The Choice of a Damping Function for Propagating Importance in Link-Based Ranking

Technical report RI-DSI N. 305-05
Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano.
2005

A version of this technical report was peer-reviewed and published in 2006, please refer to the published version for citation.

The Choice of a Damping Function for Propagating Importance in Link-Based Ranking

Ricardo Baeza-Yates
ICREA Professor
Depto. de Tecnología
Universitat Pompeu Fabra
Spain

Paolo Boldi*
Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
Italy

Carlos Castillo
Cátedra Telefónica
Depto. de Tecnología
Universitat Pompeu Fabra
Spain

Abstract

This paper studies a family of link-based algorithms that propagate page importance through links. In these algorithms there is a damping function that decreases with the distance, so a direct link implies more endorsement than a link through a long path. PageRank is the most widely known ranking function of this family.

We focus on three damping functions, having linear, exponential, and hyperbolic decay on the lengths of the paths. The exponential decay corresponds to PageRank, and the other functions are new. Our analysis includes a comparison among them and experiments for studying their behavior under different parameters.

1 Introduction

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 [28, 20, 26]; however, it quickly become clear that just counting the links was not a very reliable measure of authoritativeness (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 nearly nothing.

The PageRank technique, introduced by Page *et al.* [30], actually tries to mend this problem by looking at the importance of a page in a recursive manner. In other words, "a page with high PageRank is a page referenced by many pages with high PageRank": PageRank not only counts the direct links to a page, but also includes indirect links. The same is valid for scientific citations.

PageRank turns out to be a simple, robust and reliable way to measure the importance of web pages, and it can be computed in a very efficient way. For these reasons, most of today's commercial search engines are believed to use PageRank as an important source of ranking.

In this paper we:

- describe general ranking functions that depend on incoming paths of varying lengths,
- show that PageRank belongs to this class of functions,

*Partially supported by MIUR COFIN Project "Linguaggi formali e automi".

- show how to compute these rankings,
- suggest why 0.85 is a good choice of the parameter for PageRank, and
- compare the ranking orders induced by different ranking functions, finding ways of approximating PageRank up to very high precision.

The rest of this paper is organized as follows: Section 2 introduces the notion of functional ranking, Section 3 describes three damping functions; Section 4 compares them analytically, and Section 5 experiments with different parameters for each function. Finally, the last section presents our conclusions.

2 Functional Rankings

In this section, we introduce the notion of *functional ranking*, a general family of ranking functions that includes PageRank. To describe PageRank formally, consider a web graph of N pages. Let $\mathbf{A}_{N\times N}$ be the link matrix in this graph, $a_{i,j} = 1$ iff there is a link from page i to page j. This link matrix is seldom used as it is, mainly for two reasons:

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 [27] 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 assign more score than it has. 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, as there is 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. [15]. 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 [4], or Altavista's Eigenrank that treat them differently. Other characteristics of links, such as if they appear at the beginning or the bottom of the page, or if they appear within a certain HTML element, can also be used for non-uniform normalization [2].

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.

Dangling nodes. Special attention should be paid to the possible presence of nodes with no outgoing arcs (also known as "dangling nodes"): 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. 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 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 [6], 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 an equivalent, and actually very intriguing way of rewriting this formula, mentioned in [29] that leads to a conclusion similar to those of [8]: given a path $p = \langle x_1, x_2, \ldots, x_k \rangle$ in the graph, 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 $\text{Path}(-,i)$ 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. 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 will refer to this form of ranking as *functional ranking*, because it is parametrized by the damping function. As we have seen, generic PageRank is a functional ranking where the damping function

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

decays exponentially fast.

3 Damping Functions

In this section, we show several functional rankings by describing their damping functions. First, we show which class of damping functions generate well-defined functional rankings.

As shown in [8, 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, $\sum_{p \in \text{Path}(i,-),|p|=t+1} \text{branching}(p)$ 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

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

Using Theorem 1, $\sum_{p \in \text{Path}(-,-)} \text{branching}(p) = N$, so the latter equality is equivalent to

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

Hence, every choice of the damping function such that $\sum_{t=0}^{\infty} \text{damping}(t) = 1$ yields a well-defined normalized functional ranking. Nonetheless, 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.

3.1 Linear damping

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

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

this 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 indegree, 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 considered, so the computation is fast and the number of iterations is fixed.

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 $\sum_{k=0}^{L-1}\mathbf{R}^{(k)}$. An algorithm for computing this ranking, shown in Figure 1, arises directly from this summation.

3.2 Exponential damping: PageRank

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

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

As longer paths have less importance in the calculation of PageRank, it could be approximated by using only a few levels of links. In [10], it is shown that by using only the nodes at distance 1 from a target node (equivalent to linear damping with $L = 2$), PageRank 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 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 [30]; 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

Require: L: maximum path length, N: number of nodes, **v**: preference vector
1: **for** i : 1 ... N **do** {Initialization}
2: S[i] $\leftarrow$ R[i] $\leftarrow 2v[i]/(L+1)$
3: **end for**
4: **for** step : 1 ... L$-$1 **do** {Iteration step}
5: Aux $\leftarrow$ **0**
6: **for** i : 1 ... N **do** {Follow links in the graph}
7: **for all** j such that there is a link from i to j **do**
8: Aux[j] $\leftarrow$ Aux[j] + R[i]/outdegree(i)
9: **end for**
10: **end for**
11: **for** i : 1 ... N **do** {Add to ranking value}
12: R[i] $\leftarrow$ Aux[i] $\times \frac{(\text{L}-\text{step})}{(\text{L}-(\text{step}-1))}$
13: S[i] $\leftarrow$ S[i] + R[i]
14: **end for**
15: **end for**
16: **return** R

Figure 1: Algorithm for computing a functional ranking with linear damping.

α gets close to 1 [18, 17]. 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, [30, 16, 13, 25, 22, 21]).

3.3 Quadratic hyperbolic damping: TotalRank

Recently, a ranking method called TotalRank [5] 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:

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

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 $\mathbf{P}^{-1}$ is not singular and $\mathbf{P} \neq \mathbf{I}$.

TotalRank is as a weighted sum of the score 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:

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

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

Computation. It is known that the cost of calculating TotalRank is the same as the cost of calculating PageRank via the Power Method [6], even though some more iterations are required to obtain the same precision.

3.4 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, $\zeta(\beta) = \sum_{t=1}^{\infty} \frac{1}{t^{\beta}}$ 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. Figure 2 shows an algorithm for approximating HyperRank. 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 leads to the algorithm, 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.

Require: N: number of nodes, **v**: preference vector, β: damping parameter
1: **for** i : 1 ... N **do** {Initialization}
2: S[i] $\leftarrow$ R[i] $\leftarrow v[i]/\zeta(\beta)$
3: **end for**
4: k $\leftarrow$ 0
5: **while** not converged **do** {Iteration step}
6: k $\leftarrow$ k + 1
7: Aux $\leftarrow$ **0**
8: **for** i : 1 ... N **do** {Follow links in the graph}
9: **for all** j such that there is a link from i to j **do**
10: Aux[j] $\leftarrow$ Aux[j] + R[i]/outdegree(i)
11: **end for**
12: **end for**
13: **for** i : 1 ... N **do** {Add to ranking value}
14: R[i] $\leftarrow$ Aux[i] $\times (k/(k+1))^{\beta}$
15: S[i] $\leftarrow$ S[i] + R[i]
16: **end for**
17: **end while**
18: **return** S

Figure 2: An algorithm to compute general hyperbolic rank.

3.5 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 link to each other are more similar than pages chosen at random [12]; evidence from topical crawlers [32] shows that when doing breadth-first exploring, the topic "drifts" as the distance increases. On the same line of though, we propose to use the decrease of text similarity as an approximation to an "empirical" damping function.

To find out which is 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, 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 [3], without stemming nor stopwords removal.

The resulting averages are shown in Figure 3, 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 [35] uses the number of repeated words and phrases between pages and obtains a faster decrease in similarity.

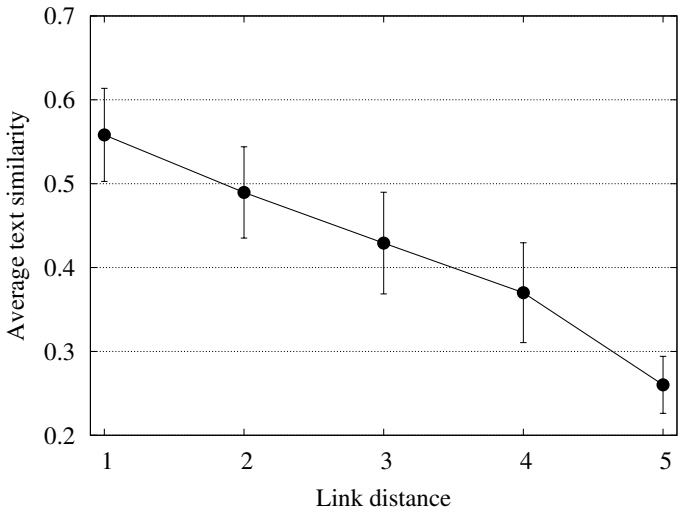

Figure 3: Link distance vs. average text similarity.

4 Comparing Damping Functions

A comparison of the damping functions described in the previous section is shown in Figure 4: of course, hyperbolic damping functions decay asymptotically more slowly than exponential damping, but notice that for short paths the latter may dominate the former in many cases.

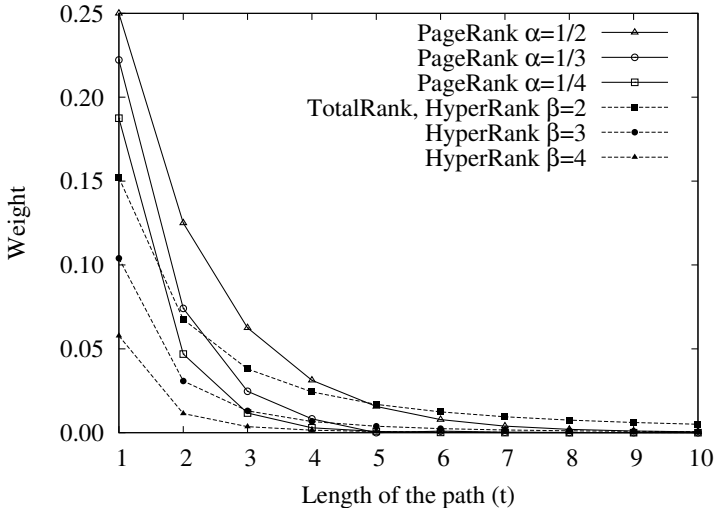

Figure 4: Weights given by the different damping functions for some values of α and β.

In this section, we aim at analyzing how similar are these functional rankings, and how could we use one of the damping functions to approximate another with a suitable choice of parameters.

4.1 Approximating TotalRank with PageRank

It has been shown experimentally that the rank correlation (Kendall's τ) between TotalRank and PageRank is maximal when $\alpha \approx 0.7$ [5]; the maximum value for τ is over 0.95, meaning that 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; more precisely, we aim at studying 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 5 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 [9] or 19 [1], 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.

4.2 Approximating HyperRank with PageRank

Now we want to approximate the weights of:

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

using the weights of:

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

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:

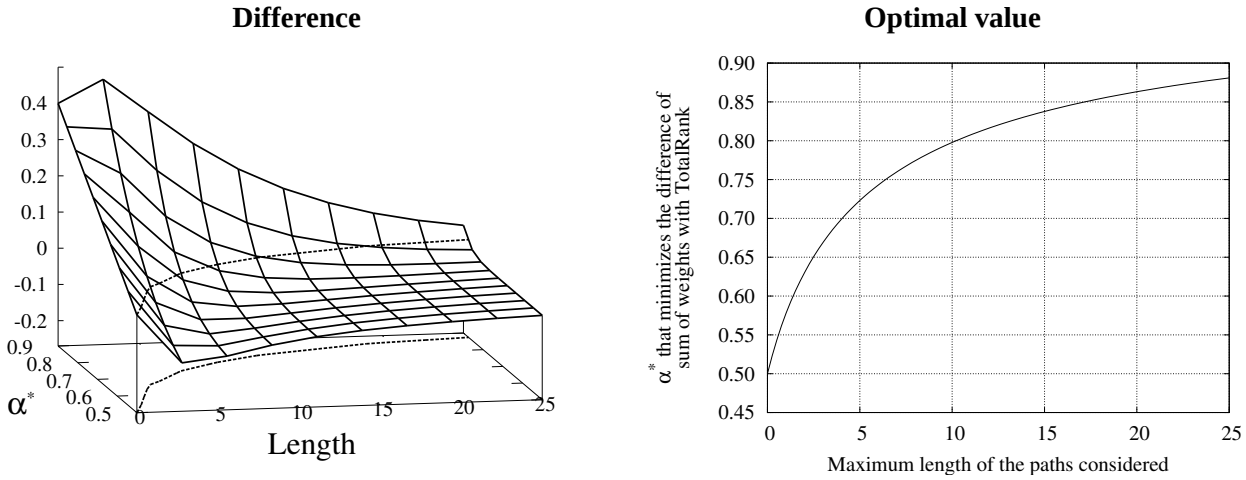

Figure 5: Best $\alpha^*(\ell)$ for minimizing the difference of the sum of weights between PageRank and TotalRank.

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

The α that minimizes the difference of weights for different values of β and of the maximum path lengths ℓ is shown in Figure 6. 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.

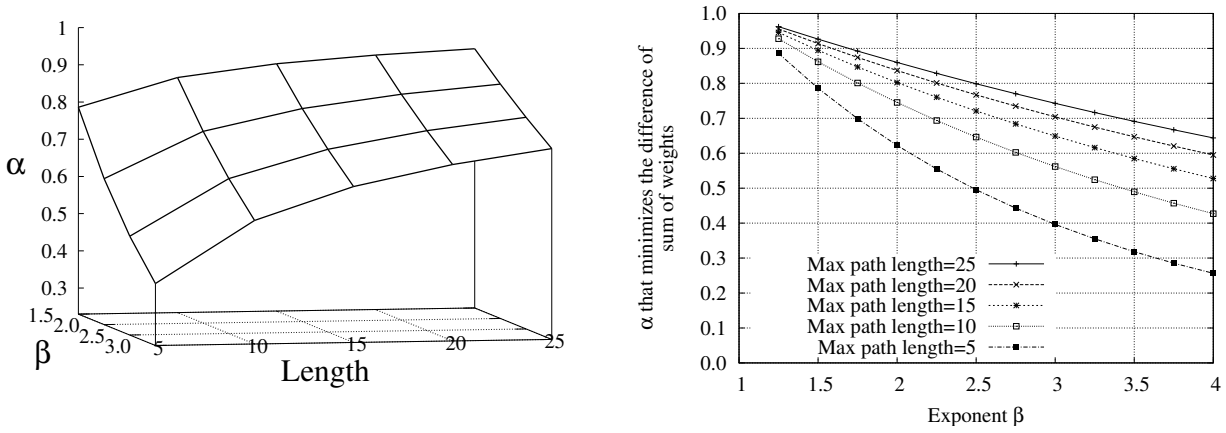

Figure 6: Best α for minimizing the difference of the sum of weights between PageRank and HyperRank with exponent β, for various path lengths.

4.3 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 in which both rankings have non-zero values. The L that minimizes the difference for a given combination of α and ℓ is

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

and we have plotted it for different values of α and ℓ in Figure 7.

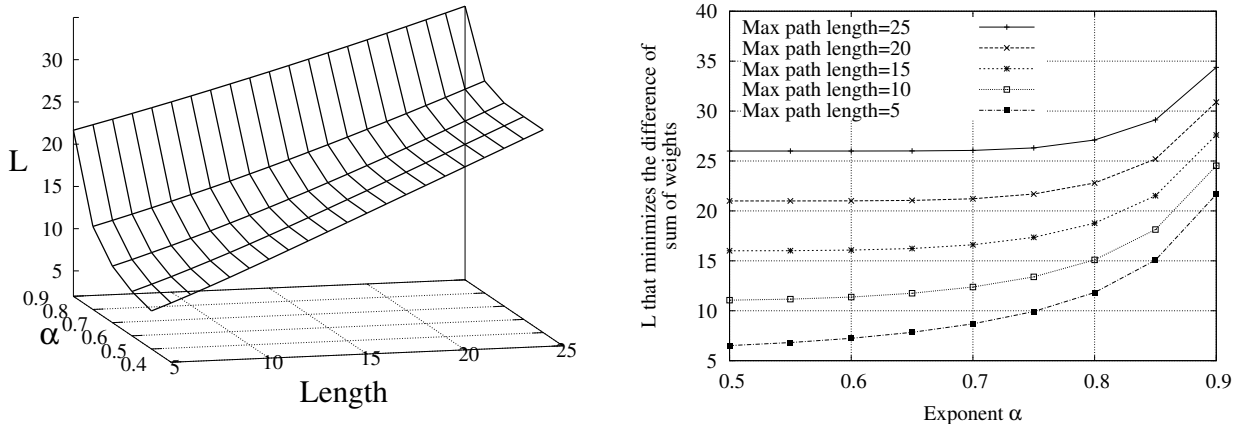

Figure 7: Best L for minimizing the difference of the sum of weights between LinearRank and LinearRank, for various path lengths.

5 Parameters for the Damping Functions

For our experiments, we used several snapshots from the Web, including the `.uk`, `.it` and `.eu.int` domains. For comparison, we also considered a synthetic scale-free network produced according to the evolving model described by Kumar *et al.* [24] (a combination of preferential attachment and random links) with the parameters suggested by Pandurangan *et al.* [31]. As far as the latter is concerned, 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.

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

5.1 Characteristic path lengths

In scale-free networks, the distances between pairs of nodes follow a Gaussian distribution [1] (the average is not given in their paper). Analytic estimations for the average distance of a graph of scale-free network of n nodes include:

- $O(\log(n))$ [34],
- $O(\log(n)/\log(np))$ in sparse graphs with p links [11],
- $1+\log(n/z_1)/\log(z_2/z_1)$ where z_1 is the average indegree, and z_2 is the average number of nodes at distance 2 [29], and
- $O(\log(n)/\log(\log(n)))$ [7].

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 8 plots the average distances found, which appear to be growing (sub)logarithmically with the size of the graph. Figure 9 shows the distribution obtained in each sample. For this experiment, we are not counting the pages without in-links.

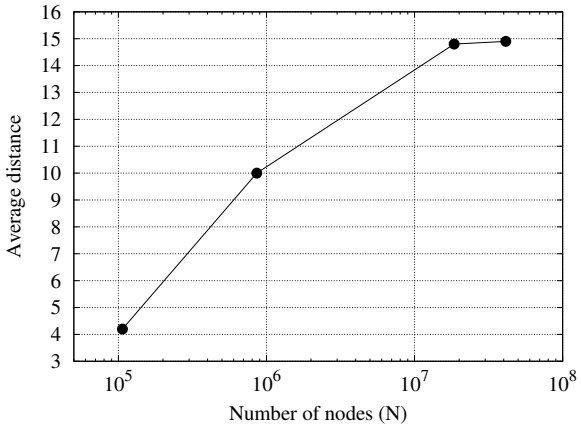

Figure 8: Average distances versus number of nodes.

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

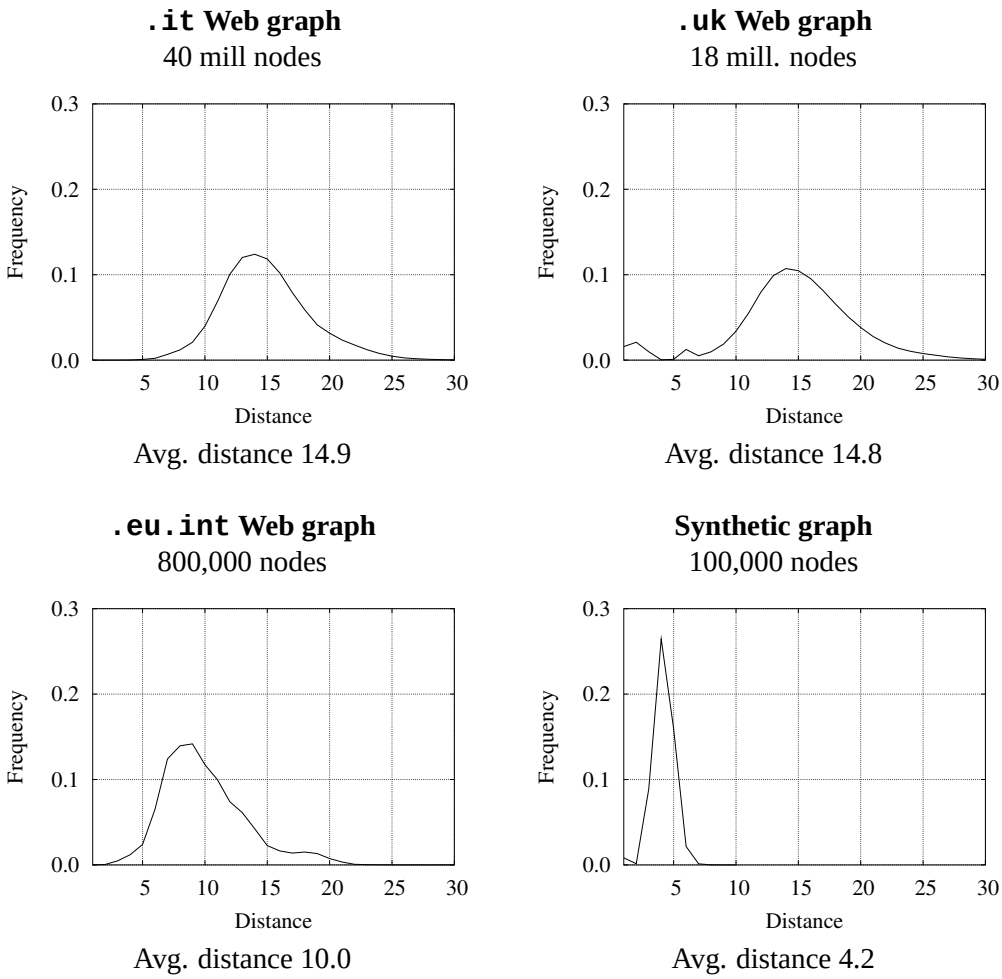

Figure 9: Distribution of the average number of nodes at a certain distance from a given node.

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, consistent with what we have done so far, is to consider that the sum of the weights up to the average path lengths of the graphs (L_1, L_2) have to be similar 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 used in 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 [14]), 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 $\alpha_2^* = 0.76$.

5.2 Damping parameters and in-degree

In this section, we are using data from the `.uk` Web graph and a 8,500-nodes synthetic graph with the same indegree and outdegree distribution than the previous 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 covariance, correlation coefficient and ranking orders (Kendall's τ). The results are shown in Figure 10.

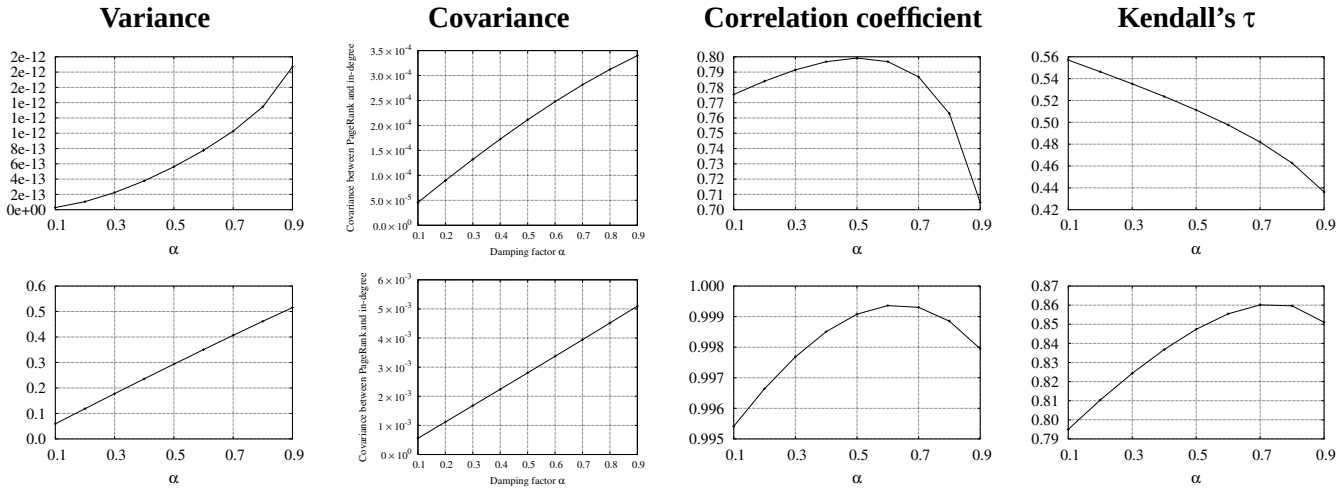

Figure 10: Variance of PageRank, and its relationship with indegree in terms of covariance, correlation coefficient and Kendall's τ coefficient, for varying values of the parameter α. Top: `.uk` web graph; bottom: synthetic graph.

The variance is higher as α increases. In regard to the relationship with indegree, for company home pages, it has been shown that the logarithm of the in-degree is correlated with PageRank [33]. Our results are consistent with this observation. The covariance monotonically increases with α both in the case of the synthetic graph and in the web graph. 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 significatively as α increases, because a large α means that longer paths have an effect in the calculation; note, however, that this phenomenon does not significantly impact on 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[10], 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, as in Figure 11 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.

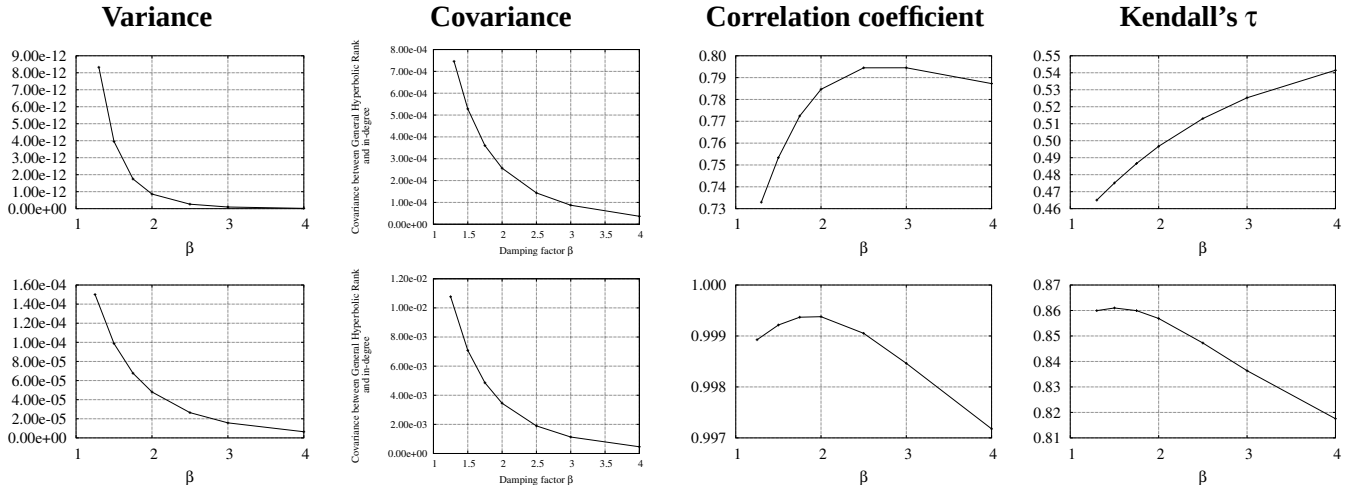

Figure 11: Variance of general hyperbolic rank for some values of β, and its relation with in-degree. Experiments have been performed on the `.uk` graph (top) and synthetic graph (bottom).

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 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 neighbours of highly linked pages have small indegree).

5.3 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 comparison, we used Kendall's τ and data from the U. K. Web graph. Figure 12 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.

In the figure, we can see that the rankings obtained with HyperRank and PageRank can be almost equivalent (Kendall's $\tau \geq 0.95$), moreover, the analysis shown in section 4.2 considering only paths of lengths less than 5, provides a very good approximation for the optimum 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 its convergence is not faster than the one of PageRank.

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 13.

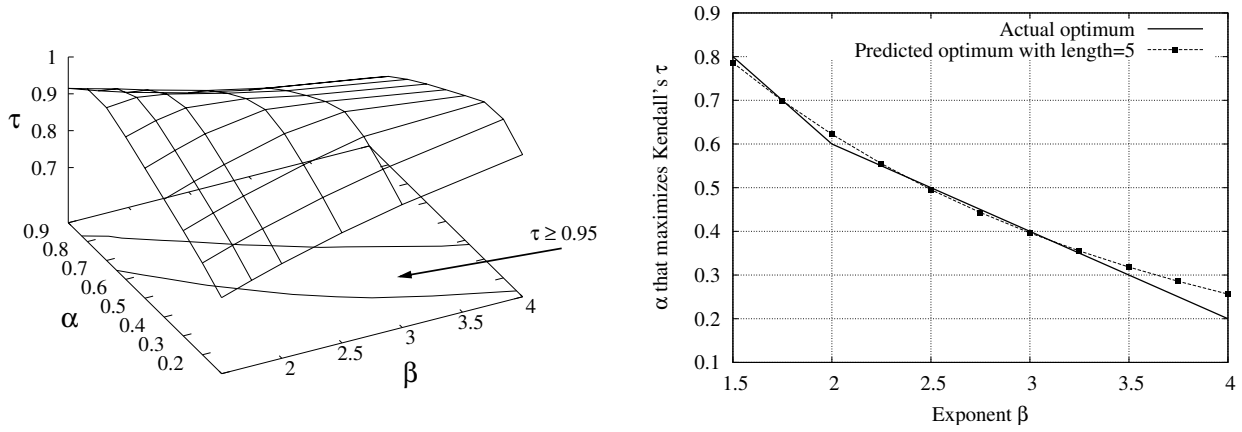

Figure 12: Comparison (using Kendall's τ) between PageRank and HyperRank, with various damping parameters in the U.K. web graph. The optimum predicted in the analysis with $\ell = 5$ is very close to the real one.

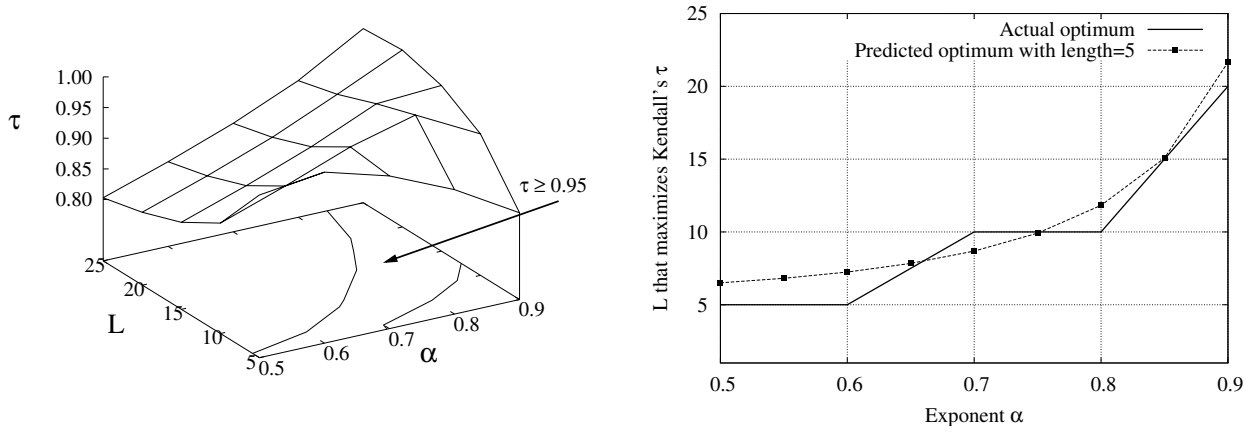

Figure 13: Comparison (using Kendall's τ) between PageRank and LinearRank in the U.K. web graph, with various damping parameters. Again, the predicted optimum with $\ell = 5$ is very close to the actual optimum.

The predicted optimum given in section 4.3 with $\ell = 5$, this is, 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.

6 Conclusions

In this paper we have defined a broad class of link-based ranking algorithms based on the contribution of damping factors along all different paths reaching a page. We introduce four particular damping decays: lin-

ear, exponential, quadratic hyperbolic and general hyperbolic, where exponential is equivalent to PageRank and quadratic hyperbolic to TotalRank.

We studied the differences and similarities among these ranking algorithms, and we have found that:

- Functional rankings using different damping functions can provide similar orderings, if the parameters are chosen carefully.
- LinearRank can be used for calculating a ranking that is as good as PageRank, but with a fixed, and smaller, number of iterations.
- The parameters for the damping functions depend on the characteristic of path lengths in the graph, which is known to grow sub-logarithmically on the size of the graph.

More work is needed to find other damping functions that compute rankings similar to PageRank but are easier and faster to compute. We use 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 this moment, but the functional ranking with linear damping we have presented can provide a good approximation with few iterations. Also, the approach we have presented could be also applied to multivalued ranking functions such as HITS [23] and topic-sensitive PageRank [19] 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 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.

Acknowledgements

We would like to thank Dániel Fogaras for a valuable discussion about TotalRank that motivated part of this research.

References

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

[2] R. Baeza-Yates and E. Davis. 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, 2004. ACM Press.

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

[4] K. Bharat and M. 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*, pages 104–111, Melbourne, Australia, August 1998. ACM Press, New York.

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

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

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

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

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

[10] Y.-Y. Chen, Q. Gan, and T. Suel. 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, 2004. ACM Press.

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

[12] B. D. Davison. 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, 2000. ACM Press.

[13] G. H. Golub and C. Greif. Arnoldi-type algorithms for computing stationary distribution vectors, with application to pagerank. Technical Report SCCM-04-15, Stanford University, 2004.

[14] A. Gulli and A. Signorini. 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, 2005. ACM Press.

[15] S. W. Haas and E. S. Grams. 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, 1998. ACM Press.

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

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

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

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

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

[21] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank, 2003.

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

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

[24] 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 (FOCS)*, pages 57–65, Redondo Beach, CA, USA, 2000. IEEE CS Press.

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

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

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

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

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

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

[31] G. Pandurangan, P. Raghavan, and E. Upfal. 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, August 2002. Springer.

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

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

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

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