Generic Damping Functions for Propagating Importance in Link-Based Ranking

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain
& Santiago, Chile

Paolo Boldi*
Università degli Studi
di Milano
Milan, Italy

Carlos Castillo†
Università di Roma
"La Sapienza"
Rome, Italy

Abstract

This paper introduces a family of link-based ranking algorithms that propagate page importance through links. The algorithms include a damping function which decreases with distance, thus a direct link implies greater endorsement that 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 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.

*Partially supported by MIUR COFIN Project "Linguaggi formali e automi" and by the EC Project DELIS.

†Currently at Yahoo! Research Barcelona.

1 Introduction

While traditional Information Retrieval (IR) methods are used by web search engines to some extent, the web is much more extensive, dynamic and less coherent than traditional text collections [Arasu et al., 2001]. 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, 2000].

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. [Broder et al., 2000], in which a global structure was described based on the presence of a large strongly connected component. 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, described in [Donato et al., 2005, Baeza-Yates et al., 2004].

A related macroscopic description is the *Jellyfish structure* described in [Tauro et al., 2001] for autonomous systems in the Internet topology. According to this 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, 2001, Barabási, 2002], and is based on the observation that the distribution of the degree on the Web is very skewed, not showing the typical Poisson distribution observed in classical random graphs [Erdõs and Rényi, 1960]. In scale-free networks, such as the Web, the distribution of the number of links of a page p follows a power-law:

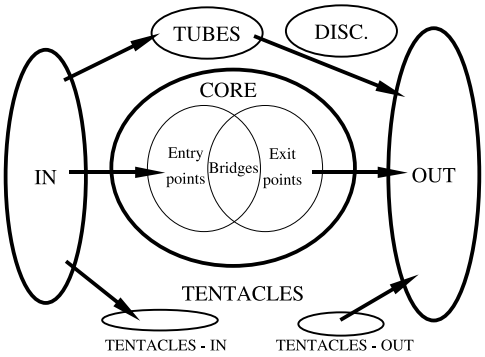

Figure 1: Schematic depiction of the macroscopic "bow-tie" structure of the Web [Broder et al., 2000].

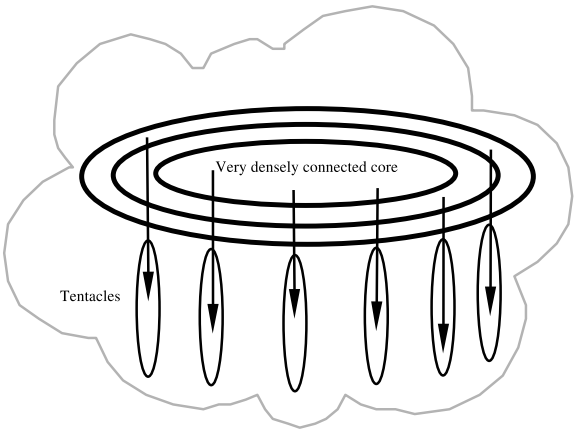

Figure 2: Schematic depiction of the macroscopic "jellyfish" structure of the Internet [Tauro et al., 2001].

$$Pr(\text{page } p \text{ has } k \text{ links}) \propto k^{-\theta} \tag{1}$$

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., 2000], and can be explained in part by a "preferential attachment" process [Barabási and Albert, 1999], 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.

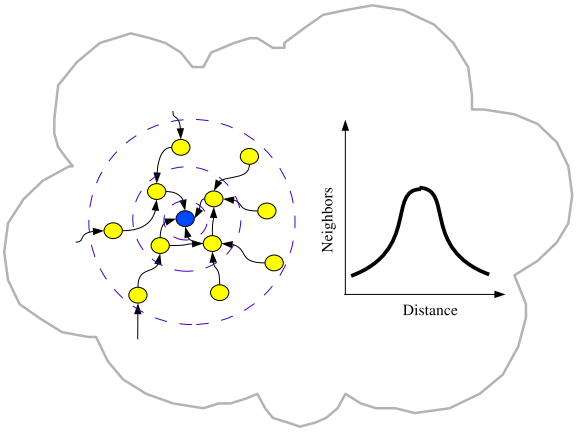

Figure 3: Schematic depiction of the “hop-plot”: a plot of the number of neighbors at different distances.

The class of functional rankings which 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.

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 we have described.

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 are of the main problems of Web IR, thus requiring a method of estimating relevance.

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 links for ranking web pages [Marchiori, 1997, Joo and Myaeng, 1998, Li, 1998]; however, it quickly become 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

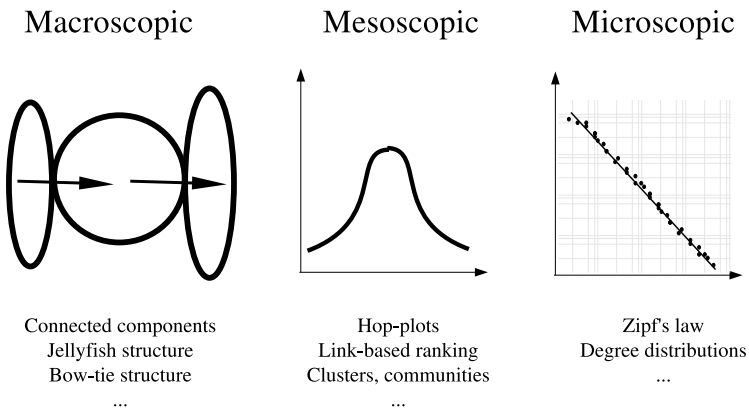

Figure 4: Levels of link-based analysis.

easy to manipulate in the context of the web, where creating a page costs almost nothing.

The PageRank technique, introduced by Page *et al.* [Page et al., 1998], 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". 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 web pages, 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. There are other well-known methods for link-based ranking that we do not discuss here, such as HITS [Kleinberg, 1999, Bharat and Henzinger, 1998] or SALSA [Lempel and Moran, 2001]; for a survey of them see [Borodin et al., 2005].

1.3 Our contribution

In this paper we describe general ranking functions that depend on incoming paths of varying length, and 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 do 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 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., 2006].

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 pages crawl of the `.uk` domain in 2002, a 40-million pages crawl of the `.it` domain in 2004 , `.it` and a 860,000-pages 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.* [Kumar et al., 2000] (a combination of preferential attachment and random links) with parameters suggested by Pandurangan *et al.* [Pandurangan et al., 2002]. 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-nodes graph without disconnected nodes.

To compare ranking orders among different ranking functions, we used Kendall's τ [Kendall and Gibbons, 1990]: this is one of the most widely used and intuitive nonparametric correlation indices, that has recently received much attention within the web community for its possible applications to rank aggregation [Fagin et al., 2003b, Fagin et al., 2003a, Dwork et al., 2001] and for determining the convergence speed in the computation of PageRank [Kamvar et al., 2003b]. 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, that 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, $a_{i,j} = 1$ iff there is a link from page i to page j. This link matrix is hardly ever used as it is, mainly as 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 web page authors to create pages with many out-links; this is the reason why a metaphor of "voting" is enforced [Lifantsev, 2000] 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 web page can only decide how to divide its own score among the pages it leads to, 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, 1998]. 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, 1998] that treat them differently. Other characteristics of links, such as the exploration level at which they appear in Web sites [Liu and Ma, 2005], 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, 2004].

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., 2004]. 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 a 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 observed in [Fogaras, 2003, Boldi et al., 2005], 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 matricial form:

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

There is, in fact, an equivalent, very intriguing way of rewriting this formula, mentioned in [Newman et al., 2001] that leads to a conclusion similar to those of [Brinkmeier, 2006]: given a path, that is, a sequence of edges in the graph $p = \langle x_1, x_2, \ldots, x_k\rangle$, 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 outdegree, 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 $\mathrm{Path}(-,i)$ 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:

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

or equivalently

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

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 [Lifantsev, 2000] model in which the damping factor is a matrix of *voting trust* that is fixed during the computation, 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 $\mathrm{damping}(t)\frac{1}{N}\mathbf{1}$; in this paper we analyze only the latter form. Fogaras [Fogaras, 2003] 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. 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

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

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

3.3 Characteristic path lengths

In scale-free networks, the distances between pairs of nodes follow a Gaussian distribution [Albert et al., 1999]. Analytic estimations for the average

distance of a graph of scale-free network of n nodes include: $O(\log(n))$ [Watts and Strogatz, 1998]; $O(\log(n) \ / \log(np))$ in sparse graphs with p links [Chung and Lu, 2001]; $1 + \log(n/z_1)/\log(z_2/z_1)$ where z_1 is the average in-degree, and z_2 is the average number of nodes at distance 2 [Newman et al., 2001]; and $O(\log(n)/\log(\log(n)))$ [Bollobás and Riordan, 2004].

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., 2005].

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 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.

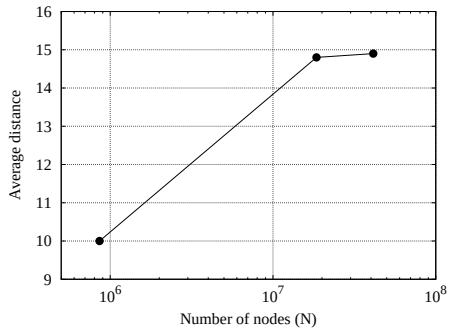

Figure 5: Average distances versus number of nodes in four Web graphs.

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, e-mail 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.

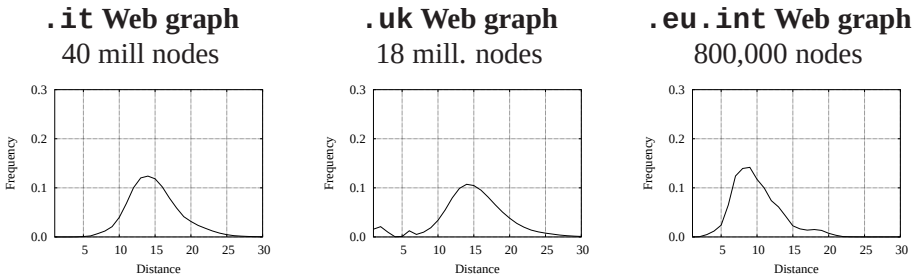

Figure 6: Distribution of the average number of nodes at a certain distance from a given node, in three Web samples.

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 weights 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:

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

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, 2005]), 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 3 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 web pages 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. As shown in [Brinkmeier, 2006, Corollary 2.4], for every pair of nodes i and j, and for every length t

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

A more general property holds:

Theorem 1. *For every node i and every length t*

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

Proof. 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 outdegree 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:

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

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

∎

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 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.

Computation. For calculating functional rankings, we use the general algorithm shown in Figure 7; the next sections provide details on the initialization, stop condition and iteration steps for each calculation.

```
Require: N: number of nodes, v: preference vector
 1: for i : 1 ... N do {Initialization}
 2:    S[i] ← R[i] ← START
 3: end for
 4: for k : 1 ... ∞ do {Iteration step}
 5:    if STOP then
 6:       break
 7:    end if
 8:    Aux ← 0
 9:    for i : 1 ... N do {Follow links in the graph}
10:       for all j such that there is a link from i to j do
11:          Aux[j] ← Aux[j] + R[i]/outdegree(i)
12:       end for
13:    end for
14:    for i : 1 ... N do {Add to ranking value}
15:       R[i] ← Aux[i] × DAMP(k)
16:       S[i] ← 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.

4.1 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 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.

Computation. 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:

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

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. In [Chen et al., 2004], it is 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. 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.

Computation. 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.* in their original paper [Page et al., 1998]; 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, 2003b, Haveliwala and Kamvar, 2003a]. Other methods to compute PageRank have been proposed, some of them using techniques for the solution of systems of linear equations, some other 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., 1998, Haveliwala, 1999, Golub and Greif, 2004, Lee et al., 2004, Kamvar et al., 2003c, Kamvar et al., 2003a]).

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

$$\begin{aligned} \text{START} &: (1-\alpha)v[i] \\ \text{STOP} &: \text{convergence} \\ \text{DAMP(k)} &: \alpha \end{aligned}$$

4.3 Quadratic hyperbolic damping: TotalRank

Recently, a ranking method called TotalRank [Boldi, 2005] has been proposed. 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:

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

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 applying Theorem 1.27 of [Rudin, 1986].

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

Computation. 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., 2005], 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 done 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.

Computation. 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 which are linked to each other share a greater degree of similarity than pages chosen at random [Davison, 2000]; evidence from topical crawlers [Srinivasan et al., 2005] 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. In [Menczer, 2004] it is shown that text similarity and link distance are anti-correlated up to 4-5 links.

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, 1999], without stemming or stop-word removal.

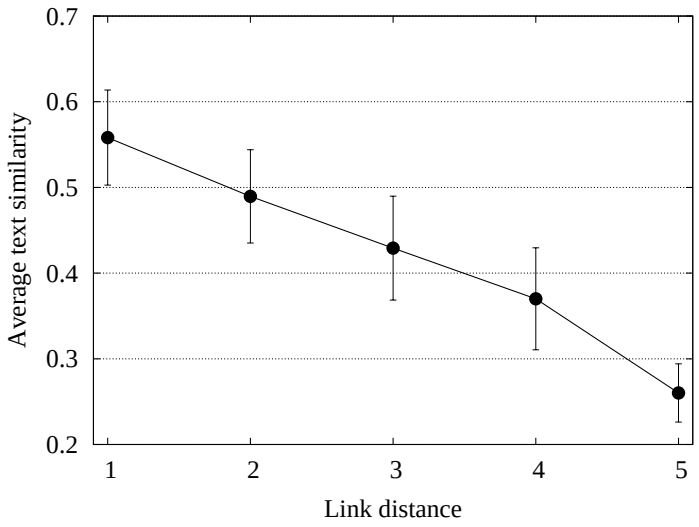

Figure 8: Link distance vs. average text similarity in a sample of 18 million pages from the .uk domain. A link distance of one means direct linking. The text similarity appears to decrease linearly in the first few levels.

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., 2004] uses the number of repeated words and phrases between pages and obtains 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, 2004]. 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 in [Menczer, 2004] is that for different communities, the link structure could be different. For instance, academic Web pages 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 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.

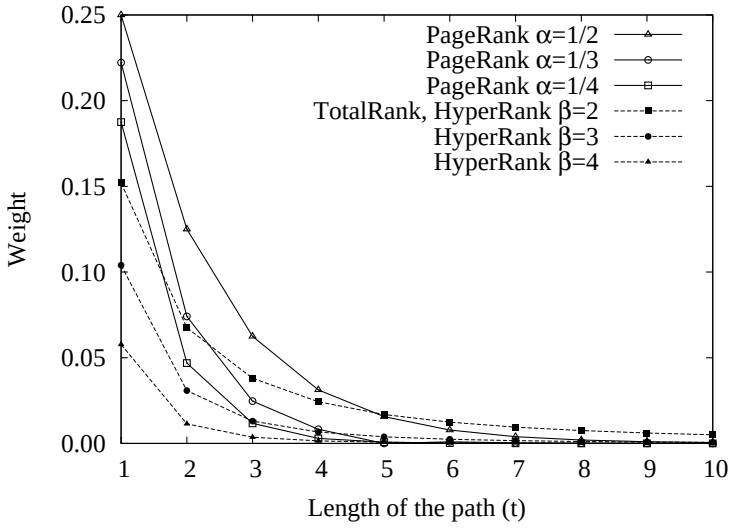

Figure 9: Weights given by the different damping functions for some values of α and β.

We can empirically observe that the ranking ordering 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.

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, 2005]; 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:

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

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 we are interested in, 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, has been estimated in about 16 [Broder et al., 2000] or 19 [Albert et al., 1999], but clearly today 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.

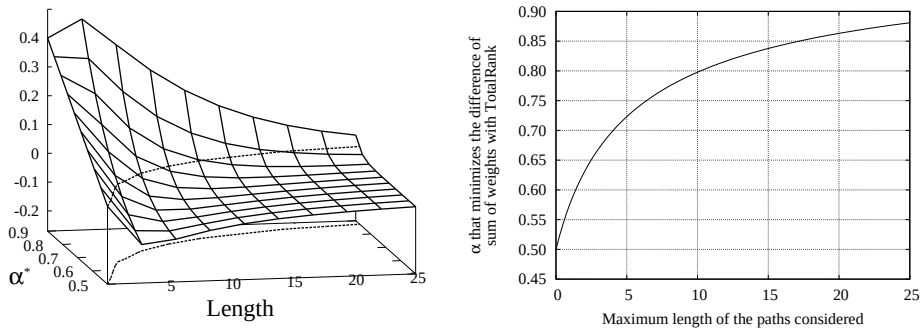

Figure 10: Left: difference of the sum of the weights for various combinations of α and ℓ. Right: $\alpha^*(\ell)$ for minimizing the difference of the sum of weights between PageRank and TotalRank.

5.2 Approximating PageRank with HyperRank

Now we want to approximate the weights of:

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

using the weights of:

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

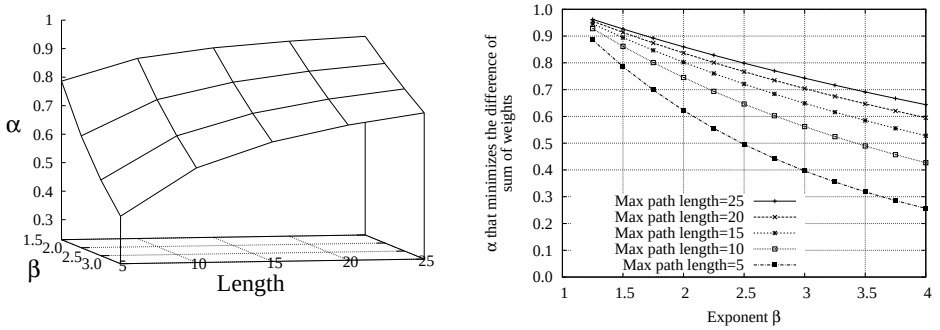

Figure 11: Left: best α for minimizing the difference of the sum of weights between PageRank and HyperRank, for various parameter combinations. Right: 2-D view of the same plot.

and 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:

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

The α that minimizes the difference of weights 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.

5.4 Experimental comparison of ranking orders

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

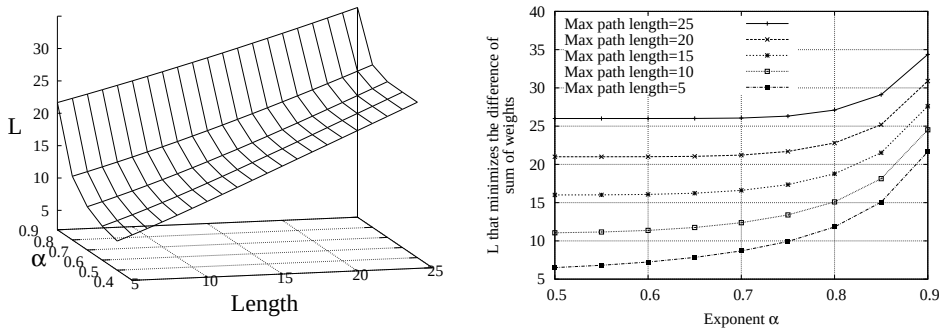

Figure 12: Left: best L for minimizing the difference of the sum of weights between LinearRank and PageRank, for various parameter combinations. Right: 2-D view of the plot on the left.

Figure 13 shows how PageRank compares with HyperRank for various pairs of α and β. In the limit $\alpha, \beta \to 1$ both rankings are equivalent, and they remain similar in a large region of the parameter space.

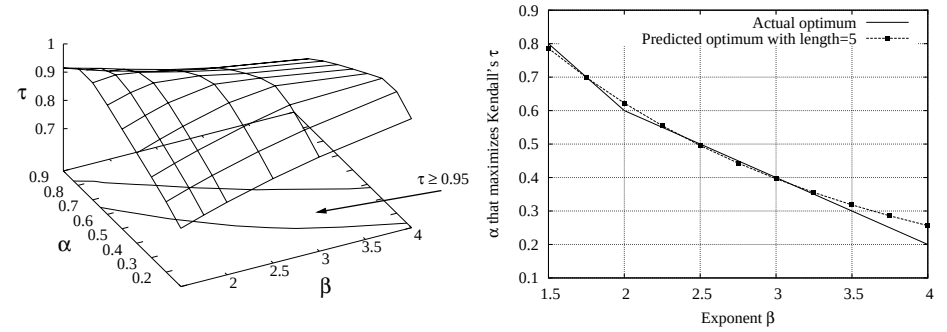

Figure 13: Comparison (using Kendall's τ) between PageRank and HyperRank, with various damping parameters in the U.K. web graph. The optimum predicted in the analysis with $\ell = 5$ is very close to the real one.

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 proposed in [Fagin et al., 2003b]: 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.

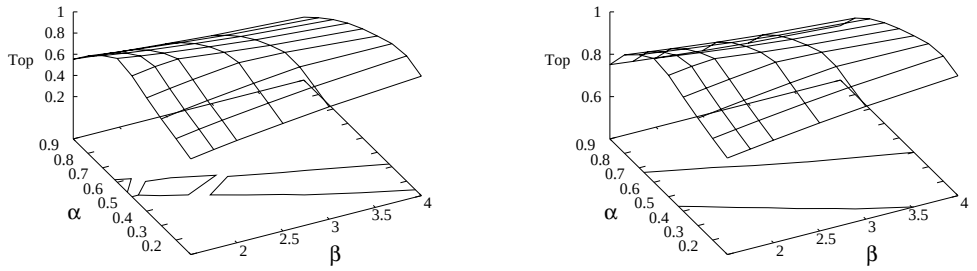

Figure 14: Comparison (using top-k intersection metric) between PageRank and HyperRank, with various damping parameters in the U.K. web graph, for $k = 1\,000$ and $k = 100\,000$.

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 $\alpha = 0.8 \ldots 0.9$, paths of roughly 10 to 20 links should be considered to obtain rankings that are almost equivalent.

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 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

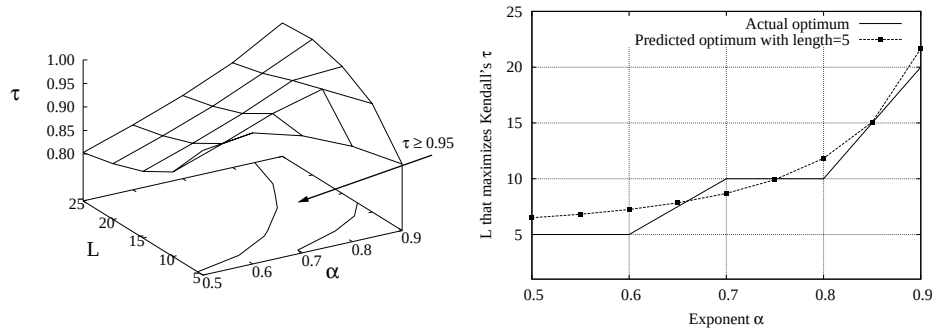

Figure 15: Comparison (using Kendall's τ) between PageRank and LinearRank in the U.K. web graph, with various damping parameters. Again, the predicted optimum with $\ell = 5$ is very close to the actual optimum.

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 precision of both ranking functions for Information Retrieval tasks, is a problem that merits further experiments and investigation.

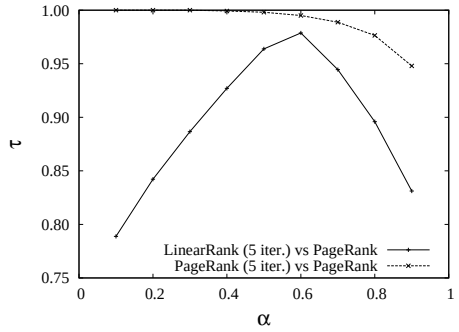

Figure 16: Comparison (using Kendall's τ) between PageRank stopped after 5 iterations and LinearRank with $\ell = 5$ in the U.K. web graph, with various damping parameters.

5.5 Comparison with in-degree

In this section, we study the behavior of the ranking functions for different values of their parameters.

In this section, we are using data from the .uk Web graph and a 8,500-nodes 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 correlation coefficient and ranking orders (Kendall's τ). The results are shown in Figure 17.

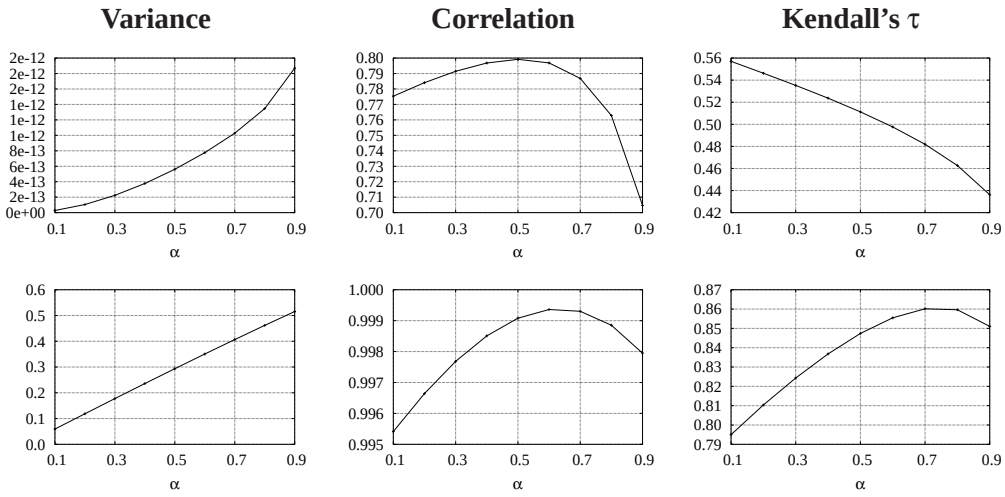

Figure 17: Variance of PageRank, and its relationship with in-degree in terms of correlation coefficient and Kendall's τ coefficient, for varying values of the parameter α. Top: .uk web graph; bottom: synthetic graph.

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., 2003]. Our results are consistent with this observation. Not surprisingly, using the generative model the correlations are higher. 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, however, 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 high correlation was foreseeable because, as shown in [Chen et al., 2004], even approximating PageRank with just only 1 level of links gets 70% of accuracy.

The behavior of the Kendall's τ coefficient which measures the similarity between ranking orders is the opposite than 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).

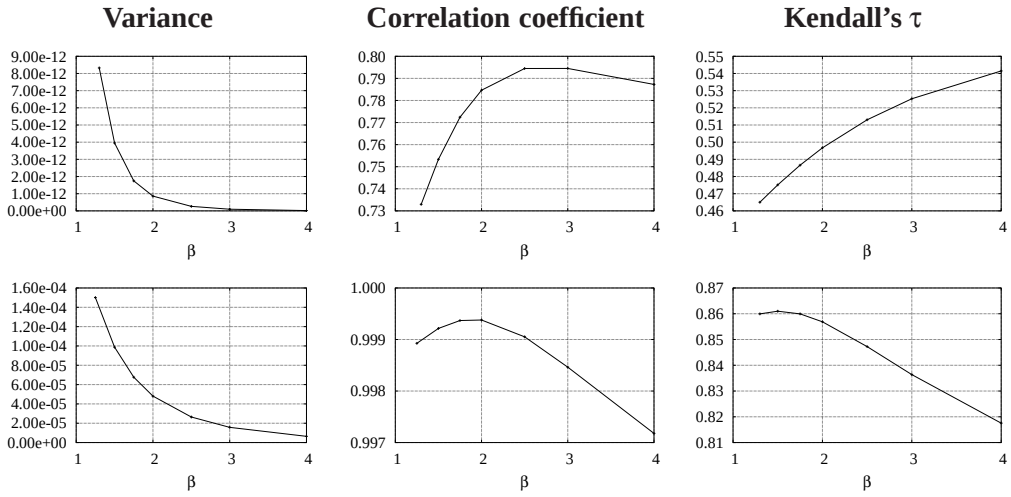

Figure 18: Variance of general hyperbolic rank for some values of β, and its relation with in-degree. Experiments have been performed on the `.uk` graph (top) and synthetic graph (bottom).

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 which 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 according and we re-ordered the query results according to the scores resulting from different link-based rankings 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, 1999] 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 "correct" results in our case are taken from the quality assessments included in this reference collection. This is shown in Figure 19 (a).

Another indicator that is usually measured is recall, which is the proportion of correct results which 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 (b).

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 (that requires 39 iterations) with LinearRank at distance 5 (that 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.

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 ranking 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 sub-graph that is generated at query time.

[1] Available 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`.

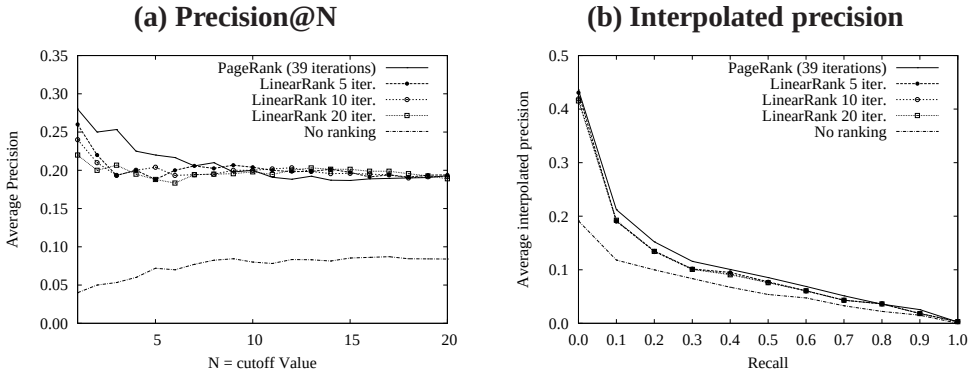

Figure 19: Evaluation of LinearRank and PageRank in the WebTREC collection.

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 which we have presented can provide a good approximation with few iterations. Moreover, the approach we have taken could be also applied to multivalued ranking functions such as HITS [Kleinberg, 1999] and topic-sensitive PageRank [Haveliwala, 2002] 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 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 is further studied in [Becchetti et al., 2006] by using a truncated exponential damping function for spam detection.

We have use 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 to Karen Whitehouse for valuable feedback on a previous version of this paper.

The authors also thank the support from ICREA and the Cátedra Telefónica at Universitat Pompeu Fabra. This work was partially supported by the EU IST-FET project 001907 (DELIS).

References

[Albert et al., 1999] Albert, R., Jeong, H., and Barabási, A. L. (1999). Diameter of the world wide web. *Nature*, 401:130–131.

[Arasu et al., 2001] Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Raghavan, S. (2001). Searching the web. *ACM Transactions on Internet Technology (TOIT)*, 1(1):2–43.

[Baeza-Yates et al., 2006] Baeza-Yates, R., Boldi, P., and Castillo, C. (2006). Generalizing pagerank: Damping functions for link-based ranking algorithms. In *Proceedings of ACM SIGIR*, pages 308–315, Seattle, Washington, USA. ACM Press.

[Baeza-Yates et al., 2004] Baeza-Yates, R., Castillo, C., and Jean, F. S. (2004). *Web Dynamics*, chapter Web Dynamics, Structure and Page Quality, pages 93–109. Springer.

[Baeza-Yates and Davis, 2004] Baeza-Yates, R. and Davis, E. (2004). Web page ranking using link attributes. In *Alternate track papers & posters of the 13th international conference on World Wide Web*, pages 328–329, New York, NY, USA. ACM Press.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

[Barabási, 2002] Barabási, A.-L. (2002). *Linked: The New Science of Networks*. Perseus Books Group.

[Barabási and Albert, 1999] Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[Becchetti et al., 2006] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. (2006). Using rank propagation and probabilistic counting for link-based spam detection. In *Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD)*, Pennsylvania, USA. ACM Press.

[Bharat and Henzinger, 1998] Bharat, K. and Henzinger, M. R. (1998). 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*, pages 104–111, Melbourne, Australia. ACM Press, New York.

[Boldi, 2005] Boldi, P. (2005). Totalrank: ranking without damping. In *Poster proceedings of the 14th international conference on World Wide Web*, pages 898–899, Chiba, Japan. ACM Press.

[Boldi et al., 2005] Boldi, P., Santini, M., and Vigna, S. (2005). Pagerank as a function of the damping factor. In *Proceedings of the 14th international conference on World Wide Web*, pages 557–566, Chiba, Japan. ACM Press.

[Bollobás and Riordan, 2004] Bollobás, B. and Riordan, O. (2004). The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34.

[Borodin et al., 2005] Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P. (2005). Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, 5(1):231–297.

[Brinkmeier, 2006] Brinkmeier, M. (2006). PageRank revisited. *ACM Transaction on Internet Technologies*, 6(3):257–279.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web: Experiments and models. In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands. ACM Press.

[Callaway et al., 2000] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2000). Network Robustness and Fragility: Percolation on Random Graphs. *Physical Review Letters*, 85(25):5468–5471.

[Chen et al., 2004] Chen, Y.-Y., Gan, Q., and Suel, T. (2004). Local methods for estimating pagerank values. In *Proceedings of the thirteenth ACM conference on Information and knowledge management (CIKM)*, pages 381–389, New York, NY, USA. ACM Press.

[Chung and Lu, 2001] Chung, F. and Lu, L. (2001). The diameter of random sparse graphs. *Adv. Appl. Math.*, 26:257–279.

[Davison, 2000] Davison, B. D. (2000). Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 272–279, Athens, Greece. ACM Press.

[Donato et al., 2005] Donato, D., Leonardi, S., Millozzi, S., and Tsaparas, P. (2005). Mining the inner structure of the web graph. In *Eigth international workshop on the Web and databases WebDB*, Baltimore, USA.

[Dwork et al., 2001] Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the tenth international conference on World Wide Web*, pages 613–622. ACM Press.

[Eiron et al., 2004] Eiron, N., Curley, K. S., and Tomlin, J. A. (2004). Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA. ACM Press.

[Erdõs and Rényi, 1960] Erdõs, P. and Rényi, A. (1960). Random graphs. *Publication of the Mathematical Institute of the Hungarian Acadamy of Science*, 5.

[Fagin et al., 2003a] Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A., and Williamson, D. P. (2003a). Searching the workplace

web. In *Proceedings of the twelfth international conference on World Wide Web*, pages 366–375. ACM Press.

[Fagin et al., 2003b] Fagin, R., Kumar, R., and Sivakumar, D. (2003b). Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 28–36. Society for Industrial and Applied Mathematics.

[Fogaras, 2003] Fogaras, D. (2003). Where to start browsing the web? In *Innovative Internet Community Systems*, volume 2877 of *Lecture Notes in Computer Science*, pages 65–79, Leipzig, Germany.

[Golub and Greif, 2004] Golub, G. H. and Greif, C. (2004). . Technical Report SCCM-04-15, Stanford University.

[Gulli and Signorini, 2005] Gulli, A. and Signorini, A. (2005). The indexable Web is more than 11.5 billion pages. In *Poster proceedings of the 14th international conference on World Wide Web*, pages 902–903, Chiba, Japan. ACM Press.

[Haas and Grams, 1998] Haas, S. W. and Grams, E. S. (1998). Page and link classifications: connecting diverse resources. In *DL '98: Proceedings of the third ACM conference on Digital libraries*, pages 99–107, New York, NY, USA. ACM Press.

[Haveliwala, 1999] Haveliwala, T. (1999). Efficient computation of pagerank. Technical report, Stanford University.

[Haveliwala and Kamvar, 2003a] Haveliwala, T. and Kamvar, S. (2003a). The condition number of the pagerank problem. Technical Report 36, Stanford University.

[Haveliwala and Kamvar, 2003b] Haveliwala, T. and Kamvar, S. (2003b). The second eigenvalue of the google matrix. Technical Report 20, Stanford University.

[Haveliwala, 2002] Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proceedings of the Eleventh World Wide Web Conference*, pages 517–526, Honolulu, Hawaii, USA. ACM Press.

[Huberman, 2001] Huberman, B. A. (2001). *The Laws of the Web: Patterns in the Ecology of Information*. The MIT Press.

[Joo and Myaeng, 1998] Joo, W.-K. and Myaeng, S. H. (1998). Improving retrieval effectiveness with hyperlink information. In *Proceedings of International Workshop on Information Retrieval with Asian Languages (IRAL)*, Singapore.

[Kamvar et al., 2003a] Kamvar, S., Haveliwala, T., Manning, C., and Golub, G. (2003a). Exploiting the block structure of the web for computing pagerank.

[Kamvar et al., 2003b] Kamvar, S. D., Haveliwala, T. H., Manning, C. D., and Golub, G. H. (2003b). Extrapolation methods for accelerating pagerank computations. In *Proceedings of the twelfth international conference on World Wide Web*, pages 261–270. ACM Press.

[Kamvar et al., 2003c] Kamvar, S. D., Haveliwala, T. H., Manning, C. D., and Golub, G. H. (2003c). Extrapolation methods for accelerating pagerank computations. In *Proceedings of the twelfth international conference on World Wide Web*, pages 261–270. ACM Press.

[Katz, 1953] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43.

[Kendall and Gibbons, 1990] Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods*. Edward Arnold.

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

[Kumar et al., 2000] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. (2000). Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 57–65, Redondo Beach, CA, USA. IEEE CS Press.

[Lee et al., 2004] Lee, C. P., Golub, G. H., and Zenios, S. A. (2004). A fast two-stage algorithm for computing pagerank and its extensions. Technical report, Stanford University.

[Lempel and Moran, 2001] Lempel, R. and Moran, S. (2001). Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160.

[Leskovec et al., 2005] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA. ACM Press.

[Li, 1998] Li, Y. (1998). Toward a qualitative search engine. *IEEE Internet Computing*.

[Lifantsev, 2000] Lifantsev, M. (2000). Voting model for ranking Web pages. In Graham, P. and Maheswaran, M., editors, *Proceedings of the International Conference on Internet Computing*, pages 143–148, Las Vegas, Nevada, USA. CSREA Press.

[Liu and Ma, 2005] Liu, T.-Y. and Ma, W.-Y. (2005). Webpage importance analysis using conditional markov random walk. In *The IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 515–521, Compiegne, France. ACM Press.

[Marchiori, 1997] Marchiori, M. (1997). The quest for correct information of the Web: hyper search engines. In *Proc. of the sixth international conference on the Web*, Santa Clara, USA.

[Menczer, 2004] Menczer, F. (2004). Lexical and semantic clustering by web links. *Journal of the American Society for Information Science and Technology*, 55(14):1261–1269.

[Newman et al., 2001] Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(2 Pt 2).

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project.

[Pandurangan et al., 2002] Pandurangan, G., Raghavan, P., and Upfal, E. (2002). Using Pagerank to characterize Web structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*, volume 2387 of *Lecture Notes in Computer Science*, pages 330–390, Singapore. Springer.

[Rudin, 1986] Rudin, W. (1986). *Real and Complex Analysis*. McGraw-Hill Science/Engineering/Math.

[Srinivasan et al., 2005] Srinivasan, P., Pant, G., and Menczer, F. (2005). A general evaluation framework for topical crawlers. *Information Retrieval*, 8(3):417–447.

[Tauro et al., 2001] Tauro, L., Palmer, C., Siganos, G., and Faloutsos, M. (2001). A simple conceptual model for the internet topology. In *Global Internet*, San Antonio, Texas, USA. IEEE CS Press.

[Upstill et al., 2003] Upstill, T., Craswell, N., and Hawking, D. (2003). Predicting fame and fortune: Pagerank or indegree? In *Proceedings of the Australasian Document Computing Symposium, ADCS2003*, pages 31–40, Canberra, Australia.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

[Wu et al., 2004] Wu, F., Huberman, B. A., Adamic, L. A., and Tyler, J. R. (2004). Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327–335.