500-node synthetic graph. We first measured the variance of the values from the ranking function, as we consider that a high variance is good in a ranking function as the relative values differ more. We also measured the relationship between the ranking function and in-degree for different values of the parameters in terms of the correlation coefficient and ranking orders (Kendall’s τ). The results are shown in Figure 17.

The variance is higher as α increases. As far as the relationship with in-degree is concerned, for company home pages it has been observed that the logarithm of the in-degree is correlated with PageRank [Upstill et al. 03]. Our results are consistent with this observation. Not surprisingly, the correlations are higher in the generative model. We observe a maximum correlation at $\alpha = 0.7$ in the synthetic graph and at $\alpha = 0.5$ in the web graph. We also notice that the correlation drops significantly as α gets larger, because a large α means that longer paths have an effect in the calculation; note, that this phenomenon does not significantly affect the correlation coefficient that is still very large.

A high correlation between PageRank and in-degree is bad from the point of view of a search engine, because it makes link spam easier. In particular, as the correlation coefficient is higher in the .uk web graph near 0.5, if we choose α close to this value, we are helping link spammers. Note, however, that a

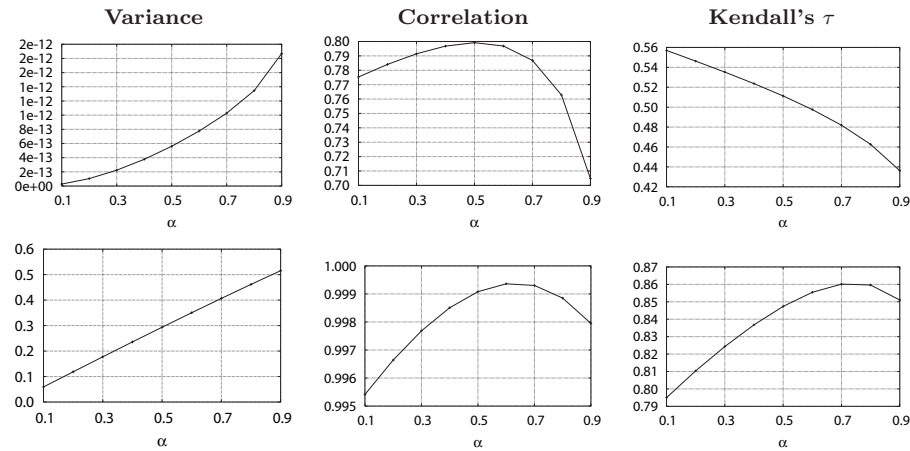


Figure 17. Variance of PageRank (left), and its relationship with in-degree in terms of the correlation coefficient (middle) and Kendall's τ coefficient (right), for varying values of the parameter α , in the .uk web graph (top) and the synthetic graph (bottom).

high correlation was foreseeable because (as shown in [Chen et al. 04]), even approximating PageRank with only one level of links gets 70% accuracy.

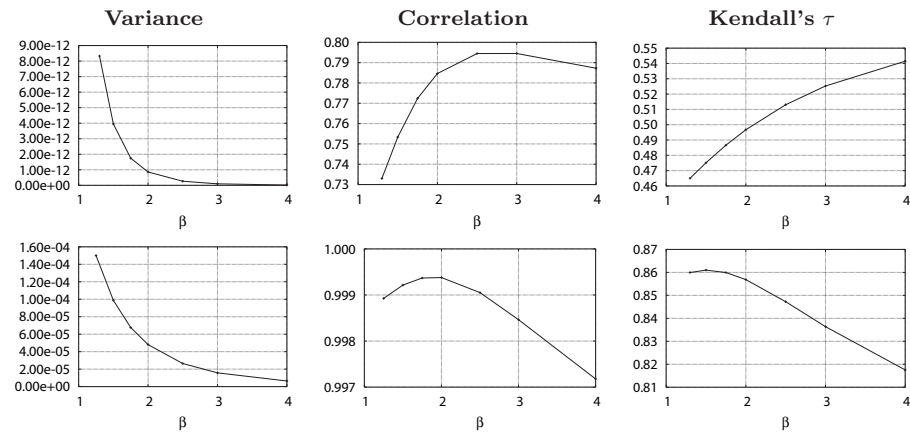


Figure 18. Variance of general hyperbolic rank for some values of β (left), and its relationship with in-degree in terms of the correlation coefficient (middle) and Kendall's τ coefficient (right). Experiments were performed on the .uk graph (top) and the synthetic graph (bottom).

The behavior of the Kendall’s τ coefficient that measures the similarity between ranking orders is the opposite of the one observed in the real graph. This also happens for HyperRank: in Figure 18 we made the same measurements for this functional ranking, and the results were consistent (the graph seems inverted because a low value of β has the same effect as a high value of α : longer paths have more importance in the calculation).

The differences in the behavior of the ranking order in the synthetic graph might be explained by the fact that the generative model we are using does not capture some of the properties that might be relevant for the ranking order under a functional ranking such as the clustering coefficient. Also, the synthetic graph is *assortative* (highly linked pages are mostly linked to other highly linked pages), while the real web graph is *disassortative* (most of the neighbors of highly linked pages have small in-degree).

6. Experimental Evaluation of Precision

Finally, we would like to determine if a damping function that does not decay exponentially as PageRank does still induces a ranking function that is appropriate for information retrieval tasks.

With this aim in mind, we used the webTREC Gov2 collection.¹ This collection consists of about 25 million documents obtained in 2004 from a crawl of a large subset of the .gov (U.S. government) domain. The most important characteristic of this collection is that it includes relevance judgments for a set of information retrieval tasks.

We picked 50 of them at random and manually created keyword queries for this evaluation, following the policy used in the standard ad hoc TREC tasks. We then used the Managing Gigabytes for Java (`mg4j`) framework to select from the collection 1000 pages matching each query accordingly and we reordered the query results according to the scores resulting from different link-based ranking strategies.

At this point, we shifted our attention to the LinearRank ranking that uses linear damping to see if LinearRank with a small number of iterations can provide a ranking that is competitive with PageRank. On this graph, the PageRank calculation took 39 iterations to converge on the L1-norm of the difference between two iterations to less than 10^{-6} .

For the evaluation we computed the standard precision and recall measures [Baeza-Yates and Ribeiro-Neto 99] and averaged them across all queries. Precision at result number N (also denoted as “precision at N ” or simply “P@N”) is the fraction of correct results in the first N results returned by the system; the

¹Available from the University of Glasgow for research purposes. For inquiries about this collection, see http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm.

“correct” results in our case are taken from the quality assessments included in this reference collection. This is shown in Figure 19(top).

Another indicator that is usually measured is *recall*, which is the proportion of correct results that the system finds among the total number of correct pages. It is customary to interpolate the precision for different recall levels, and this is shown in Figure 19(bottom).

Of course, using link ranking improves the precision over no ranking at all, and PageRank and LinearRank behave very similarly. For instance, if we compare the PageRank (which requires 39 iterations) with LinearRank at distance 5 (which requires 5 iterations only), we observe that the precision of the first element is 8% better for PageRank, of the first five elements is 17% better for PageRank, but for the first ten elements it is 2% better for LinearRank. From that point over, both rankings are roughly equivalent.

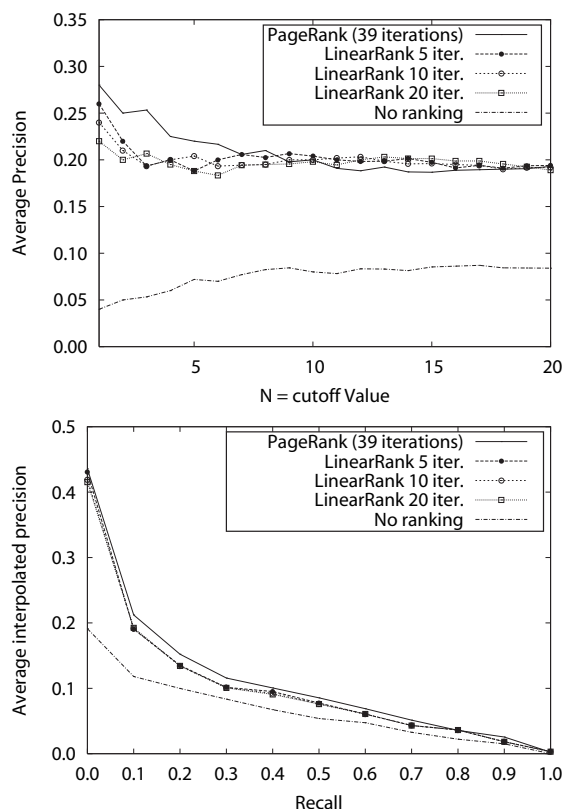


Figure 19. Evaluation of LinearRank and PageRank in the webTREC collection based on precision at N (top) and interpolated precision (bottom).

This means that LinearRank at distance 5 can provide a level of precision for information retrieval tasks that is quite similar to that of PageRank. This is applicable in contexts where link-based rankings cannot be computed in advance but a computation at query time is necessary. For instance, this occurs if we need to analyze links over a subgraph that is generated at query time.

7. Conclusions

In this paper we have defined a broad class of link-based ranking algorithms based on the contribution of damping factors along all the different paths reaching a page. We introduced four particular damping decays: linear, exponential, quadratic hyperbolic, and general hyperbolic, where exponential is equivalent to PageRank and quadratic hyperbolic to TotalRank.

We studied the differences and similarities between these ranking algorithms, and we found that functional rankings using different damping functions can (if the parameters are chosen carefully) provide similar orderings. LinearRank can be used for calculating a ranking that is as good as PageRank for IR tasks. Also, the parameters for the damping functions depend on the characteristic path lengths in the graph, which are known to grow sub-logarithmically on the size of the graph.

More work needs to be done in order to find other damping functions that compute rankings similar to PageRank but are easier and faster to compute. We use a global ranking similarity, but another measure could be the ranking similarity in the top 20 results of real queries. In this setting our results can change, so future work will include this variation.

Because of their high cost, link-based ranking methods that involve iterative calculations at query time are probably not used by large-scale search engines at present, but the functional ranking with linear damping that we have presented can provide a good approximation with few iterations. Moreover, the approach that we have taken could also be applied to multivalued ranking functions such as HITS [Kleinberg 99] and topic-sensitive PageRank [Haveliwala 02] to obtain, for instance, a method for approximating the hubs and authority scores using less iterations and a linear damping function.

Our approach also helps to understand how easy or difficult it is to collude many pages to modify the ranking of a given page. Clearly, there are many different factors: path lengths, damping function, branching degrees, and the number of colluded pages. The graph structure of the collusion will affect those factors, and we plan to analyze them. In particular, under the assumption that is easier to “spam” closer links, PageRank damping is more affected by collusion than the rest of the damping functions presented here. This idea has been further studied by using a truncated exponential damping function for spam detection [Becchetti et al. 06].

We have used damping coefficients that are described by a simple function, but we do not need this restriction. We could learn the best coefficients for an ad-hoc damping function from a given web collection. The aim, in this case, would be to optimize the precision in a sample of queries and their relevant answers, or to accurately filter spam given a sample of spam pages.

Acknowledgments. We would like to thank Dániel Fogaras for a valuable discussion about TotalRank that motivated part of this research. We also thank Karen Whitehouse for her thorough revision of the English. This work was partially supported by the EU IST-FET project 001907 (DELIS), by the EC Project DELIS, and by MIUR COFIN Project “Linguaggi formali e automi.” The authors also thank the support from ICREA and the Càtedra Telefónica at Universitat Pompeu Fabra.

References

- [Albert et al. 99] Réka Albert, Hawoong Jeong, and Albert L. Barabási. “Diameter of the World Wide Web.” *Nature* 401 (1999), 130–131.
- [Arasu et al. 01] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. “Searching the Web.” *ACM Transactions on Internet Technology (TOIT)* 1:1 (2001), 2–43.
- [Baeza-Yates and Davis 04] Ricardo Baeza-Yates and Emilio Davis. “Web Page Ranking Using Link Attributes.” In *Proceedings of the 13th International World Wide Web Conference, Alternate Track Papers & Posters*, pp. 328–329. New York: ACM Press, 2004.
- [Baeza-Yates and Ribeiro-Neto 99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Reading, MA: Addison Wesley, 1999.
- [Baeza-Yates et al. 04] Ricardo Baeza-Yates, Carlos Castillo, and Felipe S. Jean. “Web Dynamics, Structure and Page Quality.” In *Web Dynamics*, edited by M. Levene and A. Pulovassilis, pp. 93–109. New York: Springer, 2004.
- [Baeza-Yates et al. 06] Ricardo Baeza-Yates, Paolo Boldi, and Carlos Castillo. “Generalizing PageRank: Damping Functions for Link-Based Ranking Algorithms.” In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 308–315. New York: ACM Press, 2006.
- [Barabási 02] Albert-László Barabási. *Linked: The New Science of Networks*. New York: Perseus Books Group, 2002.
- [Barabási and Albert 99] A. L. Barabási and R. Albert. “Emergence of Scaling in Random Networks.” *Science* 286:5439 (1999), 509–512.
- [Becchetti et al. 06] Luca Becchetti, Carlos Castillo, Debora Donato, Stefano Leonardi, and Ricardo Baeza-Yates. “Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection.” Paper presented at the Workshop on

- Web Mining and Web Usage Analysis (WebKDD), Pennsylvania, August 20–23, 2006. Available at http://www.dcc.uchile.cl/~ccastill/papers/becchetti_06-automatic_link_spam_detection_rank_propagation.pdf.
- [Bharat and Henzinger 98] Krishna Bharat and Monika R. Henzinger. “Improved Algorithms for Topic Distillation in a Hyperlinked Environment.” In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 104–111. New York: ACM Press, 1998.
- [Boldi 05] Paolo Boldi. “TotalRank: Ranking without Damping.” In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pp. 898–899. New York: ACM Press, 2005.
- [Boldi et al. 05] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. “PageRank as a Function of the Damping Factor.” In *Proceedings of the 14th International Conference on World Wide Web*, pp. 557–566. New York: ACM Press, 2005.
- [Bollobás and Riordan 04] Béla Bollobás and Oliver Riordan. “The Diameter of a Scale-Free Random Graph.” *Combinatorica* 24:1 (2004), 5–34.
- [Borodin et al. 05] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. “Link Analysis Ranking: Algorithms, Theory, and Experiments.” *ACM Trans. Inter. Tech.* 5:1 (2005), 231–297.
- [Brinkmeier 06] Michael Brinkmeier. “PageRank Revisited.” *ACM Transaction on Internet Technologies* 6:3 (2006), 257–279.
- [Broder et al. 00] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. “Graph Structure in the Web: Experiments and Models.” In *Proceedings of WWW9, Computer Networks* 33:1–6 (2000), 309–320.
- [Callaway et al. 00] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. “Network Robustness and Fragility: Percolation on Random Graphs.” *Physical Review Letters* 85:25 (2000), 5468–5471.
- [Carrière and Kazman 97] Jeromy Carrière and Rick Kazman. “WebQuery: Searching and Visualizing the Web through Connectivity.” *Computer Networks and ISDN Systems* 29:8 (1997), 1257–1267.
- [Chen et al. 04] Yen-Yu Chen, Qingqing Gan, and Torsten Suel. “Local Methods for Estimating PageRank Values.” In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp. 381–389. New York: ACM Press, 2004.
- [Chung and Lu 01] Fan Chung and Linyuan Lu. “The Diameter of Random Sparse Graphs.” *Adv. Appl. Math.* 26 (2001), 257–279.
- [Davison 00] Brian D. Davison. “Topical Locality in the Web.” In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 272–279. New York: ACM Press, 2000.
- [Donato et al. 05] Debora Donato, Stefano Leonardi, Stefano Millozzi, and Panayiotis Tsaparas. “Mining the Inner Structure of the Web Graph.” In *Proceedings of the Eighth International Workshop on the Web and Databases*, pp. 145–150, 2005. Available at <http://webdb2005.uhasselt.be/papers/P-9.pdf>.

- [Dwork et al. 01] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. “Rank Aggregation Methods for the Web.” In *Proceedings of the Tenth International Conference on World Wide Web*, pp. 613–622. New York: ACM Press, 2001.
- [Eiron et al. 04] Nadav Eiron, Kevin S. Curley, and John A. Tomlin. “Ranking the Web Frontier.” In *Proceedings of the 13th International Conference on World Wide Web*, pp. 309–318. New York: ACM Press, 2004.
- [Erdős and Rényi 60] Paul Erdős and Alfred Rényi. “On the Evolution of Random Graphs.” *Publication of the Mathematical Institute of the Hungarian Academy of Science* 5 (1960), 17–61.
- [Fagin et al. 03a] Ronald Fagin, Ravi Kumar, Kevin S. McCurley, Jasmine Novak, D. Sivakumar, John A. Tomlin, and David P. Williamson. “Searching the Workplace Web.” In *Proceedings of the 12th International Conference on World Wide Web*, pp. 366–375. New York: ACM Press, 2003.
- [Fagin et al. 03b] Ronald Fagin, Ravi Kumar, and D. Sivakumar. “Comparing Top k Lists.” In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 28–36. Philadelphia: Society for Industrial and Applied Mathematics, 2003.
- [Fogaras 03] Dniel Fogaras. “Where to Start Browsing the Web?” In *Innovative Internet Community Systems*, Lecture Notes in Computer Science 2877, pp. 65–79. Berlin: Springer, 2003.
- [Golub and Greif 04] Gene H. Golub and Chen Greif. “Arnoldi-type Algorithms for Computing Stationary Distribution Vectors, with Application to PageRank.” Technical Report SCCM-04-15, Stanford University, 2004.
- [Gulli and Signorini 05] Antonio Gulli and Alessio Signorini. “The Indexable Web is More than 11.5 Billion Pages.” In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pp. 902–903. New York: ACM Press, 2005.
- [Haas and Grams 98] Stephanie W. Haas and Erika S. Grams. “Page and Link Classifications: Connecting Diverse Resources.” In *Proceedings of the Third ACM Conference on Digital Libraries*, pp. 99–107. New York: ACM Press, 1998.
- [Haveliwala and Kamvar 03a] Taher Haveliwala and Sepandar Kamvar. “The Condition Number of the PageRank Problem.” Technical Report 36, Stanford University, 2003.
- [Haveliwala and Kamvar 03b] Taher Haveliwala and Sepandar Kamvar. “The Second Eigenvalue of the Google Matrix.” Technical Report 20, Stanford University, 2003.
- [Haveliwala 99] Taher Haveliwala. “Efficient Computation of PageRank.” Technical report, Stanford University, 1999.
- [Haveliwala 02] Taher H. Haveliwala. “Topic-Sensitive PageRank.” In *Proceedings of the 11th International Conference on World Wide Web*, pp. 517–526. New York: ACM Press, 2002.
- [Huberman 01] Bernardo A. Huberman. *The Laws of the Web: Patterns in the Ecology of Information*. Cambridge, MA: MIT Press, 2001.

- [Joo and Myaeng 98] W-K Joo and S. H. Myaeng. “Improving Retrieval Effectiveness with Hyperlink Information.” Paper presented at the International Workshop on Information Retrieval with Asian Languages, Singapore, October 15–16, 1998.
- [Kamvar et al. 03a] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. “Exploiting the Block Structure of the Web for Computing PageRank.” Technical report, Stanford University, 2003.
- [Kamvar et al. 03b] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. “Extrapolation Methods for Accelerating PageRank Computations.” In *Proceedings of the 12th International Conference on World Wide Web*, pp. 261–270. New York: ACM Press, 2003.
- [Katz 53] L. Katz. “A New Status Index Derived from Sociometric Analysis.” *Psychometrika* 18 (1953), 39–43.
- [Kendall and Gibbons 90] Maurice Kendall and Jean Dickinson Gibbons. *Rank Correlation Methods*, Fifth Edition. New York: Oxford University Press, 1990.
- [Kleinberg 99] Jon M. Kleinberg. “Authoritative Sources in a Hyperlinked Environment.” *Journal of the ACM* 46:5 (1999), 604–632.
- [Kumar et al. 00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. “Stochastic Models for the Web Graph.” In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pp. 57–65. Los Alamitos, CA: IEEE Press, 2000.
- [Lee et al. 04] Chris P. Lee, Gene H. Golub, and Stefanos A. Zenios. “A Fast Two-Stage Algorithm for Computing PageRank and Its Extensions.” Technical report, Stanford University, 2004.
- [Lempel and Moran 01] R. Lempel and S. Moran. “SALSA: The Stochastic Approach for Link-Structure Analysis.” *ACM Trans. Inf. Syst.* 19:2 (2001), 131–160.
- [Leskovec et al. 05] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. “Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations.” In *Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 177–187. New York: ACM Press, 2005.
- [Li 98] Yanhong Li. “Toward a Qualitative Search Engine.” *IEEE Internet Computing* 2:4 (1998), 24–29.
- [Lifantsev 00] Maxim Lifantsev. “Voting Model for Ranking Web Pages.” In *Proceedings of the International Conference on Internet Computing*, edited by Peter Graham and Muthucumaru Maheswaran, pp. 143–148. CSREA Press, 2000.
- [Liu and Ma 05] Tie-Yan Liu and Wei-Ying Ma. “Webpage Importance Analysis Using Conditional Markov Random Walk.” In *The IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 515–521. New York: ACM Press, 2005.
- [Marchiori 97] Massimo Marchiori. “The Quest for Correct Information of the Web: Hyper Search Engines.” *Proceedings of WWW6, Computer Networks and ISDN Systems* 29:8–13 (1997), 1225–1236.
- [Menczer 04] Filippo Menczer. “Lexical and Semantic Clustering by Web Links.” *Journal of the American Society for Information Science and Technology* 55:14 (2004), 1261–1269.

- [Newman et al. 01] M. E. Newman, S. H. Strogatz, and D. J. Watts. “Random Graphs with Arbitrary Degree Distributions and Their Applications.” *Phys. Rev. E (Part 2: Stat. Nonlin. Phys.)* 64:2 (2001), 026118.
- [Page et al. 98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. “The PageRank Citation Ranking: Bringing Order to the Web.” Technical report, Stanford Digital Library Technologies Project, 1998.
- [Pandurangan et al. 02] Gopal Pandurangan, Prabhakara Raghavan, and Eli Upfal. “Using PageRank to Characterize Web Structure.” In *Computing and Combinatorics: 8th Annual International Conference, COCOON 2002, Singapore, August 15–17, 2002, Proceedings*, Lecture Notes in Computer Science 2387, pp. 330–390. Berlin: Springer, 2002.
- [Rudin 86] Walter Rudin. *Real and Complex Analysis*. New York: McGraw-Hill, 1986.
- [Srinivasan et al. 05] Padmini Srinivasan, Gautam Pant, and Filippo Menczer. “A General Evaluation Framework for Topical Crawlers.” *Information Retrieval* 8:3 (2005), 417–447.
- [Tauro et al. 01] L. Tauro, C. Palmer, G. Sigamos, and M. Faloutsos. “A Simple Conceptual Model for the Internet Topology.” In *Proceedings of the IEEE Global Telecommunications Conference GLOBECOM '01*, Vol. 3, pp. 1667–1671. Los Alamitos, CA: IEEE Press, 2001.
- [Upstill et al. 03] Trystan Upstill, Nick Craswell, and David Hawking. “Predicting Fame and Fortune: PageRank or Indegree?” In *Proceedings of the Australasian Document Computing Symposium, ADCS2003*, pp. 31–40. Canberra, Australia: Australian Computer Society, 2003.
- [Watts and Strogatz 98] Duncan J. Watts and Steven H. Strogatz. “Collective Dynamics of Small-World Networks.” *Nature* 393:6684 (1998), 440–442.
- [Wu et al. 04] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. “Information Flow in Social Groups.” *Physica A: Statistical and Theoretical Physics* 337:1-2 (2004), 327–335.

Ricardo Baeza-Yates, Yahoo! Research, Ocata 1, 08003 Barcelona, Spain
(rby@yahoo-inc.com)

Paolo Boldi, Università degli Studi di Milano, Dipartimento di Scienze dell’ Informazione, via Comelico 39/41, 20135 Milan, Italy (boldi@dsi.unimi.it)

Carlos Castillo, Università di Roma, “La Sapienza,” Dipartimento di Informatica e Sistemistica, via Ariosto 25, 00185 Rome, Italy (chato@yahoo-inc.com)

Received March 1, 2007; accepted July 5, 2007.