Information Retrieval, 8, 245–264, 2005

Rank-Stability and Rank-Similarity of Link-Based Web Ranking Algorithms in Authority-Connected Graphs

R. LEMPEL rlempel@il.ibm.com
IBM Research Lab, Haifa, Israel

S. MORAN* moran@cs.technion.ac.il
Department of Computer Science, Technion, Haifa, Israel

Abstract. Web search algorithms that rank Web pages by examining the link structure of the Web are attractive from both theoretical and practical aspects. Today's prevailing link-based ranking algorithms rank Web pages by using the dominant eigenvector of certain matrices—like the co-citation matrix or variations thereof. Recent analyses of ranking algorithms have focused attention on the case where the corresponding matrices are irreducible, thus avoiding singularities of reducible matrices. Consequently, rank analysis has been concentrated on *authority connected* graphs, which are graphs whose co-citation matrix is irreducible (after deleting zero rows and columns). Such graphs conceptually correspond to thematically related collections, in which most pages pertain to a single, dominant topic of interest.

A link-based search algorithm A is *rank-stable* if minor changes in the link structure of the input graph, which is usually a subgraph of the Web, do not affect the ranking it produces; algorithms A, B are *rank-similar* if they produce similar rankings. These concepts were introduced and studied recently for various existing search algorithms.

This paper studies the rank-stability and rank-similarity of three link-based ranking algorithms—PageRank, HITS and SALSA—in authority connected graphs. For this class of graphs, we show that neither HITS nor PageRank is rank stable. We then show that HITS and PageRank are not rank similar on this class, nor is any of them rank similar to SALSA.

Keywords: Web IR, citation, link analysis

1. Introduction

The *link-structure* of the Web has been the focus in Web information retrieval research over the last few years. In particular, many novel algorithms for performing Web search and retrieval based on link-structure analysis were proposed (e.g. *HITS* (Kleinberg 1999) and the ensuing CLEVER project (Chakrabarti et al. 1999a, 1999b, 1998b, 1998a), *PageRank* (Brin and Page 1998, Haveliwala 2002), *SALSA* (Lempel and Moran 2001b)). These algorithms demonstrated empirically that link-structure analysis can improve search and rank techniques on the Web (Amento et al. 2000, Silva et al. 2000). Accordingly, search engines (such as *Google*[1] and *AltaVista*[2]) nowadays incorporate extensive link analysis in

*This research was supported by the Fund for the Promotion of Research at the Technion, and by the Barnard Elkin Chair in Computer Science.

their ranking algorithms, that determine the order in which the search results are displayed to the user (Brin and Page 1998, Arasu et al. 2001).

The demonstrated improvements in precision that link-analyzing approaches achieved on the Web have prompted a more rigorous investigation of the theoretical properties of those algorithms. In particular, the robustness, stability, and, to a lesser extent, the theoretical effectiveness of such algorithms has been examined in several recent works (Ng et al. 2001, Azar et al. 2001, Achlioptas et al. 2001, Borodin et al. 2001, Chien et al. 2002, Lee 2002). This paper focuses on two properties: rank-stability and rank-similarity, which are described next.

1.1. Stability and rank stability

Stability is an important quality of any information retrieval algorithm, and link-based ranking algorithms are no exception. Loosely speaking, an algorithm is stable when small perturbations of its input do not dramatically alter its output—in the IR case, the set of retrieved results and their scores. The Web's graph (the pages and their interconnecting hyperlinks) changes constantly as pages are added, deleted and modified. Important, authoritative pages, however, are not as transient and volatile, and link-based algorithms should be able to consistently identify them even as the underlying graph of the Web evolves.

Web-IR algorithms must cope with more than the natural evolution of the Web. With the growing phenomenon of *search engine spamming* (Marchiori 1997, Dwork et al. 2001, Henzinger et al. 2002), the algorithms should remain focussed on the true high quality Web pages while malicious and adversarial attempts are made to bias their output. In the context of this work, we are especially concerned with the proliferation of *link spamming* (Lempel and Moran 2001b, Davison 2000). Link spammers are Web authors that create pages and interconnecting links with the intention of biasing search engine link-based rankings of pages, rather than for delivering content or facilitating human browsing. Link spammers, however, control just a small portion of Web pages and so cannot alter the Web's graph in a radical or global fashion. They only introduce local noise into the graph, and so link-based algorithms should be able to withstand link spamming to some extent.

To the best of our knowledge, all of the link-based ranking algorithms proposed so far operate by first assigning numerical scores to Web pages, and then ranking the pages by those scores. This process gives rise to two different notions of stability: that of the scores, and that of the induced rankings. Following the terminology of Borodin et al. (2001), we will refer to the former notion as *stability*, while the latter notion will be termed *rank-stability*. The issue of stability was looked upon from three angles, which differ with respect to the graphs on which stability was examined.

One approach, which is applied in two related works—(Azar et al. 2001) and (Achlioptas et al. 2001)—assumes that the Web's structure obeys some topic-driven stochastic models. In Azar et al. (2001) it is argued that HITS is stable under a certain model of the Web's link structure. In Achlioptas et al. (2001), a broader model of the Web was presented, modeling not only the linkage patterns between pages but also their textual contents and the relevance of each page to each topic. Under this broad model, the authors devised a query-driven algorithm that is provably effective: the algorithm, with high probability, results in a score vector that is very close to the relevance vector of the pages with respect to the query. This result also implies the stability of their algorithm, since (with high probability) its output is close to a constant vector (per query).

Ng et al. (2001) examined the effect of small changes in the analyzed graph on the score vectors of PageRank, HITS, and *Subspace HITS* (a variation of HITS presented there). For PageRank, they studied the L_1-change to the scores when modifying the outgoing links of a set P of pages. They bounded this change by a linear function of the aggregate score of all pages in P. For HITS, they showed that the L_2-change to the scores is bounded by a function of (1) the number of links added/deleted, (2) the maximal out-degree of any page, and (3) the *eigengap* of the co-citation matrix of the graph. Lee (2002) argued that an algorithm is stable if the L_1 change in its score vector following perturbations is bounded by a linear function of the sum of the scores of the perturbed nodes. For that definition, he showed that PageRank and Randomized HITS are stable for all graphs, SALSA is stable on authority-connected graphs, and HITS is not stable on authority-connected graphs (and hence is not stable on general graphs). Neither of these works examined how the perturbations affect the rankings that are induced by the score vectors. That aspect was looked upon by Chien et al. (2002), where it was shown that when a link is added to a graph, (1) the PageRank of the node receiving the link rises, and (2) the same node's rank cannot decrease with respect to its rank prior to the change.

This paper follows the approach introduced by Borodin et al. (2001), which examined how perturbations of the input graphs affect the output of the algorithms under a *worst-case* analysis. They considered arbitrary Web graphs (that do not adhere to any particular model or conditions), and examined whether it is possible to find instances on which algorithms are unstable. They also introduced the notion of *rank stability*, which measures the volatility of the rankings induced by the (assigned scores of the) algorithms. They applied the definitions to many algorithms, and attained many (mostly negative, instability) results.

In Ng et al. (2001), Borodin et al. (2001) and Lee (2002) it was noted that the stability of some algorithms may depend on whether the graph being analyzed is *authority connected*—a concept that was introduced in Borodin et al. (2001) and will be formally defined in Section 3. The authors of Borodin et al. (2001) raise the question whether their negative results remain true when the discussion is limited to authority connected graphs.

1.2. Rank similarity

As noted earlier, many link-based ranking algorithms were proposed over the last few years. This variety of algorithms raises several practical questions: are the results of different ranking algorithms substantially different on some (or on most) input graphs? Are some algorithms clearly more effective than others? Is it possible to characterize cases where algorithms outperform each other, or at least disagree with each other?

Due to the lack of an agreed-upon theoretical framework or experimental testbed, the effectiveness of algorithms was usually demonstrated by comparing the outputs of several algorithms on several queries (e.g., Bharat and Henzinger (1998), Chakrabarti et al. (2001) and Borodin et al. (2001)). As a by-product of such comparisons, it was noted that different algorithms often rank different pages as the top resources for the same query. In contrast, when Amento et al. (2000) compared HITS, PageRank and ranking by In-degree on several carefully constructed Web subgraphs, the three schemes produced very similar rankings. This unexpected similarity was due, according to the authors, to the manner in which the examined Web graphs were assembled.

Analytical and experimental evidence that HITS and SALSA may produce inherently different rankings was shown in the context of the *TKC (Tightly-Knit Community) Effect* ((Lempel and Moran 2001b), see also Section 3.3). There, an infinite set of graph instances on which the rankings of HITS and SALSA strongly disagree was constructed. Furthermore, real Web subgraphs on which the two algorithms produced disagreeing rankings due to this effect, were reported.

Borodin et al. (2001), in addition to examining the stability and rank-stability of individual algorithms, also defined the notions of *similar* and *rank similar* algorithms. These notions measure the resemblance between the scores/rankings produced by pairs of algorithms under a worst-case approach. Again, when applying these definitions to pairs of algorithms, most of their attained results were of negative nature (non-similarity).

1.3. This work

We extend the results of Borodin et al. (2001) by proving that HITS is not rank-stable on the class of authority-connected graphs. A similar result is shown for PageRank, which was not analyzed in Borodin et al. (2001). We then show that HITS and PageRank are not rank similar on this class, nor is any of them rank similar to SALSA.

This paper is organized as follows. Section 2 briefly describes the ranking algorithms PageRank, HITS and SALSA. Section 3 brings the definitions of rank-stability and rank-similarity from Borodin et al. (2001), and presents the known results concerning these notions. Section 4 details our extension of those results, and discusses the significance of these extensions. Section 5 brings our conclusions and ideas for future research.

2. Link-based ranking algorithms for Web pages

This section provides a brief overview of three link based ranking algorithms: PageRank (Brin and Page 1998), HITS (Kleinberg 1999) and SALSA (Lempel and Moran 2001b). We first bring an overview on the three algorithms and the differences between them. Then, we technically describe how each of the algorithms ranks the pages (=nodes) of a directed graph $G = (V, E)$ where $|V| = N$.

PageRank defines a random walk with random jumps over the (entire) Web graph. The states of the random walk are Web pages, and the score of each page is defined as its value in the stationary distribution of the random walk (Gallager 1996). Thus, the PageRank score of a page can be interpreted as a *global*, topic-independent importance rating of that page.

HITS and SALSA, on the other hand, are topic-specific, *local* ranking algorithms: they operate on a small portion of the Web where resources pertaining to a specific topic are likely to exist, by analyzing the link structure of that Web subgraph and assigning hub and authority scores to its pages. A page is an authority on a topic if it contains high quality, valuable information on it. A page is a hub on a topic if it links to good authorities on it, i.e. it is a list of quality resources on the topic. This paper focuses on the authority scores produced by the two algorithms; HITS defines the authority score of each page to be the corresponding value in the normalized principal eigenvector of the input graph's co-citation matrix. SALSA combines aspects from both HITS and PageRank, and performs a certain random walk which converges to the authority scores of the pages[3]. More on the relation between HITS and SALSA can be found in Borodin et al. (2001).

2.1. *PageRank*

PageRank (Brin and Page 1998) is an important part of the ranking function of the Google search engine. The PageRank of a page p (denoted $PR(p)$) is the probability of visiting p in a random walk of the entire Web, where the set of states of the random walk is the set of pages, and each random step is of one of two types:

1. Choose a Web page uniformly at random, and jump to it.
2. From the given state s, choose uniformly at random an outgoing link of s and follow that link to the destination page.

The first type of state transitions, the jumps to random Web pages, is needed since the Markov chain which is implied by the link-structure of the Web is separable rather than ergodic. In particular, it has absorbing states (pages that have no outgoing links). However, incorporating random jumps introduces a (small) probability of transition from any page a to any page b, even in absence of a Web link $a \to b$, thus giving rise to an ergodic Markov chain.

PageRank chooses a parameter $d,\ 0 < d < 1$, and each state transition is of the first transition type with probability d and of the second type with probability $1 - d$.[4] The PageRanks obey the following formula:

$$\mathrm{PR}(p) = \frac{d}{N} + (1-d)\left(\sum_{q:q \to p} \frac{\mathrm{PR}(q)}{\text{out degree of } q} \right)$$

Thus, the PageRank of a page grows with the importance (=PageRanks) of the pages which point to it. Since the random walk was defined on the entire Web, PageRank is a *global*, topic independent measure of each page's information content.

2.2. *Kleinberg's hyperlink-induced topic search (HITS)*

HITS assigns each page $s \in V$ a pair of weights, a hub-weight $h(s)$ and an authority weight $a(s)$, based on the following two principles:

- The quality of a hub is determined by the quality of the authorities it points at.
- The quality of an authority is determined by the quality of the hubs which link to it.

Technically, the weights are initialized to 1, and are updated by repeating the following three operations until convergence:

1. Update the authority weight of each page s (the $\mathcal{I}$ operation): $a(s) \leftarrow \sum_{\{x \mid (x,s) \in E\}} h(x)$
2. Update the hub weight of each page s (the $\mathcal{O}$ operation): $h(s) \leftarrow \sum_{\{x \mid (s,x) \in E\}} a(x)$
3. Normalize the authority weights and the hub weights.

Let W_G denote the adjacency matrix of G. $W_G^T W_G$ is the (symmetric) co-citation matrix of G. Pages are ranked according to their authority weights, which converge to the coordinates of the normalized principal eigenvector [5] of $W_G^T W_G$.

2.3. SALSA

SALSA, the Stochastic Approach for Link Structure Analysis, also assigns separate hub and authority scores to each page. These scores are based on two random walks performed on G, the *authority walk* and the *hub walk*. We describe here the authority walk. Intuition suggests that authoritative pages should be visible (linked) from many pages G. Thus, a random walk on this subgraph will visit those pages with high probability. Formally, the states of the authority walk are the nodes of G with at least one incoming link. Let v be such a node, and let $q_1, \ldots, q_k$ be the nodes that link to v. A transition from v involves picking a random index i uniformly over $\{1, 2, \ldots, k\}$, and selecting a new state from the outgoing links of q_i (again, randomly and uniformly). Thus, the transition involves traversing two Web links, the first of which is traversed backwards (from destination to source) and the second is traversed forwards. Let π denote the stationary distribution of the random walk described above, when the initial distribution is uniform over all states. The score of each page (=state) v is π_v (pages that have no incoming links attain a score of 0). It was shown in Lempel and Moran (2001b) that on authority connected graphs, π_v is directly proportional to the in-degree of v in G.

3. Definitions, notations and known results

Let $G = (V, E)$ be a directed graph representing a set of Web-pages and their interconnecting links. We now define the terms rank-similarity and rank-stability of link-based ranking algorithms for the Web, and the concept of authority-connected graphs (Borodin et al. 2001). Our definitions, although at times rephrased, are equivalent to those given in Borodin et al. (2001).

3.1. *Rank-stability and rank-similarity*

Definition 1. Let v_1, v_2 be N-dimensional real vectors. The *ranking distance*, d_r, between v_1 and v_2 is defined as follows:

$$d_r(v_1, v_2) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{I}_{v_1,v_2}(i,j)$$

$$\text{where } \mathbf{I}_{v_1,v_2}(i,j) = \begin{cases} 1 & v_1(i) < v_1(j) \quad \text{and } v_2(i) > v_2(j) \\ 0 & \text{otherwise} \end{cases}$$

For example,

$$d_r(\langle 2,4,6,8\rangle\ ,\ \langle 2,9,5,3\rangle) = \frac{3}{16}$$

due to the pairs $(i, j) \in \{\,(2,3), (2,4), (3,4)\,\}$. The ranking distance d_r is a normalized version of the *Kendall tau distance* between two rankings on the same set of objects (Diaconis 1988, Dwork et al. 2001).

In what follows, $\mathcal{G}$ denotes a set of directed graphs, and $\mathcal{G}_N$ is the subset of N-node graphs in $\mathcal{G}$. Let A_1 and A_2 be two link-based ranking algorithms which assign $|V|$-dimensional weight vectors $A_1(G)$, $A_2(G)$ to the nodes of the graph $G \in \mathcal{G}_N$. The weights of $A_i(G)(i = 1, 2)$ induce rankings on the nodes of G.

Definition 2. Two ranking algorithms A_1 and A_2 are *rank-similar* on $\mathcal{G}$ if [6]

$$\lim_{N\to\infty}\max_{G\in\mathcal{G}_N} d_r(A_1(G), A_2(G)) \longrightarrow 0$$

Intuitively, when two algorithms are found to be rank-similar, they should no longer be compared with each other in terms of search quality, but rather in terms of runtime performance (or ease of implementation). The more complex algorithm of the two usually becomes redundant.

Definition 3. An algorithm A is *rank stable* on $\mathcal{G}$ if for every fixed k, we have

$$\lim_{N\to\infty}\max_{\{G_1,G_2\in\mathcal{G}_N | d_e(G_1,G_2)\le k\}} d_r(A(G_1), A(G_2)) \longrightarrow 0$$

where $d_e(G_1, G_2) \stackrel{\triangle}{=} |(E_1 \cup E_2)\backslash(E_1 \cap E_2)|$.

Whereas in Definition 2 we compare the rankings induced by two different algorithms on the nodes of the same graph, Definition 3 fixes the algorithm and compares the rankings it induces on the nodes of a pair of graphs that differ in at most k edges.

Note that while the above concepts were defined in terms of link-based algorithms, corresponding ideas can be developed and studied in the context of classic IR as well. For example, fixing a corpus and a query, one can compare the document rankings induced by different similarity measures and retrieval models.

3.2. Authority-connected graphs: Definition and motivation

Let $G = (V, E)$ be a directed graph (representing some Web-subgraph). Two nodes $p, q \in V$ are *co-cited* if there exists a node r that links to both p and q. We say that p and q are connected by a *co-citation path* if there exist nodes $p = v_0, v_1, \ldots, v_{k-1}, v_k = q$ such that (v_{i-1}, v_i) are co-cited for all $i = 1, \ldots, k$. V_{in} will denote all nodes in V with at least one incoming edge.

Definition 4. A directed graph $G = (V, E)$ is called *authority connected* if for all $p, q \in V_{in}$, there exists a co-citation path connecting p and q.

We will examine rank stability and rank similarity on the class of authority connected graphs.

One of the most basic premises of link analysis is that a co-citation of two pages implies a topical connection between them. Thus, when two pages (with at least one incoming link) are connected by a co-citation path, there is a "semantic line of thought" that leads from the idea expressed in the first page to the topic covered in the second. Conversely, distinct authority-connected components of a graph conceptually correspond to neighborhoods of pages that pertain to different topics or concepts: not a single page links to pages of two distinct components. Thus, asking a link-based algorithm to rank the pages of graphs that are not authority connected is basically asking it to compare the relative importance of pages on unrelated concepts ("is p_1 a better geography resource than p_2 is an authority on sports"). By examining rankings on authority connected graphs, we ensure that the relevance of all pages will be measured with respect to the same bar. For more on the significance of such graphs, see Section 4.1.

3.3. Known results

3.3.1. Rank-stability and rank-similarity. Let $\bar{\mathcal{G}}$ denote the set of all directed graphs. The following were shown in Borodin et al. (2001): (1) HITS is not rank-stable on $\bar{\mathcal{G}}$, (2) HITS and SALSA are not rank-similar on $\bar{\mathcal{G}}$, and (3) SALSA is not rank-stable on $\bar{\mathcal{G}}$ *but is* rank-stable and L_1-stable on authority-connected graphs.

The instability and non-similarity results were shown on graphs with two disjoint components. Such graphs, which are obviously not authority connected, conceptually correspond to collections of Web pages that pertain to multiple, unrelated topics. For the instability results, the small perturbations considered in the proofs caused HITS and SALSA to shift their preference from the pages of one component to those of the other. The non-similarity result involved proving that HITS and SALSA prefer pages of different components. The authors of Borodin et al. (2001) leave open the question whether the negative results (1), (2) remain true when the discussion is limited to authority connected graphs. We answer this affirmatively in Section 4.

3.3.2. The Tightly-Knit Community (TKC) effect. The TKC effect highlights an important difference between HITS and SALSA: HITS favors groups of pages that have many "internal" co-citations, while SALSA prefers pages with many inlinks. A tightly-knit community is a small but highly interconnected set of pages. Roughly speaking, the TKC effect occurs when such a community scores high in link-analyzing algorithms, even though its pages are not authoritative on the topic, or pertain to just one aspect of the topic.

It was shown in Lempel and Moran (2001b), both theoretically and experimentally, that HITS is more sensitive than SALSA to this effect. The theoretical analysis involved the construction of authority-connected graphs containing both a small tightly knit community (with many interconnecting hubs) and a large, less densely connected community. It was proven that HITS ranks the authorities of the small, tightly knit community higher than it ranks the authorities of the larger community, while SALSA prefers the authorities of the larger community. The demonstrated qualitative difference between the two algorithms, however, does not prove that the algorithms are not rank similar under the definition given in the previous subsection: the ranking distance d_r demonstrated by the constructed graphs is $\mathcal{O}(\frac{1}{\sqrt{N}})$, which approaches zero as N (the number of pages with at least one incoming link) grows. The experimental part in Lempel and Moran (2001b) involved the collection of several real-life Web graph, on which HITS and SALSA gave different results. It was shown there that the rankings of HITS were biased towards the pages of tightly-knit communities.

4. Results

Our work focuses on the class of authority connected graphs, which we denote by $\mathcal{G}^{AC}$. Section 4.2 shows that neither HITS nor PageRank is rank stable on $\mathcal{G}^{AC}$. Thus, SALSA is the only algorithm of the three algorithms we consider here that is rank stable on $\mathcal{G}^{AC}$. Furthermore, we show there that no pair of these algorithms is rank similar on $\mathcal{G}^{AC}$. The results concerning HITS and SALSA complement the knowledge on general graphs (Section 3.3); PageRank was not previously examined in this context. Before diving into the mathematical details of the proofs, Section 4.1 discusses the significance of our results.

4.1. Significance of results

4.1.1. Focusing on authority-connected graphs. As detailed in Section 3, today's prevailing link analysis algorithms rank Web pages according to the entries of the principal eigenvector of some matrix representation of the corresponding web graphs. Accordingly, the algorithms differ only in the manner in which these matrices are constructed. Specifically, PageRank uses a stochastic version of the adjacency matrix of the Web, HITS uses the co-citation matrix, and SALSA uses a stochastic version of the co-citation matrix of the corresponding graph. It is possible to identify two extreme approaches in papers studying properties of the rankings induced by these algorithms:[7] one approach, used for proving positive results, assumes that the Web's structure, and hence the resulting matrices, obey certain regular, well behaved stochastic properties (e.g., Azar et al. (2001) and Achlioptas et al. (2001)). The other approach, used in Borodin et al. (2001) when proving negative results, considers Web graphs—and hence, matrices—of arbitrary structure. This paper attempts to narrow the gap between these approaches by restricting the worst case analysis of Borodin et al. (2001) to a class of well behaved graphs—the authority connected graphs, which are exactly those graphs whose co-citation matrices are irreducible.[8]

Informally, authority-connected graphs represent a natural testbed for examining theoretical properties of link-based algorithms, since every two possible authorities (pages with at least one inlink) are connected by a co-citation path, and thus are thematically related. Furthermore, from a technical point of view, it can be argued that studying the scores and rankings of HITS on authority-disconnected graphs, is of limited interest. Perron-Frobenius theory (Gallager 1996, Horn and Johnson 1985) implies that HITS will assign positive scores to all authorities in authority-connected graphs. However, when the graph contains distinct authority-connected components, HITS will assign zero scores to the pages in all but one of these components.[9] This means that any perturbation which disconnects a component or merges distinct components will cause significant changes in the induced scores and rankings. Indeed, the rank-instability/nonsimilarity for HITS and SALSA on general graphs in Borodin et al. (2001) involved such perturbations.

The significance of restricting matrix-based link analyses to irreducible matrices has been noted before. PageRank was purposely defined in a manner which guarantees that the resulting stochastic matrix is irreducible, by incorporating random jumps into the random walk over the Web's graph (Brin and Page 1998). As the underlying graph of the Web is not strongly connected (Broder et al. 2000), had PageRank's random walk been limited to traversing Web links alone, it would not have been ergodic (Gallager 1996). The addition of random jumps effectively links every page to all other pages, implying strong connectivity, and in turn, ergodicity. Several proposed variations of HITS avoided the singularities of reducible matrices that stem from authority-disconnected graphs by applying HITS on modified versions of the co-citation matrix. Randomized HITS, for example, incorporated random jumps (in the spirit of PageRank) into HITS. These jumps essentially transform every graph into an authority-connected graph, and contribute to the stability of Randomized HITS' scores (Ng et al. 2001, Lee 2002). Farahat et al. (2001) used exponentiated input to HITS in order to transform every weakly connected Web graph into an authority-connected graph.

4.1.2. Revisiting the definition of the ranking distance. The definitions of rank stability and rank similarity are based on graph pairs which attain the highest rank distance. It is not clear, however, that the ranking distance d_r is an appropriate measure of the difference of two rankings. The current definition is unweighted: any time the two compared rankings disagree on the relative ranking of a pair of pages, the pair contributes 1 to d_r. One could argue that on the Web, rank changes at the top of the rankings should count much more than rank changes closer to the bottom: for many queries on search engines, many thousands of results are ranked. However, users only view the very top of the list of results, and any change in the order of the bottom half of the rankings is inconsequential and will go unnoticed.

While alternative definitions of the ranking distance might better suit the problem of ranking Web pages, we argue that our results will stand under any reasonable definition.

In all of our results, the differences between the two compared rankings involve (at the very least) swapping the bottom half of the rankings with its upper half. In some cases (see Proposition 1), the two rankings are essentially opposites. Since any reasonable function will judge such extreme changes to the upper half of the rankings as different, our results are applicable to alternative definitions of the ranking distance.

4.1.3. Stability of scores does not imply rank stability. The notion of stability of an algorithm, which pertains to the volatility of the scores which the algorithm assigns to pages, has been studied more than the corresponding notion of rank-stability. In particular, there are positive results concerning stability where there are no results at all concerning rank-stability. For example, Lee (2002) showed that the PageRank algorithm produces stable scores: when the perturbation of the graph involves changing the outlinks of a single page p, the L_1 change in the score vector of PageRank is bounded by a linear function of the PageRank of p (prior to the change). An interesting question is whether stability implies rank-stability, and Proposition 2 shows that this is not the case. There, a change in one outlink of a very low ranking page essentially turns the entire rankings upside down.

4.2. Proofs of results

This section provides the technical details of our rank-nonsimilarity and rank-instability results. When proving that two algorithms are not rank-similar, we construct infinitely many examples where the two rankings being compared differ significantly. Similarly, when proving that an algorithm is not rank-stable, we construct infinitely many cases where the relative order of a significant fraction of the $\binom{N}{2}$ pairs of nodes is reversed.

Proposition 1. *HITS is not rank stable on the class of authority connected graphs $\mathcal{G}^{AC}$.*

Proof: Consider the following graphs G_1 and G_2 (G_2 is shown in figure 1). Both graphs contain $N \triangleq 2n + 3$ nodes: n authorities named $a_1, a_2, \ldots, a_n$, $n + 1$ "fixed" hubs named $h_0, h_1, \ldots, h_n$, and 2 "flipping" hubs h^*, h^{**}. Both graphs contain the following $2n$ links:

- $h_0 \rightarrow a_1, h_n \rightarrow a_n$.
- For all $i = 1, \ldots, n - 1$: $h_i \rightarrow a_i, h_i \rightarrow a_{i+1}$.

Clearly, G_1 and G_2 are authority connected. The difference between the graphs is that in G_1, both h^* and h^{**} link to a_1, while in G_2 these two flipping hubs link to a_n. Note that G_1 and G_2 are isomorphic, where the unique isomorphism between them involves reversing the identities of the n authorities and of the $n + 1$ fixed hubs.

Let $x_1, \ldots, x_n$ denote the HITS authority weights of $a_1, \ldots, a_n$ under G_2. Recall that these weights are the entries of the principal eigenvector of the co-citation matrix of G_2 (see Section 2.2). In what follows we prove that $x_1 < x_2 < \cdots < x_n$. By the isomorphism of G_1 and G_2, we conclude that the rankings of $a_1, \ldots, a_n$ on G_1 and G_2 are completely

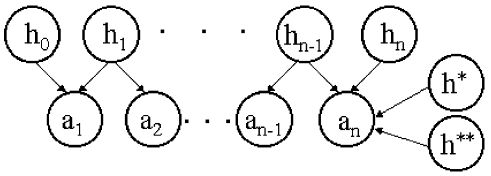

Figure 1. The graph G_2.

reversed and are in complete disagreement. Hence,

$$d_r(\mathit{HITS}(G_1), \mathit{HITS}(G_2)) = \frac{n(n-1)}{2(2n+3)^2}, \quad d_e(G_1, G_2) = 4$$

proving the non-rank-stability of HITS on $\mathcal{G}^{AC}$.

By the definition of G_2, the (irreducible) co-citation matrix of the n authorities is as follows:

$$\begin{pmatrix}
2 & 1 & 0 & 0 & 0 & \dots & & & & & 0\\
1 & 2 & 1 & 0 & 0 & \dots & & & & & 0\\
0 & 1 & 2 & 1 & 0 & \dots & & & & & 0\\
 & & \vdots & & & \dots & & & \vdots & & \\
0 & & & & & \dots & 0 & 1 & 2 & 1 & 0\\
0 & & & & & \dots & 0 & 0 & 1 & 2 & 1\\
0 & & & & & \dots & 0 & 0 & 0 & 1 & 4
\end{pmatrix}$$

Denote the principal eigenvalue of the matrix by λ. By the Perron-Frobenius theorem (Gallager 1996, Horn and Johnson 1985), λ, as well as $x_1, \dots, x_n$, are positive. This easily implies that λ is greater than any element on the main diagonal of the matrix, hence $\lambda > 4$.

The first line of the co-citation matrix implies the following equation on x_1 and x_2:

$$2x_1 + x_2 = \lambda x_1 \;\Rightarrow\; x_2 = (\lambda - 2)x_1 \;\;\Rightarrow x_2 > x_1$$

We now proceed by induction to show that $x_{i+1} > x_i$ for $i = 2, \dots n-1$. By the i'th row of the co-citation matrix,

$$x_{i-1} + 2x_i + x_{i+1} = \lambda x_i \;\Rightarrow\; x_{i+1} = (\lambda - 2)x_i - x_{i-1} > 2x_i - x_{i-1} > x_i\ .$$

Therefore, $x_1 < x_2 < \cdots < x_n$. □

In Lempel and Moran (2001a) we use a construction similar to that of Proposition 1 to prove that HITS' scores are not L_1-stable on authority-connected graphs, extending a previous result concerning HITS' L_1-instability on general graphs from Borodin et al. (2001).

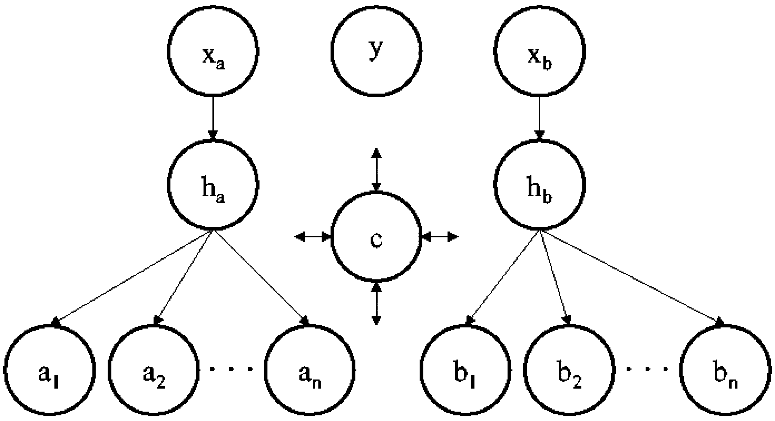

Figure 2. The graph G on which G_a and G_b are based.

Proposition 2. *PageRank is not rank stable on the class of authority connected graphs* $\mathcal{G}^{AC}$.

Proof: Consider the following graph $G = (V, E)$ (shown in figure 2):

$$\begin{aligned} V &= \{c,\ x_a, y, x_b,\ h_a, h_b,\ a_1, a_2, \ldots, a_n,\ b_1, b_2, \ldots, b_n\} \\ E &= \{x_a \to h_a, x_b \to h_b\} \cup \{h_a \to a_i, h_b \to b_i \mid i = 1, \ldots, n\} \\ &\quad \cup \{c \to v, v \to c \mid v \in V \setminus \{c\}\} \end{aligned}$$

Define $G_a \triangleq (V, E \cup \{y \to h_a\})$, $G_b \triangleq (V, E \cup \{y \to h_b\})$. Both G_a and G_b are authority connected (through the connectivity of the vertex c).

Let $PR_a(v), PR_b(v)$ $(v \in V)$ denote the PageRank of v in G_a, G_b respectively. From the definition of PageRank, it is easy to see that

$$0 < PR_a(x_a) = PR_a(y) = PR_a(x_b)\ , \text{ and so } PR_a(h_a) > PR_a(h_b)\ .$$

Therefore, $PR_a(a_i) > PR_a(b_i)$ for all $1 \leq i \leq n$. Similarly, $PR_b(a_i) < PR_b(b_i)$ for all $1 \leq i \leq n$. Thus,

$$d_r(PageRank(G_a), PageRank(G_b)) = \frac{n^2}{(2n+6)^2}, \quad d_e(G_a, G_b) = 2$$

and the result follows.

Observe that $\forall p \in \{h_a, h_b,\ a_1, \ldots, a_n,\ b_1, \ldots, b_n\}$, $PR(y) < PR(p)$ (in either graph). Thus, the dramatic change in the rankings was caused by shifting the single outlink of y, which is a very low-ranking node. □

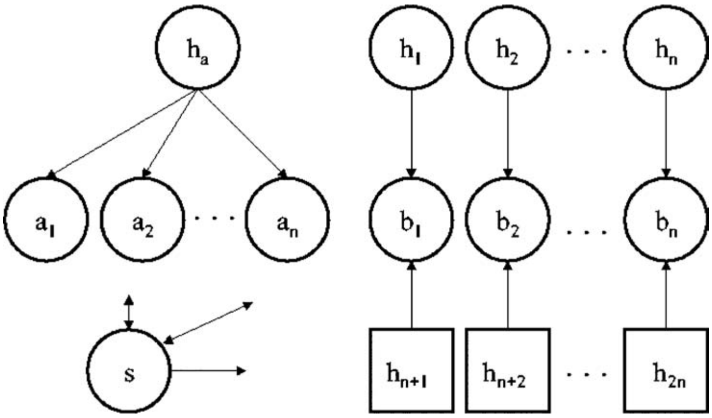

Figure 3. The graph G_3 (note that the square nodes do not link back to s).

Proposition 3. *HITS, PageRank are not rank similar on the class of authority connected graphs.*

Proof: Consider the following graph $G_3 = (V, E)$ (shown in figure 3). For ease of notation, we denote $A = \{a_1, \dots, a_n\}$ and $B = \{b_1, \dots, b_n\}$.

$$\begin{aligned} V &= \{s,\ \ h_a,\ \ h_1, h_2, \dots, h_{2n}\} \cup A \cup B \\ E &= \{s \to v \mid v \in V \setminus \{s\}\} \ \cup\ \{h_a \to a \mid a \in A\} \ \cup\ \{h_a \to s\} \\ &\quad \cup \{h_i \to b_i, h_{i+n} \to b_i \mid i = 1, \dots, n\} \\ &\quad \cup \{a_i \to s, b_i \to s, h_i \to s \mid i = 1, \dots, n\} \end{aligned}$$

Thus, G_3 consists of $N \triangleq 4n + 2$ nodes. It is easy to see that G_3 is authority-connected (through the connectivity of the node s).

We examine the relative rankings of the A nodes and the B nodes. We first note that HITS ranks the A-nodes higher than it ranks the B-nodes. The proof relies on the structure of the rows that correspond to the A and B nodes in the co-citation matrix of G_3, $M_{G_3} = [m_{i,j}]$:

$$m_{i,j} = m_{j,i} = \begin{cases} 2 & i, j \in A \\ 1 & i \in A, j \in V \setminus A \\ 3 & i, j \in B, i = j \\ 1 & i \in B, j \in V \setminus \{i\} \end{cases}$$

HITS' authority weights of all A-nodes will be equal, and will be denoted by a. Likewise, the (all-equal) authority scores of the B-nodes will be denoted by b. The authority scores of all nodes are positive. In particular, $a, b > 0$. Let λ denote the principal eigenvalue of

M_{G_3}. The rows in M_{G_3} that correspond to the A and B nodes give rise to the following two equations:

$$\lambda a = 2na + nb + T \tag{1}$$

$$\lambda b = na + (n+2)b + T \tag{2}$$

Where T is the sum of the (positive) authority scores of the $V \setminus (A \cup B)$-nodes. The first equation implies that $\lambda > 2n$. Subtracting (2) from (1), we have (for all $n > 2$)

$$na - 2b = \lambda(a-b) \Rightarrow \frac{a}{b} = \frac{\lambda - 2}{\lambda - n} > 1 \Rightarrow a > b$$

Thus, HITS ranks the A-nodes higher than it ranks the B-nodes.

As for PageRank, by arguments similar to those of Proposition 2, we have

$$PR(h_a) = PR(h_i) \quad i = 1, 2, \ldots, 2n$$

Since b_i is linked to by s, h_i and h_{i+n} while a_i is linked to by s and h_a, we conclude that PageRank prefers the B nodes over the A nodes. Thus,

$$d_r(\text{HITS}(G_3), \text{PageRank}(G_3)) \geq \frac{n^2}{(4n+2)^2} \ ,$$

proving that the two algorithms are not rank similar. □

Proposition 4. *HITS and SALSA are not rank similar on the class of* authority connected *graphs.*

Proof: Consider the graph G_3 defined in Proposition 3. Since G_3 is authority-connected, SALSA will rank the nodes by their in-degree. The in-degree of all $a \in A$ is 2 while the in-degree of each $b \in B$ is 3. Thus, SALSA (like PageRank) prefers the B-nodes over the A-nodes while HITS prefers the A-nodes over the B-nodes, and the result follows. □

Proposition 5. *PageRank and SALSA are not rank similar on the class of* authority connected *graphs* $\mathcal{G}^{AC}$.

Proof: Let d be the random jump parameter of PageRank and let $t \triangleq \lceil \frac{4}{3} + \frac{2}{1-d} \rceil$. For all $n \geq t$, consider the following graph $G_5 = (V, E)$ (again, $A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_n\}$):

$$\begin{aligned} V &= \left\{s,\ x_1, x_2, \ldots, x_t,\ y,\ h_a, h_b^1, h_b^2\right\} \cup A \cup B \\ E &= \{x_i \to h_a \mid i = 1, \ldots, t\} \cup \left\{y \to h_b^1, y \to h_b^2\right\} \\ &\quad \cup \{h_a \to a \mid a \in A\} \cup \left\{h_b^1 \to b, h_b^2 \to b \mid b \in B\right\} \\ &\quad \cup \left\{s \to v \mid v \in V \Big\backslash \left\{s, h_b^2\right\}\right\} \cup \{v \to s \mid v \in V \setminus \{s\}\} \end{aligned}$$

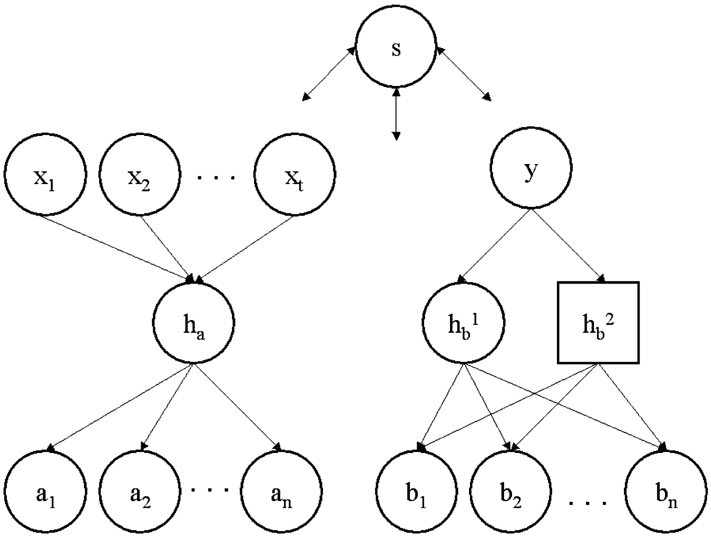

Figure 4. The graph G_5 (note that s does not link to the square node h_b^2).

G_5 contains $N \triangleq 2n + t + 5 \leq 3n + 5$ nodes (see figure 4), and is clearly authority connected. Since the in-degree of all A-nodes is 2 while the in-degree of all B-nodes is 3, SALSA ranks the B-nodes higher than it ranks the A-nodes.

Let $PR(v)$ denote the PageRank of node v. By the definition of PageRank,

$$\begin{aligned} PR(a_1) &= PR(a_2) = \cdots = PR(a_n) \\ &= \frac{d}{N} + (1-d)\left[\frac{PR(h_a)}{n+1} + \frac{PR(s)}{N-2}\right] \end{aligned} \tag{3}$$

$$\begin{aligned} PR(b_1) &= PR(b_2) = \cdots = PR(b_n) \\ &= \frac{d}{N} + (1-d)\left[\frac{PR(h_b^1) + PR(h_b^2)}{n+1} + \frac{PR(s)}{N-2}\right] \end{aligned} \tag{4}$$

Subtracting (4) from (3), we have

$$\begin{aligned} PR(a_1) - PR(b_1) &= \cdots = PR(a_n) - PR(b_n) \\ &= \frac{1-d}{n+1}\left[PR(h_a) - PR(h_b^1) - PR(h_b^2)\right] \end{aligned}$$

Therefore, proving $PR(h_a) - PR(h_b^1) - PR(h_b^2) > 0$ will suffice to show that PageRank prefers the A-nodes over the B-nodes.

Let $p \overset{\triangle}{=} PR(y)$. Since $PR(y) = PR(x_1) = PR(x_2) = \cdots = PR(x_t)$, we have

$$PR(h_a) = \frac{d}{N} + (1-d)\left[\frac{tp}{2} + \frac{PR(s)}{N-2}\right] \tag{5}$$

$$PR\left(h_b^1\right) + PR\left(h_b^2\right) = \frac{2d}{N} + (1-d)\left[\frac{2p}{3} + \frac{PR(s)}{N-2}\right] \tag{6}$$

Note that $p > \frac{d}{N}$ (like the PageRank of every node in G_5). Thus, subtracting (6) from (5) yields

$$\begin{aligned} PR(h_a) - PR\left(h_b^1\right) - PR\left(h_b^2\right) &= p\,(1-d)\left(\frac{t}{2} - \frac{2}{3}\right) - \frac{d}{N} \\ &> \frac{d}{N}\left[(1-d)\left(\frac{t}{2} - \frac{2}{3}\right) - 1\right] \end{aligned}$$

By our choice of t,

$$t \geq \frac{4}{3} + \frac{2}{1-d}, \quad \text{and so} \quad \frac{t}{2} \geq \frac{2}{3} + \frac{1}{1-d}, \quad \text{or} \quad (1-d)\left(\frac{t}{2} - \frac{2}{3}\right) \geq 1$$

which proves that $PR(h_a) - PR(h_b^1) - PR(h_b^2) > 0$.

We conclude that $d_r(\text{SALSA}(G_5), \text{PageRank}(G_5)) \geq \frac{n^2}{(3n+5)^2}$, and the result follows. □

5. Conclusions

This work examined the notions of rank-stability and rank-similarity of link-based ranking algorithms on authority-connected graphs. Three specific algorithms were examined: PageRank, HITS and SALSA. Previous work has already shown that SALSA is rank-stable on authority-connected graphs. In this paper it was shown that both PageRank and HITS are not rank-stable on authority-connected graphs, and that no pair of the three algorithms is rank-similar on such graphs.

As noted in the Introduction, rank-instability and rank-nonsimilarity are worst-case notions. While our results do not necessarily reflect the instability or non-similarity of PageRank, HITS or SALSA on the "typical" Web graph, they do provide theoretical insight on why some of these algorithms are potentially vulnerable to link spamming attacks—as demonstrated experimentally in Lempel and Moran (2001b). This research should be complemented by the average-case analysis of these (and other) algorithms, coupled with the study of realistic models for the Web graph—an area of research which has seen much activity (Kleinberg et al. 1999, Pandurangan et al. 2002, Azar et al. 2001, Achlioptas et al. 2001, Kumar et al. 2000, Ruhl et al. 2001). Another issue of importance, under-explored so far in the IR literature, is a methodical analysis of the stability of the running times of the algorithms as their input is perturbed (see, for example, Dominich and Tuza (2003)).

Stability of running times is especially important in the context of link-analyses on graphs of Web scale, as these algorithms demand considerable computational resources. In terms of HITS and PageRank, studying running time stability translates to investigations of how the eigengaps of the respective matrices fluctuate as the Web graph undergoes minor changes.

One particular research direction we find interesting is examining the possible rank-similarity of PageRank and SALSA on the real Web. It is well-known that the distribution of in-degrees of Web pages follows a power-law (Barabasi and Albert 1999, Kleinberg et al. 1999, Broder et al. 2000). A study by Pandurangan et al. (2002) revealed that the distribution of PageRank also follows a power-law. Furthermore, the exponent of both distributions is the same (2.1), and so these two measures essentially follow the same distribution. While the motivation behind PageRank and HITS was that in-degree is not a sufficient indication of a page's authority or importance, the identical distributions of the in-degrees and PageRank suggest that ultimately, in-degree might be an effective approximation to PageRank. In the terms used in this paper, the ranking distance d_r between PageRank and the in-degree measure (the fraction of pairs whose relative ranking is different between the two measures) could very well be small. Carrying this argument further, the ranking distance between SALSA and PageRank on real Web graphs[10] could also be small. We leave this for future experimental research.

Notes

1. http://www.google.com
2. http://www.altavista.com
3. While both HITS and SALSA assign hub and authority scores to each page, they use different approaches and there is a qualitative difference between the scores (of either type) produced by them.
4. This rule assumes that all Web pages have at least one outgoing link. This will indeed be the case in all the examples concerning PageRank given in this paper.
5. The (normalized) eigenvector which corresponds to the eigenvalue of highest magnitude.
6. In this and the following definition it is assumed that the max operation is performed on nonempty sets.
7. As opposed to studies of stability properties of the *scores* produced by the algorithms, surveyed in Section 1.
8. Note that negative (rank instability/nonsimilarity) results on such graphs trivially imply negative results for the entire set of graphs as well.
9. As long as the principal eigenvalue of the co-citation matrix of the graph has multiplicity 1.
10. SALSA is not equivalent to the in-degree measure on graphs that are not authority-connected.

References

Achlioptas D, Fiat A, Karlin A and McSherry F (2001) Web search through hub synthesis. In: Proc. 42nd Annual Symposium on Foundations of Computer Science (FOCS 2001), Las Vegas, Nevada.

Amento B, Terveen L and Hill W (2000) Does "Authority" mean quality? predicting expert quality ratings of web documents. In: Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece.

Arasu A, Cho J, Garcia-Molina H, Paepcke A and Raghavan S (2001) Searching the Web. ACM Transactions on Internet Technology, 1(1):2–43.

Azar Y, Fiat A, Karlin A, McSherry F and Saia J (2001) Spectral analysis of data. In: Proc. 33rd annual ACM Symposium on Theory of Computing (STOC 2001), Crete, Greece.

Barabasi A-L and Albert R (1999) Emergence of scaling in random networks. Science, 286:509–512.

Bharat K and Henzinger MR (1998) Improved algorithms for topic distillation in a hyperlinked environment. In: Proc. 21'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Borodin A, Roberts GO, Rosenthal JS and Tsaparas P (2001) Finding authorities and hubs from link structures on the world wide web. Submitted for publication. Extended abstract appeared in Proc. 10th International World Wide Web Conference, pp. 415–429.

Brin S and Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Proc. 7th International WWW Conference, pp. 107–117.

Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A and Wiener J (2000) Graph structure in the web. In: Proc. 9th International WWW Conference, pp. 309–320.

Chakrabarti S, Dom B, Gibson D, Kleinberg J, Kumar S, Raghavan P, Rajagopalan S and Tomkins A (1999a) Hypersearching the web. Scientific American.

Chakrabarti S, Dom B, Gibson D, Kleinberg J, Kumar S, Raghavan P, Rajagopalan S and Tomkins A (1999b) Mining the link structure of the WWW. IEEE Computer.

Chakrabarti S, Dom B, Gibson D, Kleinberg JM, Raghavan P and Rajagopalan S (1998a) Automatic resource list compilation by analyzing hyperlink structure and associated text. In: Proc. 7th International WWW Conference.

Chakrabarti S, Dom B, Gibson D, Kumar S, Raghavan P, Rajagopalan S and Tomkins A (1998b) Spectral filtering for resource discovery. In: ACM SIGIR workshop on Hypertext Information Retrieval on the Web.

Chakrabarti S, Joshi M and Tawde V (2001) Enhanced topic distillation using text, markup tags, and hyperlinks. In: Proc. 24'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 208–216.

Chien S, Dwork C, Kumar R, Simon D and Sivakumar D (2002) Link evolution: Analysis and algorithms. In: Workshop on Algorithms and Models for the Web Graph (WAW). Vancouver, Canada.

Davison BD (2000) Recognizing nepotistic links on the web. Technical Report WS-00-01, Artificial Intelligence for Web Search.

Diaconis P (1988) Group Representation in Probability and Statistics. IMS Lecture Series 11, Institute of Mathematical Statistics.

Dominich S and Tuza Z (2003) Computational aspects of connectionist interaction information retrieval. In: Proc. ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR). Toronto, Canada.

Dwork C, Kumar R, Naor M and Sivakumar D (2001) Rank aggregation methods for the web. In: Proc. 10th International World Wide Web Conference, pp. 613–622.

Farahat A, LoFaro T, Miller JC, Rae G, Schaefer F and Ward LA (2001) Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records. In: Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA.

Gallager RG (1996) Discrete Stochastic Processes. Kluwer Academic Publishers.

Haveliwala TH (2002) Topic-Sensitive PageRank. In: Proc. 11th International WWW Conference (WWW2002).

Henzinger MR, Motwani R and Silverstein C (2002) Challenges in web search engines. SIGIR Forum, 36(2).

Horn RA and Johnson CR (1985) Matrix Analysis. Cambridge University Press.

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J. ACM, 46(5):604–632.

Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S and Tomkins AS (1999) The web as a graph: Measurements, models and methods. In: Proc. of the Fifth International Computing and Combinatorics Conference, pp. 1–17.

Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins AS and Upfal E (2000) Stochastic models for the web graph. In: Proc. 41st Annual Symposium on Foundations of Computer Science (FOCS 2000), Redondo Beach, California. pp. 57–65.

Lee HC (2002) When the hyperlinked environment is perturbed. In: Workshop on Algorithms and Models for the Web Graph (WAW), Vancouver, Canada.

Lempel R and Moran S (2001a) Rank-stability and rank-similarity of web link-based ranking algorithms. Technical Report CS-2001-22 (revised version), Dept. of Computer Science, Technion—Israel Institute of Technology.

Lempel R and Moran S (2001b) SALSA: The stochastic approach for link-structure analysis. ACM Transactions on Information Systems, 19(2):131–160.

Marchiori M (1997) The quest for correct information on the web: Hyper search engines. In: Proc. 6th International WWW Conference.

Ng AY, Zheng AX and Jordan MI (2001) Stable algorithms for link analysis. In: Proc. 24'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 258–266.

Pandurangan G, Raghavan P and Upfal E (2002) Using pageRank to characterize web structure. In: Proc. 8th Annual International Computing and Combinatorics Conference, pp. 330–339.

Ruhl M, Bharat K, Chang B-W and Henzinger M (2001) Who links to whom: Mining linkage between web sites. In: IEEE International Conference on Data Mining (ICDM), pp. 51–58.

Silva I, Ribeiro-Neto B, Calado P, Moura E and Ziviani N (2000) Link-based and content-based evidential information in a belief network model. In: Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 96–103.