Traps and Pitfalls of Topic-Biased PageRank

Paolo Boldi* Roberto Posenato† Massimo Santini Sebastiano Vigna
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Italy
and† Dipartimento di Informatica, Università degli Studi di Verona, Italy

Abstract

We discuss a number of issues in the definition, computation and comparison of PageRank values that have been addressed sparsely in the literature, often with contradictory approaches. We study the difference between *weakly* and *strongly* preferential PageRank, which patch the dangling nodes with different distributions, extending analytical formulae known for the strongly preferential case, and corroborating our results with experiments on a snapshot of 100 millions of pages of the `.uk` domain. The experiments show that the two PageRank versions are poorly correlated, and results about each one cannot be blindly applied to the other; moreover, our computations highlight some new concerns about the usage of exchange-based correlation indices (such as Kendall's τ) on approximated rankings.

1 Introduction

This paper started with an attempt to reproduce the correlation data published by Haveliwala [1] about rankings biased towards different topics (where the correlation was computed using a measure similar to Kendall's τ); such seminal work has been receiving some attention lately, as in [2, 3]. The bias was introduced using a *preference vector*, that is, by assuming that upon teleportation (see below for definitions) one does not land in a node chosen uniformly at random, but rather according to a given distribution.

During our attempts, we met significant difficulties due to the number of different ways in which PageRank can be defined and computed, and to the lack of public data over which to replicate the experiments. Following the incongruences in the literature, we were led to study in great detail the way in which PageRank depends on the preference vector and on the way dangling nodes are patched to obtain the final Markov chain. Also the way in which correlation indices are computed, and their depencence on the precision of the computation, turned out to be decisive.

We report the results obtained along our way. All our experiments use publicly available data gathered by UbiCrawler [4] on the `.uk` domain in the context of the EU project DELIS [5]. The topic-bias data we use are derived from the ODP [6] hierarchy. We believe such a public, well-defined data set is essential to continue research on personalised (and, in particular, topic-based) ranking.

First of all, we provide analytical formulae for *weakly preferential* and *strongly preferential* PageRank—two variants frequently found in the literature in which different distributions are used to patch dangling nodes. Using the Sherman–Morrison formula we are able to extend the results of Del Corso, Gullì and Romani [7] for strongly preferential PageRank. In doing so, we introduce the notion of

*This work is partially supported by the EC Project DELIS and by MIUR PRIN Project "Automi e linguaggi formali: aspetti matematici e applicativi".

pseudorank, a vector obtained using a PageRank-like matrix (which is not necessarily stochastic). Pseudoranks simplify greatly the following discussion, and present some interesting phenomena.

Then, we report experiments showing that weakly and strongly preferential PageRank can be very poorly correlated, and that results in the literature obtained using the two approaches are hardly comparable. In doing so, we use a low-level truncation technique that avoids the usual (and usually neglected in the literature) noise associated with the result of an interrupted iterative process, and we conclude by showing experimentally that such a noise may have a great impact on the computation of rank-based correlation indices.

2 PageRank

Albeit definitions of PageRank can be easily found in the literature, our purpose is precisely that of clarifying some relevant differences, so we start from scratch. Given a (web) graph G, the *row-normalised matrix* of G is the matrix P such that p_{ij} is one over the outdegree of i if there is an arc from i to j in G, zero otherwise. Note that in general P will not be stochastic, as it can have rows entirely made of zeroes.

Let us define $\boldsymbol{d}$ as the characteristic vector[1] of the dangling nodes (i.e., the vector with 1 in positions corresponding to nodes without outgoing arcs and 0 elsewhere). Let $\boldsymbol{v}$ and $\boldsymbol{u}$ be distributions[2], which we will call the *preference* and the *dangling-node* distribution.

PageRank $\boldsymbol{r}$ is defined (up to a scalar) by the eigenvector equation

$$\boldsymbol{r}^T\left(\alpha(P+\boldsymbol{d}\boldsymbol{u}^T)+(1-\alpha)\mathbf{1}\boldsymbol{v}^T\right)=\boldsymbol{r}^T,$$

that is, as the stationary state of the Markov chain $\alpha(P+\boldsymbol{d}\boldsymbol{u}^T)+(1-\alpha)\mathbf{1}\boldsymbol{v}^T$. More precisely, we have a *Markov chain with restart* [8] in which $P+\boldsymbol{d}\boldsymbol{u}^T$ is the Markov chain (that follows the natural random walk on non-dangling nodes, and moves to a node at random with distribution $\boldsymbol{u}$ when starting from a dangling node) and $\boldsymbol{v}$ is the restart vector. The *damping factor* $\alpha \in [0\,..\,1)$ decides how often the Markov chain follows the graph, and how often it moves at a random node following the preference vector $\boldsymbol{v}$. The latter behaviour is commonly called *teleportation*, referring to a well-known random-walk metaphore in which a random surfer with probability α moves along an outlink chosen uniformly at random (or, in case of a dangling node, chosen among all nodes according to distribution $\boldsymbol{u}$), and teleports to a random node chosen with distribution $\boldsymbol{v}$ with probability $1-\alpha$. In the random-surfer metaphore, PageRank is the average fraction of time the surfer spends at a given node.

3 Strongly vs. Weakly Preferential

A significant amount of recent research is devoted to studying the dependence of PageRank on the preference vector. The preference vector biases the rank towards nodes that are closer to nodes with a larger value in the preference vector. The preference vector, for instance, might depend on the user's preferences, in which case one speaks of *personalised PageRank* [2].

Clearly, the preference vector $\boldsymbol{v}$ significantly conditions PageRank. Some care must be exercised, however: real-world snapshots comprise a significant percentage of *dangling nodes* (nodes without outlinks), in particular if the graph contains the whole *frontier* of the crawl [9], rather than just the visited nodes. Hence, the way in which the surfer chooses the next node when she is at a dangling

[1] All vectors in this work are column vectors.

[2] By *distribution* we mean a vector with non-negative entries and ℓ_1-norm equal to 1.

node (i.e., the choice of $\boldsymbol{u}$) is also very relevant, and it is an issue resolved in different ways by different authors.

We distinguish clearly between *strongly preferential* PageRank, in which the preference and dangling-node distributions are identical (i.e., $\boldsymbol{u} = \boldsymbol{v}$), and correspond to a topic or personalisation bias, and *weakly preferential* PageRank, in which the preference and the dangling-node distributions are not identical, and, in principle, uncorrelated (most commonly, $\boldsymbol{u} = \mathbf{1}/n$). As we shall see, the distinction is not irrelevant, as the correlation between weakly and strongly preferential PageRank can be quite low.

As a first analytical step to understand fully the relationship between preference and dangling-node distributions we extend the closed formula given by Del Corso, Gullì and Romani [7] for strongly preferential PageRank to a general formula that applies also to weakly preferential PageRank. Using this formula, any biased, weakly preferential PageRank vector whose distributions are a linear combination of a set of base vectors [2] can be computed using the *pseudorank* vectors associated to the base vectors. The computation of a pseudorank vector requires the same amount of computational effort as for computing PageRank, but once pseudoranks have been computed it is immediate to compute and compare several different biased ranks.[3]

PageRank $\boldsymbol{r}$ is defined (up to a scalar) by the eigenvector equation

$$\boldsymbol{r}^T\big(\alpha(P + \boldsymbol{d}\boldsymbol{u}^T) + (1-\alpha)\mathbf{1}\boldsymbol{v}^T\big) = \boldsymbol{r}^T.$$

After a transposition, imposing $\boldsymbol{r}^T\mathbf{1} = 1$ and solving for $\boldsymbol{r}$, we obtain the standard closed form

$$\boldsymbol{r} = (1-\alpha)\big(I - \alpha P^T - \alpha\boldsymbol{u}\boldsymbol{d}^T\big)^{-1}\boldsymbol{v}.$$

The interesting point of this form is that it exhibits PageRank as a *linear operator* on the preference vector $\boldsymbol{v}$.

Definition 1 *Let P be a row-normalised web-graph matrix. The* pseudorank *of P with preference vector $\boldsymbol{v}$ and damping factor $\alpha \in [0\,..\,1]$ is defined as*

$$\widetilde{\boldsymbol{v}}(\alpha) = (1-\alpha)\big(I - \alpha P^T\big)^{-1}\boldsymbol{v}.$$

We note by passing that if $\boldsymbol{d} = \mathbf{0}$ (equivalently, if P is stochastic) then $\widetilde{\boldsymbol{v}}(\alpha)$ is actually the PageRank.[4]

The above definition can be extended by continuity to $\alpha = 1$, albeit the fact is not trivial. The *resolvent* of a matrix M is the linear operator $\mathscr{R}(\mu, M) = (\mu I - M)^{-1}$, defined for every μ which is not an eigenvalue of M; it can be expanded into a *Laurent series* around every eigenvalue of M [10, 11]. In particular, the expansion around 1 is

$$\mathscr{R}(\mu, M) = \frac{M^*}{\mu - 1} + \sum_{k=0}^{\infty}(\mu-1)^k Q^{k+1}$$

for a suitable matrix Q, where M^* is the *Cesáro limit*

$$M^* = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}M^k,$$

[3] Actually, some papers, such as [3], use tacitly pseudoranks as the *definition* of PageRank.

[4] The condition number in the computation of pseudoranks is the same as or better than that for PageRank, and there is no increase of computation time.

which is always defined and is equal to $\lim_{k\to\infty} M^k$ whenever the latter is defined [12]. This implies that

$$\lim_{\mu\to 1^+}(1-\mu)\mathscr{R}(\mu, M) = M^*,$$

so the pseudorank for $\alpha \to 1$ is simply $(P^*)^T \boldsymbol{v}$.

Theorem 1 *Let P the row-normalised matrix of a web graph, $\boldsymbol{v}$ the preference vector, $\boldsymbol{u}$ the dangling distribution and α the damping factor. Then, the PageRank vector $\boldsymbol{r}$ satisfies*

$$\boldsymbol{r} = \widetilde{\boldsymbol{v}}(\alpha) - \widetilde{\boldsymbol{u}}(\alpha)\frac{\boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha)}{1-\frac{1}{\alpha}+\boldsymbol{d}^T\widetilde{\boldsymbol{u}}(\alpha)}.$$

Proof. Let $\widetilde{\boldsymbol{u}}(\alpha)$ and $\widetilde{\boldsymbol{v}}(\alpha)$ be the pseudoranks of $\boldsymbol{u}$ and $\boldsymbol{v}$, and define $R = I - \alpha P^T$. By the Sherman–Morrison formula [7], we have

$$\boldsymbol{r} = (1-\alpha)(R-\alpha\boldsymbol{u}\boldsymbol{d}^T)^{-1}\boldsymbol{v} = (1-\alpha)R^{-1}\boldsymbol{v} + (1-\alpha)\frac{R^{-1}\boldsymbol{u}\boldsymbol{d}^T R^{-1}}{\frac{1}{\alpha}-\boldsymbol{d}^T R^{-1}\boldsymbol{u}}\boldsymbol{v} =$$

$$= \widetilde{\boldsymbol{v}}(\alpha) + \frac{\widetilde{\boldsymbol{u}}(\alpha)\boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha)}{\frac{1}{\alpha}-1-\boldsymbol{d}^T\widetilde{\boldsymbol{u}}(\alpha)}. \blacksquare$$

Note that the scalar values $\boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha)$ and $\boldsymbol{d}^T\widetilde{\boldsymbol{u}}(\alpha)$ have two very simple interpretation—they are the pseudorank accumulated by dangling nodes w.r.t. $\boldsymbol{v}$ and $\boldsymbol{u}$, respectively.

By properly ordering multiplications, no matrix computation is necessary to compute the formula above. When $\boldsymbol{u} = \boldsymbol{v}$, the formula reduces to the one provided in [7]:[5]

$$\boldsymbol{r} = \widetilde{\boldsymbol{v}}(\alpha)\left(1 - \frac{\boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha)}{1-\frac{1}{\alpha}+\boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha)}\right) \tag{1}$$

and the (rather surprising) consequence is that *pseudoranks are just multiples of strongly preferential ranks*. In other words, PageRank might as well be computed *without taking care of dangling nodes* by using the standard expansion

$$\widetilde{\boldsymbol{v}}(\alpha) = (1-\alpha)\big(I-\alpha P^T\big)^{-1}\boldsymbol{v} = (1-\alpha)\sum_{n=0}^{\infty}\alpha^n\big(P^T\big)^n\boldsymbol{v}.$$

Indeed, by truncating the infinite sum we obtain an approximation of the pseudorank:

$$\left\|\widetilde{\boldsymbol{v}}(\alpha) - (1-\alpha)\sum_{n=0}^{k}\alpha^n\big(P^T\big)^n\boldsymbol{v}\right\| = \left\|(1-\alpha)\sum_{n=k+1}^{\infty}\alpha^n\big(P^T\big)^n\boldsymbol{v}\right\| \le \alpha^{k+1}.$$

The fact that the above formula approximates well PageRank up to a constant factor shows that actually PageRank is related more to a *diffusion* than to a *mixing* phenomenon. In other words, even if the PageRank definition is in term of Markov chains, its value can be computed also by a cumulative process in which the preference vector is broadcast to the neighbours using a decay factor α.

Pseudoranks are computed from their preference vector using a linear operator: as a consequence, both weakly and strongly preferential PageRank are quickly computable if, for instance, $\widetilde{\boldsymbol{e}_i}(\alpha)$ is known for some base $\boldsymbol{e}_i$ of the vector space. This property is noted in [2] for strongly preferential PageRank, but Theorem 1 shows that the statement is true also in the weakly preferential case, albeit the dependence on $\boldsymbol{u}$ *is not linear*, so weakly preferential PageRank vectors do not obey the simple linear laws for what matters the dangling node distribution.

[5] The reader should note that our formula has some difference in signs w.r.t. the original paper, where it was calculated incorrectly.

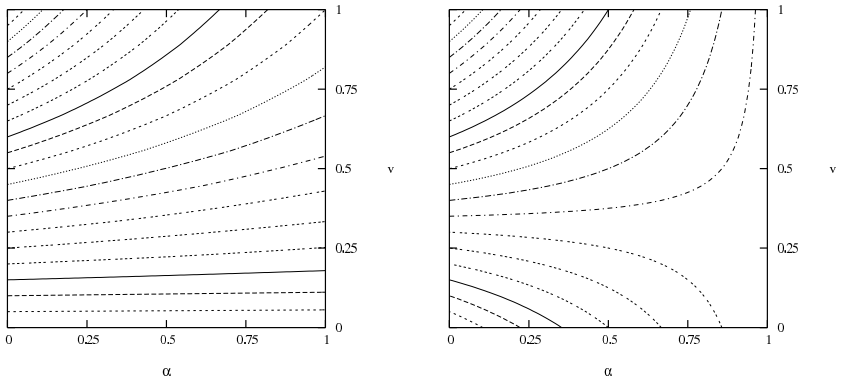

Figure 1: The contour plot of $r_1 - r_2$ for the strongly preferential (left) and weakly preferential (right) case for the worked example.

3.1 A worked example, and some observations

Let us consider the simple example of a graph with two nodes, and a single arc going from the first to the second node. With an arbitrary norm-one vector $\boldsymbol{x} = (x, 1-x)^T$ we have

$$\widetilde{\boldsymbol{x}}(\alpha) = (1-\alpha)(I - \alpha P^T)^{-1}\boldsymbol{x} = \begin{pmatrix} (\alpha-1)x \\ (1-\alpha)(1+(\alpha-1)x) \end{pmatrix},$$

and $\boldsymbol{d}^T\widetilde{\boldsymbol{x}}(\alpha) = (\alpha-1)x$. Note that, for every preference vector $\boldsymbol{v}$, $\lim_{\alpha\to 1^-} \boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha) = 0$, and this is not by chance: $\boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha)$ has a limit as $\alpha \to 1$ because it is a rational function of α, and looking at (1) it is clear that $\boldsymbol{d}^T\widetilde{\boldsymbol{v}}(\alpha)$ cannot converge to any limit different from 0, or otherwise the strongly preferential PageRank would itself converge to the zero vector.

To complete the example, for two arbitrary norm-one vectors $\boldsymbol{v} = (v, 1-v)^T$ and $\boldsymbol{u} = (u, 1-u)^T$ we have that the denominator in Theorem 1 evaluates to $(\alpha-1)(1/\alpha + u)$, giving

$$\boldsymbol{r} = \frac{1}{\alpha u + 1}\begin{pmatrix} v + \alpha(u-v) \\ (\alpha-1)v + 1 \end{pmatrix}.$$

The limit when α approaches 1 is

$$\lim_{\alpha\to 1^-} \boldsymbol{r} = \frac{1}{1+u}\begin{pmatrix} 1 \\ u \end{pmatrix}.$$

A precise estimate of the difference between strongly and weakly preferential PageRank can be obtained by considering the difference $r_1 - r_2$ between the rank values of the two nodes:

$$r_1 - r_2 = \frac{2(1-\alpha)v + \alpha u - 1}{\alpha u + 1};$$

a contour plot of this difference for the strongly preferential and weakly preferential cases (as a function of α and v) is given in Figure 1: note that the behaviours in the two cases are significantly different, in particular when $\alpha > 0.5$ (an area that is quite important, since $\alpha = .85$ is the value that is customarily adopted for PageRank computation).

4 Experiments

Is the difference between strongly and weakly preferential significant also when only ranks are considered instead of rank values? To answer this question, we ran a number of experiments on a crawl

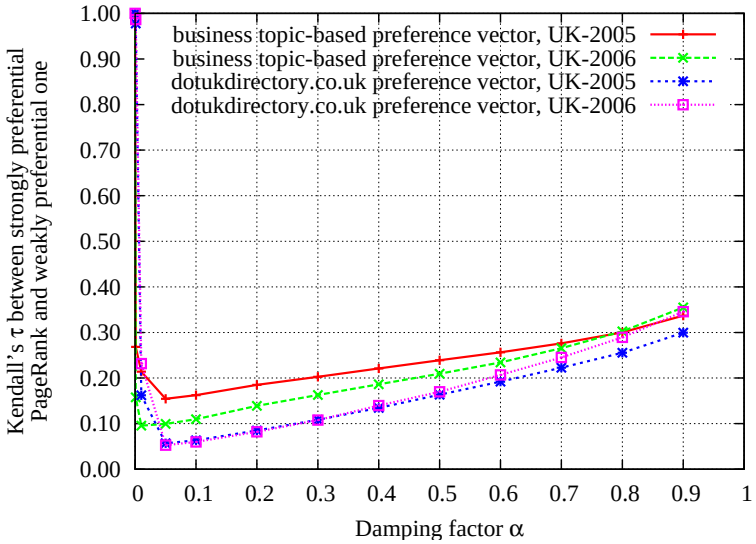

Figure 2: Kendall's τ between weakly and strongly preferential PageRank computed on the UK-2005 and UK-2006 graphs using the ODP "business" topic-based preference vector and the `http://dotuk.directory.co.uk/` page-based preference vector.

of about 100 million pages of the `.uk` domain gathered for the DELIS project [5]. For comparisons we used Kendall's τ, a classical nonparametric correlation index that has recently received much attention within the web community for its possible applications to rank aggregation [13, 14, 15, 16] and for determining the convergence speed in the computation of PageRank [17]. Here we follow exactly the definition given in [16].

In Figure 2 we show the values of Kendall's τ (in dependence of α) for weakly and strongly preferential PageRank where the preference distribution is in one case concentrated on a single node, and in the other case it is uniformly distributed among the nodes in the Open Directory Project [6] "Business" category. In both cases, $\boldsymbol{u}$ was set to the uniform distribution. The correlation between the two values is always very low (except, of course, when $\alpha \approx 0$).

The interesting phenomenon is that correlation *increases* as α increases, whereas intuition would suggest the opposite behaviour: as α increases, the graph becomes more relevant, so the structural differences between using $\boldsymbol{v}$ or the uniform distribution to patch dangling nodes should be more visible.

To understand whether the low correlation is due to topic concentration or to the number of nonzero entries, in Figure 3 we show the comparison of the same kind of values calculated on UK-2006 graph using an additional vector obtained by randomly shuffling the topic-based vector. Our experiments show that the latter exhibits the same behaviour, but with much higher correlation. In other words, topics matter.

5 More Precision Might End in Less Precision

Weak and strong preference is not the only issue met along our way. Correlation measures such as Kendall's τ are based on the number of discordancies among ranks, but the point that appears to have been completely missed in the literature (including that previously contributed by the authors [16]) is

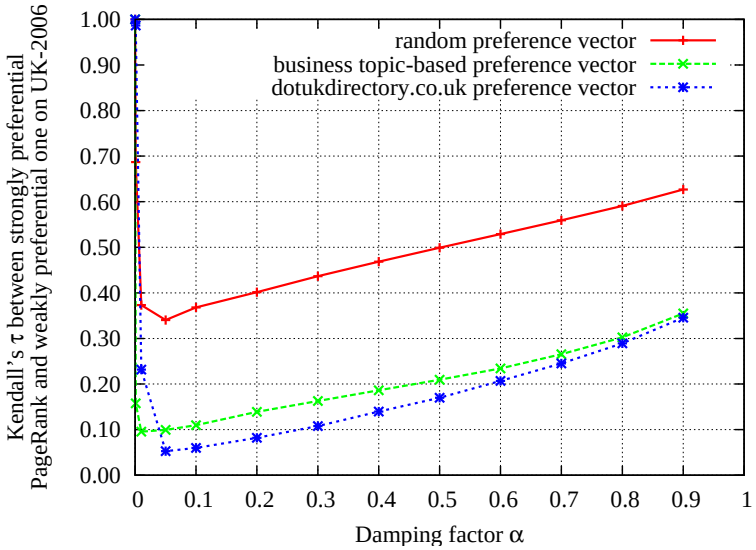

Figure 3: Kendall's τ between weakly and strongly preferential PageRanks calculated on UK-2006 using three different preference vectors.

that the computation of ranks is almost always the result of interrupting an iterative process (e.g., the power method). The interruption is usually based on a threshold satisfied by the ℓ_1 or ℓ_2 measure.

As a result, a number of correct digits appearing in the rank values is hard to predict, as it just depends on the computational process. The abovementioned norms guarantee on *average* a certain number of significant digits, but unless the much more demanding ℓ_∞ measure is used, almost no guarantee can be provided for the rank value of a single node.

In the case several very close values appear in the PageRank vector, the effect of such an unpredictable precision turns out to be catastrophic, in particular with certain computational methods (such as Gauss-Seidel). Namely, the value of Kendall's τ is strongly influenced by the number of significant digits considered in its computation.

To prove the impact of this observation experimentally, we present data obtained by working out the strongly preferential PageRank computation in a standard fashion, using the Gauss-Seidel method, for a certain preference vector. We stopped the computation at different stages, having every time a known (lower bound on the) number of correct digits in the computed ranks, that we denote by p, and then we computed Kendall's τ using only a limited number of digits in the ranks. To limit the number of significant digits we used, we turned each floating point-number into its bitwise IEEE 754 representation, and manipulated it directly so to delete all digits beyond a certain threshold. This procedure, applied with threshold θ, has the effect of *batching* all values in the interval $[\, j2^{-\theta} \,..\, (j+1)2^{-\theta}\,)$, into the value $j2^{-\theta}$. The net effect is that several rank values that appeared to be discordant because of unpredictable noise in the last digits are now considered as concordant. (We remark that due to the size of the data we use, these computations require thousands of hours of CPU time.)

The resulting graphs (an example is presented in Figure 4) are quite surprising: even the τ of a certain PageRank *computed against the same vector, but with a different precision* can go down as low as 0.2. Of course, as far as the computation of τ uses no more digits than those that are guaranteed to be correct, the correlation is 1, but it rapidly drops as soon as more digits are considered; in particular, computing τ blindly (i.e., without any form of batching) can bring essentially to random results. In a slogan: *more precision might end in less precision*. One must be always careful about the actual number of significant digits of each rank—using an ℓ_∞-measure guaranteeing the number of digits

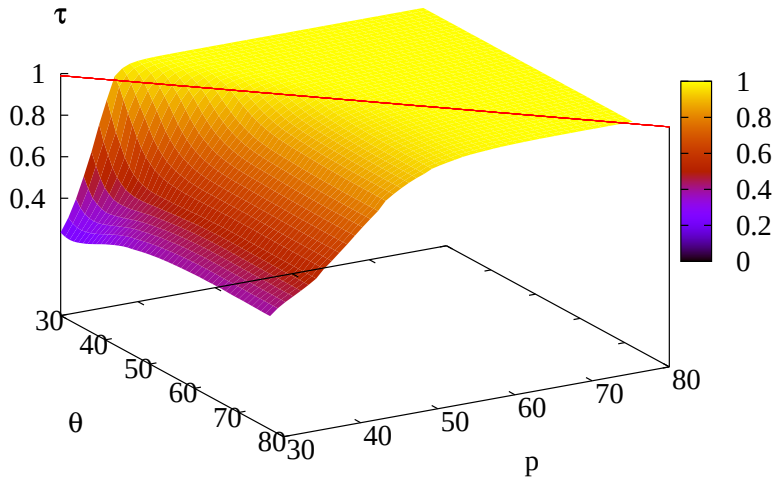

Figure 4: Values of Kendall's τ: when rank values are batched using more bits than the number of significant bits guaranteed by the PageRank computation, the value of τ drops significantly. These data are determineted from the `.uk` web graph, using the ODP "adult" topic-based preference vector, $\alpha = .85$ and the Gauss-Seidel method. p is the number of correct binary digits, θ is the number of digits used to determine τ.

used in the computation of correlation indices is a safe choice.

More evidence is needed to corroborate the data we present. But already our preliminary results show that order-based correlation indices must be managed with great care, and have probably given rise to biased results in the past.

References

[1] Haveliwala, T.H.: Topic-sensitive PageRank. In: The eleventh International Conference on World Wide Web Conference, ACM Press (2002) 517–526

[2] Jeh, G., Widom, J.: Scaling personalized web search. In: WWW '03: Proceedings of the 12th international conference on World Wide Web, New York, NY, USA, ACM Press (2003) 271–279

[3] Csalogány, K., Fogaras, D., Rácz, B., Sarlós, T.: Towards scaling fully personalized PageRank: Algorithms, lower bounds, and experiments. Internet Math. **2** (2006) 333–358

[4] Boldi, P., Codenotti, B., Santini, M., Vigna, S.: Ubicrawler: A scalable fully distributed web crawler. Software: Practice & Experience **34** (2004) 711–726

[5] DELIS: Dynamically Evolving Large-scale Information Systems EC FP6 project. (`http://delis.upb.de/`)

[6] ODP: Open Directory Project. (`http://dmoz.org/`)

[7] Del Corso, G., Gullì, A., Romani, F.: Fast PageRank computation via a sparse linear system. Internet Math. **2** (2006)

[8] Boldi, P., Lonati, V., Santini, M., Vigna, S.: Graph fibrations, graph isomorphism, and PageRank. RAIRO Inform. Théor. **40** (2006) 227–253

[9] Eiron, N., McCurley, K.S., Tomlin, J.A.: Ranking the web frontier. In: Proceedings of the 13th conference on World Wide Web, ACM Press (2004) 309–318

[10] Lasserre, J.B.: A formula for singular perturbations of Markov chains. Journal of Applied Probability **31** (1994) 829–833

[11] Yosida, K.: Functional Analysis. Springer–Verlag (1980) Sixth Edition.

[12] Iosifescu, M.: Finite Markov Processes and Their Applications. John Wiley & Sons (1980)

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

[14] Fagin, R., Kumar, R., McCurley, K.S., Novak, J., Sivakumar, D., Tomlin, J.A., Williamson, D.P.: Searching the workplace web. In: Proceedings of the twelfth international conference on World Wide Web, ACM Press (2003) 366–375

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

[16] Boldi, P., Santini, M., Vigna, S.: Do your worst to make the best: Paradoxical effects in PageRank incremental computations. Internet Math. **2** (2005) 387–404

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