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

Preliminary Version

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

ABSTRACT

The stability of Web link-based ranking algorithms was examined in recent works. Among the aspects investigated were the notions of rank stable algorithms and rank similar algorithms. Of special interest are stability results on a particular class of graphs, called authority connected graphs.

This report considers three link-based ranking algorithms: PageRank, HITS and SALSA. We extend previous results 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.

1 Introduction

Two recent works have defined and analyzed robustness and stability aspects of link-based ranking algorithms. Ng et al. ([5]) examined how small changes in the examined graph affect the weight vectors of PageRank, HITS, and Subspace HITS (a variation of HITS they presented). They did not analyze, however, the effect of the changed weight vectors on the rankings which the weights induce. The work of Borodin et al. ([1]), in addition to attacking the stability issue from a different angle, also introduced the notion of rank stability: a measure of how do small changes to the analyzed graph affect the induced rankings. Furthermore, they introduced the notions of similar
and rank similar algorithms, which measure the resemblance between the weights/rankings produced by pairs of algorithms. They applied the definitions to many algorithms (and pairs of algorithms), and attained many (mostly negative) results.

Both works noted that the stability of an algorithm may depend on whether the graph being analyzed is connected or not. To this effect, the class of authority connected graphs was defined in [1]. The authors of [1] leave open the question whether their negative results remain true when the discussion is limited to authority connected graphs.

We extend the results of [1] 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.

This report is organized as follows. Section 2 brings the definitions of rank-similarity and rank-stability from [1]. Section 3 briefly describes the ranking algorithms PageRank, HITS and SALSA. Section 4 cites the relevant results from [1], and details our extension of those results.

2 Definitions and Notations

Let $G = (V,E)$ be a directed graph representing a set of Web-pages and their interconnecting links. Let $W_G$ denote the $|V| \times |V|$ adjacency matrix of $G$. $W_G^TW_G$ is the co-citation matrix of $G$. Let $M_G$ be the matrix that results from deleting any zero rows and columns of $W_G^TW_G$ (these rows and columns correspond to nodes of $G$ whose in-degree equals zero).

The following terms were defined in [1]. The definitions that follow are at times rephrased, but are equivalent to the original definitions.

**Definition 1** A graph $G$ is called authority connected if the matrix $M_G$ is irreducible.

In other words, $G$ is authority connected if the support of $M_G$ is the adjacency matrix of a connected, undirected graph. Let $G_N^C$ denote the class of authority connected graphs on $N$ nodes.

Let $A_1$ and $A_2$ be two link-based ranking algorithms for the Web. Let $A(G)$ denote the $|V|$-dimensional weight vector that algorithm $A$ assigns to the nodes of the graph $G$. $A(G)$ also induces a ranking on the nodes of $G$.

**Definition 2** 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:

\[
d_r(A_1(G_1), A_2(G_2)) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{I}_{A_1(G_1), A_2(G_2)}(i, j)
\]

where

\[
\mathbf{I}_{A_1(G_1), A_2(G_2)}(i, j) = \begin{cases} 
1 & A_1(G_1)_i < A_1(G_1)_j \text{ AND } A_2(G_2)_i > A_2(G_2)_j \\
0 & \text{otherwise}
\end{cases}
\]

Let \( \mathcal{G}_N \) denote a subset of all directed graphs with \( N \) nodes.

**Definition 3** Two ranking algorithms \( A_1 \) and \( A_2 \) are rank-similar on \( \mathcal{G}_N \) if (as \( N \to \infty \))

\[
\max_{G \in \mathcal{G}_N} d_r(A_1(G), A_2(G)) \to 0
\]

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 4** An algorithm \( A \) is rank stable on \( \mathcal{G}_N \) if for every fixed \( k \), we have (as \( N \to \infty \))

\[
\max_{G_1, G_2 \in \mathcal{G}_N \text{ s.t. } d_e(G_1, G_2) \leq k} d_r(A(G_1), A(G_2)) \to 0
\]

### 3 Link-Based Ranking Algorithms for Web Pages

This section provides a brief overview of three link based ranking algorithms: PageRank [2], HITS [3] and SALSA [4]. We describe how each of the algorithms ranks the pages (=nodes) of a graph \( G = (V, E) \) where \( |V| = N \).

#### 3.1 PageRank

PageRank ([2]) 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 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 $1 - d$. The PageRanks obey the following formula:

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

### 3.2 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 $a(s)$, using the following iterative algorithm:

1. Initialize $a(s) \leftarrow 1, h(s) \leftarrow 1$ for all pages $s \in V$.

2. Repeat the following three operations until convergence:
   - Update the authority weight of each page $s$ (the $I$ operation):
     $$a(s) \leftarrow \sum_{x | (s, x) \in E} h(x)$$
   - Update the hub weight of each page $s$ (the $O$ operation):
     $$h(s) \leftarrow \sum_{x | (x, s) \in E} a(x)$$
   - Normalize the authority weights and the hub weights.

Pages are ranked according to their authority weights. Kleinberg showed that this algorithm converges, and that the resulting authority weights are the coordinates of the normalized principal eigenvector \(^1\) of the co-citation matrix $W_G^T W_G$.

### 3.3 SALSA

SALSA, the Stochastic Approach for Link Structure Analysis, is 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).

\(^1\)The eigenvector of the matrix which corresponds to the eigenvalue of highest magnitude.
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 [4] that on authority connected graphs, \( \pi_v \) is directly proportional to the in-degree of \( v \).

4 Results

We begin this section by surveying several results from [1]. Let \( G_N \) denote the set of directed graphs with \( N \) nodes. Recall that \( G_N^{AC} \) denotes the class of authority connected graphs with \( N \) nodes, a subset of \( G_N \). The following were shown in [1]:

1. Neither HITS nor SALSA is rank stable on \( G_N \).
2. HITS and SALSA are not rank similar on \( G_N \).
3. SALSA is rank stable on \( G_N^{AC} \).

The first two items were proven by constructing graphs that are not authority connected. Those graphs, which contain several components, conceptually correspond to collections of Web pages that are relevant to multiple topics.

Our work focuses on authority connected graphs. We show that both HITS and PageRank are not rank stable on \( G_N^{AC} \). Thus, by 3 above, SALSA is the only algorithm of the three algorithms we consider here that is rank stable on \( G_N^{AC} \). Furthermore, we show that no pair of these algorithms is rank similar on \( G_N^{AC} \).

**Proposition 1** HITS is not rank stable on the class of authority connected graphs \( G_N^{AC} \).

**Proof:** Consider the following connected graphs \( G_1 \) and \( G_2 \), each consisting of \( N = 2n + 3 \) 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:

- \( h_0 \to a_1, h_n \to a_n \).
- For all \( i = 1, \ldots, n - 1 \): \( h_i \to a_i, h_i \to a_{i+1} \). This set of links causes both graphs to be 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 $w_1, w_2, \ldots, w_n$ and $x_1, \ldots, x_n$ denote the HITS authority weights of $a_1, \ldots, a_n$ under $G_1$ and $G_2$ respectively. In what follows we prove that $x_1 < x_2 < \ldots < x_n$. By the isomorphism of $G_1$ and $G_2$, we have that $w_1 > w_2 > \ldots > w_n$. Thus,

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

The rankings of $a_1, \ldots, a_n$ on $G_1$ and $G_2$ are completely reversed and are in complete disagreement, while the graphs differ in just 4 links. This proves the non-rank-stability of HITS on $G^AC_N$.

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

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

Denote the principal eigenvalue of the matrix by $\lambda$. Since the principal eigenvalue of an irreducible, non-negative matrix is greater than any element on the main diagonal of the matrix, $\lambda > 4$.

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

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

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

The authority weights are all positive. In particular, $x_1 > 0$.

Define $y_i = \frac{x_{i+1}}{x_i}$. We have $y_0 = 1$, $y_1 = \lambda - 2$, and

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

Solving the recursion yields:

$$y_i = \frac{\sqrt{\lambda^2 - 1} + \frac{z}{\sqrt{\lambda^2 - 1}}} {2\sqrt{\lambda^2 - 1}} [z + \sqrt{\lambda^2 - 1}]^i + \frac{\sqrt{\lambda^2 - 1} - \frac{z}{\sqrt{\lambda^2 - 1}}} {2\sqrt{\lambda^2 - 1}} [z - \sqrt{\lambda^2 - 1}]^i, \quad i = 0, \ldots, n-1$$
where
\[ z \triangleq \frac{\lambda - 2}{2} > 1. \]

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

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

**Proof:** Consider the following graph \( G = (V, E) \):

\[
V = \{ c, x_a, y, x_b, \ h_a, h_b, \ a_1, a_2, \ldots, a_n, \ b_1, b_2, \ldots, b_n \}
\]
\[
E = \{ x_a \rightarrow h_a, x_b \rightarrow h_b \} \cup \\
\{ h_a \rightarrow a_i, h_b \rightarrow b_i \mid i = 1, \ldots, n \} \cup \\
\{ c \rightarrow v, v \rightarrow c \mid v \in V \setminus \{ c \} \}
\]

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

\[
PR_a(x_a) = PR_a(y) = PR_a(x_b) \implies PR_a(h_a) > PR_a(h_b)
\]

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,

\[
d_e(\text{PageRank}(G_a), \text{PageRank}(G_b)) = \frac{n^2}{(2n + 6)^2}, \ d_e(G_a, G_b) = 2
\]

and the result follows. \( \square \)

**Proposition 3** HITS and PageRank are not rank similar on the class of authority connected graphs \( G^{AC}_N \).

**Proof:** Consider the following graph \( G = (V, E) \). For ease of notation, we denote \( A = \{a_1, \ldots, a_n\} \) and \( B = \{b_1, \ldots, b_n\} \).

\[
V = \{ s, \ h_a, \ h_1, h_2, \ldots, h_{2n} \} \cup A \cup B
\]
\[
E = \{ s \rightarrow v \mid v \in V \setminus \{ s \} \} \cup \\
\{ h_a \rightarrow a \mid a \in A \} \cup \\
\{ h_i \rightarrow b_i, h_{i+n} \rightarrow b_i \mid i = 1, \ldots, n \} \cup \\
\{ a_i \rightarrow s, b_i \rightarrow s, h_i \rightarrow s \mid i = 1, \ldots, n \} \cup \{ h_a \rightarrow s \}
\]
Thus, $G$ consists of $N \triangleq 4n + 2$ nodes. It is easy to see that $G$ 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 $G$, $M_G = [m_{i,j}]$:

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

The authority scores 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 $M_G$. The rows in $M_G$ that correspond to the $A$ and $B$ nodes give rise to the following two equations:

$$\lambda a = 2n \cdot a + n \cdot b + T$$
$$\lambda b = n \cdot a + (n + 2) \cdot b + T$$

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 $n > 2$)

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

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

As for PageRank, by arguments similar to those of Proposition 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,

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

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 $\mathcal{G}^{AC}_N$. 

8
Proof: Consider the graph $G$ defined in Proposition 3. Since $G$ 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 5** PageRank and SALSA are not rank similar on the class of authority connected graphs $G^N_AC$.

Proof: Let $d$ be the random jump parameter of PageRank and let

$$t = \left[ \frac{4}{3} + \frac{2}{1-d} \right]$$

For all $n \geq t$, consider the following graph $G = (V, E)$ ($A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_n\}$):

$$V = \{s, x_1, x_2, \ldots, x_t, y, h_a, h_b, h_a^1, h_b^1\} \cup A \cup B$$

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

$G$ contains $N \Delta 2n+t+5 \leq 3n+5$ nodes, 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 $v$. For all $1 \leq i < j \leq n$, $PR(a_i) = PR(a_j)$ and $PR(b_i) = PR(b_j)$. Now,

$$PR(a_1) = \frac{d}{N} + (1-d) \left[ \frac{PR(h_a)}{n+1} + \frac{PR(s)}{N-2} \right]$$

$$PR(b_1) = \frac{d}{N} + (1-d) \left[ \frac{PR(h_b^1) + PR(h_b)}{n+1} + \frac{PR(s)}{N-2} \right]$$

Subtracting (4) from (3), we have

$$PR(a_1) - PR(b_1) = \frac{1-d}{n+1} [PR(h_a) - PR(h_b^1) - PR(h_b^2)]$$

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.
Note that \( PR(x_1) = PR(x_2) = \ldots = PR(x_t) = PR(y) \), and let \( p \triangleq PR(x_1) \). Like every PageRank in \( G \), \( p > \frac{d}{N} \).

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

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

Therefore,

\[
PR(h_a) - PR(h_b^1) - PR(h_b^2) = p(1 - d)(\frac{t}{2} - \frac{2}{3}) - \frac{d}{N}
\]

\[
> \frac{d}{N} \left[ (1 - d)(\frac{t}{2} - \frac{2}{3}) - 1 \right]
\]

By our choice of \( t \),

\[
t \geq \frac{4}{3} + \frac{2}{1 - d} \implies \frac{t}{2} \geq \frac{2}{3} + \frac{1}{1 - d}
\]

or

\[
(1 - d)(\frac{t}{2} - \frac{2}{3}) \geq 1
\]

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

We conclude that

\[
d_r(SALSA(G), \text{PageRank}(G)) \geq \frac{n^2}{(3n + 5)^2},
\]

and the result follows. \( \square \)

References


