Rank-Stability and Rank-Similarity of Web Link-Based Ranking Algorithms

Ronny Lempel        Shlomo Moran
Department of Computer Science
The Technion,
Haifa 32000, Israel
email: {rlempel,moran}@cs.technion.ac.il

September 17, 2002

ABSTRACT

Web search algorithms that rank Web pages by examining the link structure of the Web are attractive from both theoretical and practical aspects. 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.

Rank-stability and rank-similarity are of special interest on the class of authority connected graphs. Such graphs conceptually correspond to thematically related collections, in which most pages pertain to a single, dominant topic of interest.

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.
1 Introduction

The link-structure of the Web has been the focus of much research over the last few years. In particular, many novel algorithms for performing Web search and retrieval based on link-structure analysis were proposed (eg. HITS[13], PageRank[6], SALSA[16]). These algorithms demonstrated empirically that link-structure analysis can improve search and rank techniques on the Web ([2],[21]). Accordingly, search engines (such as Google\(^1\) and AltaVista\(^2\)) nowadays incorporate extensive link analysis in their ranking algorithms, that determine the order in which the search results are displayed to the user [6].

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. This paper focuses on two properties: rank-stability and rank-similarity, which are described next.

Stability and Rank Stability Stability is an important quality of any information retrieval algorithm, and link-based ranking algorithms are no exception. The Web’s graph 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, however, must cope with more than the natural evolution of the Web. With the growing phenomenon of search engine spamming [17, 11], 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 [16, 9]. Link spammers are Web authors that create pages and interconnecting links with the intention of biasing search engine link-based rankings of

\(^1\)http://www.google.com

\(^2\)http://www.altavista.com
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 [5], 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. [3] and Achlioptas et al. [1] - assumes that the Web's structure obeys some topic-driven stochastic models. In [3] it is argued that HITS is stable under a certain model of the Web's link structure. In [1], 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. [18] 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.
This paper follows the approach introduced by Borodin et al. in [5], 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.

Both [18] and [5] noted that the stability of some algorithms may depend on whether the graph being analyzed is authority connected - a concept that was introduced in [5] and will be formally defined in Section 3. The authors of [5] raise the question whether their negative results remain true when the discussion is limited to authority connected graphs.

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 (eg. [4],[8],[5] and many others). 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. [2] 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 ([16], 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 produce disagreeing rankings due to this effect, were reported.

Borodin et al. [5], 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).

**This work** We extend the results of [5] by proving that neither HITS nor PageRank is rank-stable on the class of authority connected graphs. We then show that HITS and PageRank are not rank similar on this class, nor is any of them rank similar to SALSA. We also examine the score vectors produced by HITS, proving that HITS is not \(L_1\)-stable.

This paper is organized as follows. Section 2 briefly describes the ranking algorithms PageRank, HITS and SALSA. Section 3 brings the definitions of \(L_1\)-stability, rank-stability and rank-similarity from [5], and presents the known results concerning these notions. Section 4 details our extension of those results. 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 [6], HITS [13] and SALSA [16]. 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. 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 perform a certain random walk which converges to the authority scores of the pages. More on the relation between HITS and SALSA can be found in [5].
The following describes how each of the algorithms ranks the pages (=nodes) of a graph \( G = (V, E) \) where \(|V| = N\).

### 2.1 PageRank

PageRank [6] 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.

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 \( 1 - d \). \(^3\) The PageRanks obey the following formula:

\[
PR(p) = \frac{d}{N} + (1 - d) \left( \sum_{q:q\rightarrow p} \frac{PR(q)}{\text{out degree of } q} \right)
\]

### 2.2 Kleinberg’s Hyperlink-Induced Topic Search (HITS)

Each page \( s \in V \) is assigned a pair of weights, a hub-weight \( h(s) \) and an authority weight \( a(s) \). 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 \( I \) operation): \( a(s) \leftarrow \sum_{\{x|(x,s)\in E\}} h(x) \)

2. Update the hub weight of each page \( s \) (the \( O \) operation): \( h(s) \leftarrow \sum_{\{x|(s,x)\in E\}} a(x) \)

\(^3\)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.
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 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. Its states 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).

We restrict our discussion of SALSA in this context to authority connected graphs. Let $\pi$ denote the stationary distribution of the random walk described above. The score of each page (=state) $v$ is $\pi_v$ (pages that have no incoming links attain a score of 0). It was shown in [16] that on authority connected graphs, $\pi_v$ is directly proportional to the in-degree of $v$.

### 3 Definitions, Notations and Known Results

Let $G = (V, E)$ be a directed graph representing a set of Web-pages and their interconnecting links. In what follows we define the terms of rank-similarity, rank-stability and $L_1$-stability of link-based ranking algorithms for the Web, and the concept of authority-connected graphs [5]. Our definitions, although at times rephrased, are equivalent to those given in [5].

---

*The (normalized) eigenvector which corresponds to the eigenvalue of highest magnitude.*
3.1 $L_1$-Stability, Rank-Stability and Rank-Similarity

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 $G$. The weights of $A_i(G)(i = 1, 2)$ induce rankings on the nodes of $G$.

**Definition 1** Let $G_1, G_2$ be two graphs with $N$ nodes, and let $A_1, A_2$ be two ranking algorithms. The ranking distance, $d_r$, between $A_1(G_1)$ and $A_2(G_2)$ is defined as follows:

$$d_r(A_1(G_1), A_2(G_2)) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} I_{A_1(G_1), A_2(G_2)}(i, j)$$

where $I_{A_1(G_1), A_2(G_2)}(i, j) = \begin{cases} 1 & A_1(G_1)_i < A_1(G_1)_j AND A_2(G_2)_i > A_2(G_2)_j \\ 0 & otherwise \end{cases}$

The ranking distance $d_r$ is a normalized version of the Kendall tau distance between two rankings on the same set of objects $[10, 11]$.

In what follows, $\mathcal{G}$ denotes a set of directed graphs, and $\mathcal{G}_N$ is the subset of $N$-node graphs in $\mathcal{G}$.

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

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

Let $G_1 = (V, E_1), G_2 = (V, E_2)$ be two graphs on the same set of nodes. The edge distance, $d_e$, between $G_1$ and $G_2$ is defined as $d_e(G_1, G_2) = |(E_1 \cup E_2) \setminus (E_1 \cap E_2)|$.

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

$$\lim_{N \to \infty} \max_{\{G_i, G_2 \in \mathcal{G}_N \mid d_e(G_i, G_2) \leq k\}} d_r(A(G_1), A(G_2)) \to 0$$

When proving that two algorithms are not rank-similar [an algorithm is not rank-stable], we will construct infinitely many examples where the two rankings being compared differ significantly (the relative order of a significant fraction of the $\binom{N}{2}$ pairs of nodes is reversed).

\footnote{In this and the following definition it is assumed that the max operation is performed on nonempty sets}
While most of our results concern the stability and similarity of the rankings that are produced by the various algorithms, we also extend previous results concerning the stability of the weight vectors that induce the rankings.

**Definition 4** An algorithm $A$ is $L_1$-stable on $G$ if for every fixed $k$, we have

$$
\lim_{N \to \infty} \max_{G_1, G_2 \in \mathcal{G}_N} \min_{\gamma \geq 0} \sum_{i=1}^{N} |A(G_1)_i - \gamma A(G_2)_i| = 0
$$

where $A(G_1)$ and $A(G_2)$ are scaled to be unit vectors under the $L_1$ norm.

### 3.2 Authority-Connected Graphs

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 5** 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. Since one of the most basic premises of link analysis is that a co-citation of two pages implies a topical connection between them, distinct authority-connected components of a graph conceptually correspond to neighborhoods of pages that pertain to different topics or concepts. 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?”). We will show that different algorithms may produce disagreeing rankings even on authority connected graphs.

Similarly, examining the rank stability of an algorithm on general graphs allows the modified links to change the partitioning of the pages into authority connected components.
When examining rank stability on authority connected graphs, we ensure that the relevance of all pages will be measured with respect to the same bar.

3.3 Known Results

Rank-Stability and Rank-Similarity Let \( \mathcal{G} \) denote the set of all directed graphs. The following were shown in [5]: (1) HITS is not rank-stable on \( \mathcal{G} \), (2) HITS is not \( L_1 \)-stable on \( \mathcal{G} \), (3) HITS and SALSA are not rank-similar on \( \mathcal{G} \), and (4) SALSA is not rank-stable on \( \mathcal{G} \) 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 [5] leave open the question whether the negative results (1), (2) and (3) remain true when the discussion is limited to authority connected graphs. We answer this affirmatively in Section 4.

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 [16], both theoretically and experimentally, that HITS is more sensitive than SALSA to this effect. Specifically, authority-connected graphs containing both a small tightly-knit community (with many interconnecting hubs) and a large, less densely connected
community, were constructed. 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 $O\left(\frac{1}{\sqrt{N}}\right)$, which approaches zero as $N$ (the number of pages with at least one incoming link) grows.

The theoretical analysis was exemplified in [16] by several real-life Web graphs, on which HITS and SALSA gave different results, with the rankings of HITS biased towards the pages of tightly-knit communities.

4 Results

Our work focuses on the class of authority connected graphs, which we denote by $G^{AC}$. We show that HITS is not $L_1$-stable on $G^{AC}$, and that neither HITS nor PageRank is rank stable on $G^{AC}$. Thus, SALSA is the only algorithm of the three algorithms we consider here that is rank stable on $G^{AC}$. Furthermore, we show that no pair of these algorithms is rank similar on $G^{AC}$.

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

**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.

![Diagram of graph $G_2$]

**Figure 1:** The graph $G_2$

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 $x_1 < x_2 < \ldots < x_n$. 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 reversed and are in complete disagreement. Formally,

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

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


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

\[
\begin{pmatrix}
2 & 1 & 0 & 0 & 0 & \cdots & 0 \\
1 & 2 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 2 & 1 & 0 & \cdots & 0 \\
: & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 1 & 2 & 1 & 0 \\
0 & \cdots & 0 & 0 & 1 & 2 & 1 \\
0 & \cdots & 0 & 0 & 0 & 1 & 4 \\
\end{pmatrix}
\]

Denote the principal eigenvalue of the matrix by $\lambda$. By the Perron-Frobenius theorem [12], $\lambda$, as well as $x_1, \ldots, x_n$, are positive. This easily implies that $\lambda$ is greater than any element on the main diagonal of the matrix, hence $\lambda > 4$.

The authority weights $x_1, x_2, \ldots, x_n$ satisfy the following equations:

\[
2x_1 + x_2 = \lambda x_1 \implies x_2 = (\lambda - 2)x_1
\]
\[
x_i + 2x_{i+1} + x_{i+2} = \lambda x_{i+1} \implies x_{i+2} = (\lambda - 2)x_{i+1} - x_i, \quad i = 1, \ldots, n - 2
\]

Define $y_i \triangleq \frac{x_{i+1}}{x_i}$. We have $y_0 = 1, y_n = \lambda - 2$, and

\[
y_{i+2} = (\lambda - 2)y_{i+1} - y_i, \quad i = 0, \ldots, n - 3
\]

Solving the recursion yields:

\[
y_i = \frac{\sqrt{z^2 - 1} + z}{2\sqrt{z^2 - 1}} \left[ z + \sqrt{z^2 - 1} \right]^i + \frac{\sqrt{z^2 - 1} - z}{2\sqrt{z^2 - 1}} \left[ z - \sqrt{z^2 - 1} \right]^i, \quad i = 0, \ldots, n - 1
\]

where

\[
z \triangleq \frac{\lambda - 2}{2} > 1
\]

It is easy to see that $y_0 < y_1 < \ldots < y_{n-1}$, and therefore $x_1 < x_2 < \ldots < x_n$. \qed

**Proposition 2** \textit{HITS is not} $L_1$-stable on the class of authority connected graphs $G^{AC}$. 

\[
\text{Proposition 2}
\]
Proof: The proof is similar in spirit to the proof of Proposition 1. We add to the graph $G_2$ depicted in Figure 1 a new node $h^*$, and the edge $h^* \rightarrow a_n$. Let $\tilde{G}_2$ be the resulting graph, shown in Figure 2.

Let $M_{\tilde{G}_2}$ be the co-citation matrix of $\tilde{G}_2$, and let $\lambda$ denote the principal eigenvalue of $M_{\tilde{G}_2}$. The $n$’th row of $M_{\tilde{G}_2}$ is $(0, 0, \ldots, 0, 1, 5)$, and therefore $\lambda > 5$. Let $(x_1, x_2, \ldots, x_n)$ denote the (positive) principal eigenvector of $M_{\tilde{G}_2}$, normalized under the $L_1$ norm ($\sum_{i=1}^{n} x_i = 1$). We first prove by induction that for all $j = 1, \ldots, n - 1$, $x_{j+1} > 2.5x_j$.

As in Proposition 1, the authority weights $x_1, x_2, \ldots, x_n$ satisfy the following equations:

$$2x_1 + x_2 = \lambda x_1 \implies x_2 = (\lambda - 2)x_1$$

$$x_i + 2x_{i+1} + x_{i+2} = \lambda x_{i+1} \implies x_{i+2} = (\lambda - 2)x_{i+1} - x_i, \quad i = 1, \ldots, n - 2$$

The first equation yields $x_2 > 3x_1 > 2.5x_1$. We use the second equation to complete the induction by observing that

$$x_{j+1} > 3x_j - x_{j-1} > 3x_j - \frac{x_j}{2} = 2.5x_j$$

for $j = 2, \ldots, n - 1$. Consequently,

$$x_j > \sum_{i=1}^{j-1} x_i \quad \text{for } j = 2, \ldots, n \quad \text{and in particular } x_n > \frac{1}{2} \sum_{i=1}^{n} x_i.$$
Let \( y_1, \ldots, y_n \) the \( L_1 \)-normalized authority weights that HITS assigns to \( a_1, \ldots, a_n \) in \( \tilde{G}_1 \), the graph derived from \( \tilde{G}_2 \) by connecting \( h^*, h'^* \) and \( h^{**} \) to \( a_1 \) instead of to \( a_n \). Note that \( d_e(\tilde{G}_1, \tilde{G}_2) = 6 \) and that by symmetry, \( x_i = y_{n+1-i} \). For ease of notation, we complete the proof for even values of \( n \geq 4 \), denoting \( m \triangleq \frac{n}{2} \). The proof for odd values of \( n \) is similar.

Assume first that the scaling factor \( \gamma \) (see Definition 4) is at most 1. Then,

\[
\sum_{i=1}^{n} |x_i - \gamma y_i| \geq \sum_{i=m+1}^{n} |x_i - \gamma y_i|
\]

\[
= \sum_{i=m+1}^{n} (x_i - \gamma x_{n+1-i}) = (\sum_{i=m+1}^{n} x_i) - \gamma(\sum_{i=1}^{m} x_i)
\]

\[
\geq (\sum_{i=m+1}^{n} x_i) - (\sum_{i=1}^{m} x_i) = (\sum_{i=m+1}^{n} x_i) - (\sum_{i=1}^{m} x_i)
\]

The case where \( \gamma > 1 \) is similar, since

\[
\sum_{i=1}^{n} |x_i - \gamma y_i| \geq \sum_{i=1}^{m} |x_i - \gamma y_i| = \sum_{i=m+1}^{n} |\gamma x_i - x_{n+1-i}| > (\sum_{i=m+1}^{n} x_i) - (\sum_{i=1}^{m} x_i)
\]

and the proof continues as before. \( \square \)

**Proposition 3** 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 3):

\[
V = \{c, x_a, y, x_b, h_a, h_b, a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n\}
\]

\[
E = \{x_a \rightarrow h_a, x_b \rightarrow h_b\} \cup \{h_a \rightarrow a_i, h_b \rightarrow b_i \mid i = 1, \ldots, n\} \cup \{c \rightarrow v, v \rightarrow c \mid v \in V \setminus \{c\}\}
\]

Define \( G_a \triangleq (V, E \cup \{y \rightarrow h_a\}) \), \( G_b \triangleq (V, E \cup \{y \rightarrow h_b\}) \). 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,

\[
d_r(Pagerank(G_a), Pagerank(G_b)) = \frac{n^2}{(2n + 6)^2}, \quad d_r(G_a, G_b) = 2
\]

and the result follows. \( \square \)

**Proposition 4** 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 4). For ease of notation, we denote \( A = \{a_1, \ldots, a_n\} \) and \( B = \{b_1, \ldots, b_n\} \).

\[
\begin{align*}
V & = \{s, h_s, h_1, h_2, \ldots, h_{2n}\} \cup A \cup B \\
E & = \{s \rightarrow v \mid v \in V \setminus \{s\}\} \cup \{h_a \rightarrow a \mid a \in A\} \cup \{h_s \rightarrow s\} \cup \\
& \quad \{h_i \rightarrow b_i, h_{i+n} \rightarrow b_i \mid i = 1, \ldots, n\} \cup \{a_i \rightarrow s, b_i \rightarrow s, h_i \rightarrow s \mid i = 1, \ldots, n\}
\end{align*}
\]

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 $\lambda$ 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:

$$\begin{align*}
\lambda a &= 2na + nb + T \\
\lambda b &= na + (n + 2)b + T
\end{align*}$$

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 $n > 2$)

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

Thus, HITS ranks the $A$-nodes higher than it ranks the $B$-nodes.
As for PageRank, by arguments similar to those of Proposition 3, 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,

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

proving that the two algorithms are not rank similar. \( \Box \)

**Proposition 5** HITS and SALSA are not rank similar on the class of authority connected graphs.

*Proof:* Consider the graph \( G_3 \) defined in Proposition 4. 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. \( \Box \)

**Proposition 6** PageRank and SALSA are not rank similar on the class of authority connected graphs \( G^{AC} \).

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

\[ V = \{s, x_1, x_2, \ldots, x_t, y, h_2, h_5^1, h_5^2 \} \cup A \cup B \]

\[ E = \{x_i \rightarrow h_a \mid i = 1, \ldots, t\} \cup \{y \rightarrow h_5^1, y \rightarrow h_5^2\} \cup \{h_a \rightarrow a \mid a \in A\} \cup \{h_5^1 \rightarrow b, h_5^2 \rightarrow b \mid b \in B\} \cup \{s \rightarrow v \mid v \in V \setminus \{s, h_5^1\}\} \cup \{v \rightarrow s \mid v \in V \setminus \{s\}\} \]

\( G_5 \) contains \( N = 2n + t + 5 \leq 3n + 5 \) nodes (see Figure 5), 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,

$$PR(a_1) = PR(a_2) = \ldots = PR(a_n) = \frac{d}{N} + (1-d) \left[ \frac{PR(h_a)}{n+1} + \frac{PR(s)}{N-2} \right]$$  \hspace{1cm} (3)

$$PR(b_1) = PR(b_2) = \ldots = PR(b_n) = \frac{d}{N} + (1-d) \left[ \frac{PR(h_b)}{n+1} + \frac{PR(s)}{N-2} \right]$$  \hspace{1cm} (4)

Subtracting (4) from (3), we have

$$PR(a_1) - PR(b_1) = \ldots = PR(a_n) - PR(b_n) = \frac{1}{n+1} \left[ PR(h_a) - PR(h_b) - PR(h_b^2) \right]$$

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 $p \triangleq PR(y)$. Since $PR(y) = PR(x_1) = PR(x_2) = \ldots = PR(x_t)$, we have

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

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

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

$$PR(h_a) - PR(h_b^1) - PR(h_b^2) = p(1-d) \left( \frac{t}{2} - \frac{2}{3} \right) - \frac{d}{N} > \frac{d}{N} \left[ (1-d) \left( \frac{t}{2} - \frac{2}{3} \right) - 1 \right]$$
By our choice of $t$,

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

which proves that $PR(h_a) - PR(h_{1b}) - PR(h_{2b}) > 0$. We conclude that

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

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 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 [16]. 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 (eg. [14, 19, 3, 1, 15, 20]).

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 [14, 7]. A study by Pandurangan et al. [19] 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 \( d \) 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\(^6\) could also be small. We leave this for future experimental research.

References


---

\(^6\)SALSA is not equivalent to the in-degree measure on graphs that are not authority-connected


