SIAM REVIEW
Vol. 46, No. 2, pp. 256–268
© 2004 Society for Industrial and Applied Mathematics

# Link Analysis: Hubs and Authorities on the World Wide Web*

Chris H. Q. Ding†
Hongyuan Zha‡
Xiaofeng He†
Parry Husbands†
Horst D. Simon†

**Abstract.** Ranking the tens of thousands of retrieved webpages for a user query on a Web search engine such that the most informative webpages are on the top is a key information retrieval technology. A popular ranking algorithm is the HITS algorithm of Kleinberg. It explores the reinforcing interplay between authority and hub webpages on a particular topic by taking into account the structure of the Web graphs formed by the hyperlinks between the webpages. In this paper, we give a detailed analysis of the HITS algorithm through a unique combination of probabilistic analysis and matrix algebra. In particular, we show that to first-order approximation, the ranking given by the HITS algorithm is the same as the ranking by counting inbound and outbound hyperlinks. Using Web graphs of different sizes, we also provide experimental results to illustrate the analysis.

**Key words.** webpage ranking, random graph, co-citation, co-reference, HITS, PageRank

**AMS subject classifications.** 05C80, 15A18, 05C50, 94C15

**DOI.** 10.1137/S0036144501389218

## 1. Introduction.

The rapidly growing World Wide Web now contains more than three billion webpages of text, images, and other multimedia information. While this vast amount of information has the potential to benefit all aspects of our society, finding the relevant webpages to satisfy a user's need for information still remains an important and challenging task. Many commercial search engines have been developed and used by people all over the world. However, the relevancy of webpages returned by search engines is still lacking, and further research and development are needed to make search engines more effective as a ubiquitous information-seeking tool.

*Received by the editors May 9, 2001; accepted for publication (in revised form) July 18, 2003; published electronically May 3, 2004. This work was supported in part by the Office of Science, Office of Laboratory Policy and Infrastructure, of the U.S. Department of Energy under contract DE-AC03-76SF00098 through an LBL LDRD grant. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sirev/46-2/38921.html

†NERSC Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720 (chqding@lbl.gov, xhe@lbl.gov, pjrhusbands@lbl.gov, hdsimon@lbl.gov).

‡Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802 (zha@cse.psu.edu). The work of this author was supported in part by NSF grant CCR-9901986.

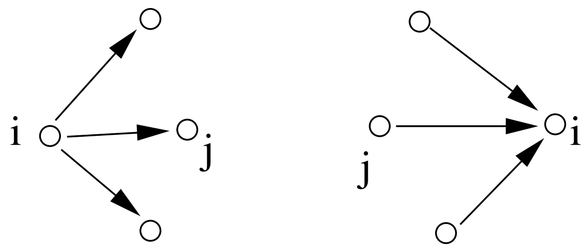


**Fig. 1** *Left: Hub webpage $p_i$ has many outbound hyperlinks. Right: Authority webpage $p_i$ has many inbound hyperlinks.*

A distinct feature of the Web is the proliferation of hyperlinks between webpages which allow a user to surf from one webpage to another with a simple click. We can model the Web as a *directed* graph with the webpages as the nodes and the hyperlinks as the directed edges. This hyperlink graph contains useful information: if webpage $p_i$ has a link pointing to webpage $p_j$, it usually indicates that the creator of $p_i$ considers $p_j$ to contain relevant information for $p_i$. Such useful opinions and knowledge are therefore registered in the form of hyperlinks. Exploring the information stored in the link graphs to infer certain relationships is an emerging field of *link analysis*. Recent introductory surveys of Web link analysis can be found in [18, 21].

A valuable and informative webpage is usually pointed to by a large number of hyperlinks; i.e., it has a large indegree (see Figure 1). Such a webpage is called an *authority* [22]. A webpage that points to many authority webpages is itself a useful resource and is called a *hub*. A hub usually has a large outdegree. In the context of literature citation, a hub is a review paper that cites many original papers, while an authority is an original seminal paper cited by many papers.

The *Hypertext Induced Topic Selection* (HITS) algorithm of Kleinberg [22] improves on the basic notions of hubs and authorities. HITS assigns importance scores to hubs and authorities, and computes them in a mutually reinforcing way: a good authority must be pointed to by several good hubs, while a good hub must point to several good authorities. Further improvements and extensions of HITS were developed in [15, 6, 10, 24, 7, 11, 26, 1, 3]. The goal of this paper is to give a detailed analysis of the HITS algorithm, focusing on the role of indegrees and outdegrees.

**2. The HITS Algorithm.** The HITS algorithm is applied to a set of webpages generated from the search engine for a query. Specifically, a subset of the top-ranked webpages together with their one-hop-away neighbors are used for analysis [22]. In the HITS algorithm, each webpage $p_i$ in the set is assigned a hub score $y_i$ and an authority score $x_i$. The intuition is that a good *authority* is pointed to by many good *hubs* and a good *hub* points to many good authorities. This mutually reinforcing relationship is represented as

$$x'_i = \sum_{j:e_{ji}\in E} y_j, \quad y'_i = \sum_{j:e_{ij}\in E} x_j; \quad x_i = x'_i/||\mathbf{x}'||, \quad y_i = y'_i/||\mathbf{y}'||, \tag{1}$$

where $||\cdot||$ stands for $L_2$ norm. Final hub and authority scores are obtained by iteratively solving (1). By ordering webpages in decreasing order according to their scores, one obtains the rankings of hubs and authorities.

The set of webpages forms a directed graph $G = (V, E)$, where webpage $p_i$ is a node in $V$ and hyperlink $e_{ij}$ is an edge in $E$. The adjacency matrix $L$ of the graph is defined as $L_{ij} = 1$ if $e_{ij} \in E$, and 0 otherwise. Authority scores on all $n$ nodes form a vector $\mathbf{x} = (x_1, \ldots, x_n)^T$, and hub scores form a vector $\mathbf{y} = (y_1, \ldots, y_n)^T$. Equation (1) can be cast into

$$\mathbf{x}' = L^T\mathbf{y}, \quad \mathbf{y}' = L\mathbf{x}; \quad \mathbf{x} = \mathbf{x}'/||\mathbf{x}||, \quad \mathbf{y} = \mathbf{y}'/||\mathbf{y}||.$$

Let $\mathbf{x}^{(t)}, \mathbf{y}^{(t)}$ denote hub and authority scores at the $t$th iteration. The iteration processes to reach the final solutions are

$$c\mathbf{x}^{(t+1)} = L^T L\mathbf{x}^{(t)}, \quad c\mathbf{y}^{(t+1)} = LL^T\mathbf{y}^{(t)}, \tag{2}$$

starting with $\mathbf{x}^{(0)} = \mathbf{y}^{(0)} = \mathbf{e} \equiv (1, \ldots, 1)^T$, where $c$ is a normalization factor so that $||\mathbf{x}|| = ||\mathbf{y}|| = 1$. Since $L^T L$ determines the authority ranking, we call $L^T L$ the authority matrix. Similarly, we call $LL^T$ the hub matrix. The final solution $\mathbf{x}^*$, $\mathbf{y}^*$ consists of the respective principal eigenvectors of the symmetric positive definite matrices $L^T L$ and $LL^T$: $L^T L\mathbf{x}^* = \lambda\mathbf{x}^*$ and $LL^T\mathbf{y}^* = \lambda\mathbf{y}^*$, which also characterize the singular value decomposition (SVD) [16] of $L$.

**3. Authority and Co-citation, Hub and Co-reference.** The hub and authority matrices have interesting connections [22] to two important concepts, co-citation and co-reference, in the fields of citation analysis and bibliometrics, which are fundamental metrics to characterize the similarity between two documents [27, 20]. Here we discuss this relationship in further detail and emphasize the important role of indegrees and outdegrees.

If two distinct webpages $p_i, p_j$ are co-cited by many other webpages, as in Figure 2, $p_i, p_j$ are likely to be related in some way. Thus co-citation is a measure of similarity. It is defined as the number of webpages that co-cite $p_i, p_j$. The co-citation between $p_i, p_j$ can be calculated as $C_{ij} = \sum_k L_{ki}L_{kj} = (L^T L)_{ij}$. The self-citation $C_{ii}$ is not defined and is usually set to $C_{ii} = 0$. Also, $C_{ij} = C_{ji}$. The indegree of webpage $p_i$ is given by $d_i = \sum_k L_{ki} = \sum_k L_{ki}L_{ki} = (L^T L)_{ii}$, since $L_{ki} = 0$ or 1. Let $D$ be the diagonal matrix of indegrees, $D = \text{diag}(d_1, d_2, \ldots, d_n)$. The link structure of $L^T L$ becomes

$$L^T L = D + C. \tag{3}$$

Thus the authority matrix is the sum of co-citation and indegree. One also sees that

$$\max(0, d_i + d_k - n) \le C_{ik} \le \min(d_i, d_k). \tag{4}$$

Thus $C_{ik} = 0$ if $d_i = 0$ or $d_k = 0$. If $d_i = 0$, the $i$th row of $L^T L$ contains all zeros. From (2), its authority score must be zero.

As shown in Figure 2, the fact that two distinct webpages $p_i, p_j$ co-reference many other webpages indicates that $p_i, p_j$ have a certain commonality. The co-reference (bibliometric coupling) measures the similarity between webpages. Let $R = (R_{ij})$ denote the co-reference, where $R_{ij}$ is defined to be the number of webpages co-referenced by $p_i$, $p_j$ and calculated as (see Figure 2) $R_{ij} = \sum_k L_{ik}L_{jk} = (LL^T)_{ij}$. The self-reference $R_{ii}$ is not defined and is set to $R_{ii} = 0$. The outdegree of node $p_i$ is $o_i = \sum_k L_{ik} = \sum_k L_{ik}L_{ik} = (LL^T)_{ii}$. Let $O = \text{diag}(o_1, o_2, \ldots, o_n)$; we have

$$LL^T = O + R. \tag{5}$$

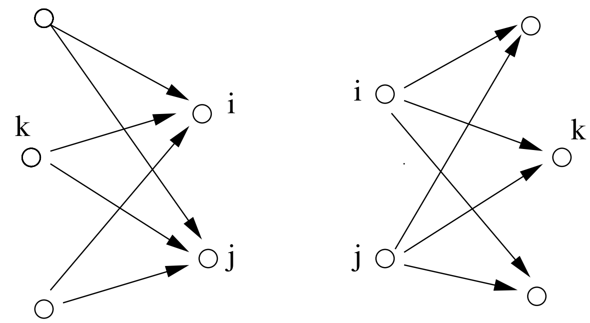


**Fig. 2** *Left: Webpages $p_i, p_j$ are co-cited by webpage $p_k$. Right: Webpages $p_i, p_j$ co-reference webpage $p_k$.*

Thus the hub matrix is the sum of the co-reference and the outdegree. We also have the inequality

$$\max(0, o_i + o_k - n) \leq R_{ik} \leq \min(o_i, o_k). \tag{6}$$

Clearly $R_{ik} = 0$ if $o_i = 0$ or $o_k = 0$. If $o_i = 0$, the $i$th row of $LL^T$ contains all zeros; from (2), its hub score must be zero.

It is interesting to note the duality relationship between hubs and authorities, and the duality between co-citations and co-references. This is similar to the duality between documents and words in information retrieval (IR). The fact that hub and authority scores are embedded in SVD resembles the latent semantic indexing in IR [12, 5].

**4. Probabilistic Analysis.** We analyze the structures of the authority and hub matrices in more detail. Equation (3) suggests an interesting and useful observation on the relationship of co-citations and indegrees: in general, nodes with large indegrees will have large co-citations with other nodes, simply because they have more in-links. Conversely, large co-citations are directly related to the large indegrees of the nodes involved.

These intuitions can be made more precise by assuming the Web graph to be a fixed-degree-sequence random graph and using probabilistic analysis on the expected value of co-citation and co-reference. This is motivated by the result of Aiello, Chung, and Lu [2], where it was proposed that the Web can be better characterized by a fixed-degree-sequence random graph, in which node degrees $\{d_1, \ldots, d_n\}$ are first given and edges are randomly distributed between nodes subject to constraints of node degrees. We have the following proposition.

PROPOSITION 1. *For fixed-degree-sequence random graphs, the expected value of co-citation is given by*

$$\langle C_{ik} \rangle = d_i d_k / (n-1). \tag{7}$$

*This is consistent with* (4).

*Proof.* We prove this relation assuming $d_i \geq d_k$. There are at most $d_k$ nonzero terms in $C_{ij} = \sum_k L_{ki} L_{kj}$, which is the inner product of the $i$th and $k$th columns of adjacency matrix $L$. Consider the case where the $q$th row in the $k$th column is 1. The probability of the corresponding position in the $i$th column being 1 is $P(L_{qi} = 1) =$

$C_{n-2}^{d_i-1}/C_{n-1}^{d_i} = d_i/(n-1)$. Here $C_{n-1}^{d_i}$ is the total number of possible patterns for $d_i$ 1s in the $i$th column, and $C_{n-2}^{d_i-1}$ is the total number of possible patterns given that there is a 1 at row $q$. Thus $\langle C_{ik}\rangle = \sum_q \langle L_{qi}L_{qk}\rangle = \sum_q^{d_k}\langle L_{qi}\rangle = d_k \cdot P(L_{qi}=1)$, and we have (7). ∎

From these analyses, we see that node $i$ with large indegree $d_i$ tends to have large co-citations with other nodes. If $d_i > d_j$, we have $\langle C_{ik}\rangle > \langle C_{jk}\rangle$ for all $k$, $k \neq i$, $k \neq j$. Thus $C_{ik}$ is more likely to be larger than $C_{jk}$, but not necessarily true in every case. We say that $C_{ik} > C_{jk}$ *on average*.

The same analysis can be applied to outdegree and co-reference for hub matrix $LL^T$. We have

$$\langle R_{ik}\rangle = o_i o_k/(n-1). \tag{8}$$

This is consistent with (6).

There are several other models for Web graph topology and indegree and outdegree distributions, such as the webpage copying model [23] and the preferential attachment model [4]. In those more complex models, the degree distributions evolve dynamically; at any given time, however, the Web graph tends to be similar to the fixed-degree-sequence random graph model and (7), (8) hold approximately.

**5. Average Case Analysis.** With the expected value of co-citations given in (7) and the relationship (3) between authorities and co-citations, we can perform an analysis for the average case in which the elements of the authority matrix are replaced by their average values. In this average case, the final ranking scores of the HITS algorithm can be obtained in closed form, providing much insight into the HITS algorithm.

To prove the results of the average case requires the spectral decomposition of a matrix which is the sum of a diagonal matrix and a rank-1 matrix: $A \equiv D + \mathbf{c}\mathbf{c}^T$. The decomposition for this type of matrix is given in Theorem 8.5.3 in Golub and van Loan [16]. That theorem requires that the diagonal entries of $D$ all be distinct. However, in our case, many entries are identical. Thus we generalize Theorem 8.5.3 [16] to this more general case.

THEOREM 2. *Spectral decomposition of the $n$-by-$n$ matrix $A \equiv D + \mathbf{c}\mathbf{c}^T$. Let $D$ be a diagonal matrix of the block form*

$$D = \operatorname{diag}(\tau_1 I_1, \ldots, \tau_\ell I_\ell), \tag{9}$$

*where $I_k, k = 1, \ldots, \ell$, is the identity matrix of size $n_k$, $\tau_k$'s are $\ell$ distinct values*

$$\tau_1 > \tau_2 > \cdots > \tau_\ell, \tag{10}$$

*and the block sizes $n_k$ satisfy $n_1 + \cdots + n_\ell = n$. Let $\mathbf{c}$ be a column vector of the block form $\mathbf{c} = [\mathbf{c}_1^T, \ldots, \mathbf{c}_\ell^T]^T$, with $\mathbf{c}_k$ being a column vector of size $n_k$, and $\mathbf{c}_k \neq 0$. Then the eigenvalues of $A \equiv D + \mathbf{c}\mathbf{c}^T$ are given by*

$$\hat\tau_1 > \underbrace{\tau_1 = \cdots = \tau_1}_{n_1-1} > \hat\tau_2 > \underbrace{\tau_2 = \cdots = \tau_2}_{n_2-1} > \cdots > \hat\tau_\ell > \underbrace{\tau_\ell = \cdots = \tau_\ell}_{n_\ell-1}. \tag{11}$$

*The eigenvector of $A$ corresponding to the eigenvalue $\hat\tau_k$ is given by*

$$\left(\frac{\mathbf{c}_1^T}{\hat\tau_k - \tau_1}, \frac{\mathbf{c}_2^T}{\hat\tau_k - \tau_2}, \ldots, \frac{\mathbf{c}_\ell^T}{\hat\tau_k - \tau_\ell}\right)^T. \tag{12}$$

*The eigenvector corresponding to the eigenvalue* $\tau_k$ *is of the form*

$$(0 \cdots 0, \mathbf{u}_k^T, 0 \cdots 0)^T, \tag{13}$$

*where* $\mathbf{u}_k$ *is an arbitrary vector of size* $n_k$ *satisfying* $\mathbf{c}_k^T \mathbf{u}_k = 0$.

*Proof.* Since $\mathbf{c}_k \neq 0$, we can find exactly $(n_k - 1)$ mutually orthogonal $\mathbf{u}_k$'s satisfying $\mathbf{c}_k^T \mathbf{u}_k = 0$. The corresponding $(n_k - 1)$ vectors of the form in (13) form an orthonormal basis for the invariant subspace of $A$ with eigenvalue $\tau_k$. In total, we have $(n_1 - 1) + \cdots + (n_\ell - 1)$ eigenvectors of this type with corresponding eigenvalues $\tau_1, \ldots, \tau_\ell$ in (11).

Now consider the $\ell$-by-$\ell$ matrix $\hat{A} \equiv \mathrm{diag}(\tau_1, \tau_2, \ldots, \tau_\ell) + \hat{\mathbf{c}}\hat{\mathbf{c}}^T$ with $\hat{\mathbf{c}} = (\|\mathbf{c}_1\|, \ldots, \|\mathbf{c}_\ell\|)^T$. It follows from (10) and Theorem 8.5.3 of [16] that $\hat{A}$ has $\ell$ distinct eigenvalues, $\hat{\tau}_1, \ldots, \hat{\tau}_\ell$, satisfying

$$\hat{\tau}_1 > \tau_1 > \hat{\tau}_2 > \tau_2 > \cdots > \hat{\tau}_\ell > \tau_\ell,$$

and the eigenvector of $\hat{A}$ corresponding to $\hat{\tau}_k$ is given by

$$\left( \frac{\|\mathbf{c}_1\|}{\hat{\tau}_k - \tau_1}, \frac{\|\mathbf{c}_2\|}{\hat{\tau}_k - \tau_2}, \ldots, \frac{\|\mathbf{c}_\ell\|}{\hat{\tau}_k - \tau_\ell} \right)^T. \tag{14}$$

For $k = 1, \ldots, \ell$, let $U_k$ be an orthonormal matrix (coordinate rotation) such that

$$U_k^T \mathbf{c}_k = \|\mathbf{c}_k\| \mathbf{z}_k \equiv \tilde{\mathbf{c}}_k,$$

where $\mathbf{z}_k = (1, 0 \cdots 0)^T$. Define $U = \mathrm{diag}(U_1, \ldots, U_\ell)$ and $\tilde{\mathbf{c}} = [\tilde{\mathbf{c}}_1^T, \ldots, \tilde{\mathbf{c}}_\ell^T]^T$. Then $U^T A U = D + \tilde{\mathbf{c}}\tilde{\mathbf{c}}^T$. By construction, the block structure of $D + \tilde{\mathbf{c}}\tilde{\mathbf{c}}^T$ matches that of $\hat{A}$. Clearly, if (14) is the eigenvector of $\hat{A}$ with the eigenvalue $\hat{\tau}_k$, then

$$\left( \frac{\|\mathbf{c}_1\| \mathbf{z}_1^T}{\hat{\tau}_k - \tau_1}, \frac{\|\mathbf{c}_2\| \mathbf{z}_2^T}{\hat{\tau}_k - \tau_2}, \ldots, \frac{\|\mathbf{c}_\ell\| \mathbf{z}_\ell^T}{\hat{\tau}_k - \tau_\ell} \right)^T \tag{15}$$

is an eigenvector of $D + \tilde{\mathbf{c}}\tilde{\mathbf{c}}^T = U^T A U$ with the same eigenvalue. To get the corresponding eigenvector of $A$, we multiply by $U$ from the left side of (15). Noting that $U_k \mathbf{z}_k = \mathbf{c}_k / \|\mathbf{c}_k\|$, this gives the eigenvector of (12). ∎

Suppose the largest $m$ $(m > 1)$ diagonal entries of $D$ are distinct. Then the $m$ corresponding eigenvectors of $D + \mathbf{c}\mathbf{c}^T$ are of the form in (12). Let $D = \mathrm{diag}(\tau_1, \tau_2, \ldots, \tau_n)$ and $\mathbf{c} = (c_1, c_2, \ldots, c_n)^T$, where the $\tau_i$'s are in nonincreasing order, in contrast with the block form of (9), (10). Let $k \leq m$. The $k$th eigenvector of $D + \mathbf{c}\mathbf{c}^T$ can be written as

$$\left( \frac{c_1}{\hat{\tau}_k - \tau_1}, \frac{c_2}{\hat{\tau}_k - \tau_2}, \ldots, \frac{c_n}{\hat{\tau}_k - \tau_n} \right)^T. \tag{16}$$

This is the case used in the average case analysis of HITS below.

We now turn to the following main result of this paper.

THEOREM 3. *Given a fixed-degree-sequence random graph, assume that* (a) *the largest* $m (m > 1)$ *indegrees are distinct,* $d_1 > \cdots > d_m > d_{m+1} \geq d_{m+2} \geq \cdots \geq d_n$, *and* (b)

$$d_i + d_j < n - 1 \quad \forall\, i, j. \tag{17}$$

*The authority matrix* $L^T L$ *for the average case has the largest* $m$ *eigenvalues* $\lambda_i, i = 1, \ldots, m$, *with the interleave relation*

$$\lambda_1 > h_1 > \lambda_2 > h_2 > \cdots > \lambda_m > h_m, \tag{18}$$

*and the corresponding eigenvectors*

$$\mathbf{u}_k = \left( \frac{d_1}{\lambda_k - h_1}, \frac{d_2}{\lambda_k - h_2}, \ldots, \frac{d_n}{\lambda_k - h_n} \right)^T, \quad k = 1, \ldots, m. \tag{19}$$

*Here* $h_i \equiv d_i - d_i^2/(n-1)$*. Analogous results hold for the hub matrix* $LL^T$*.*

*Proof.* Using (7), we have the average case authority matrix

$$\langle L^T L \rangle = \langle D \rangle + \langle C \rangle = \mathrm{diag}(h_1, h_2, \ldots, h_n) + \mathbf{d}\mathbf{d}^T/(n-1),$$

where $\mathbf{d} = (d_1, d_2, \ldots, d_n)^T$. Now $\langle L^T L \rangle$ is the sum of a diagonal matrix and a rank-1 matrix. To apply Theorem 2, it requires that $h_1 > h_2 > \cdots > h_m > h_{m+1} \geq \cdots \geq h_n$. This is satisfied, because we have

$$h_i - h_j = (d_i - d_j)[1 - (d_i + d_j)/(n-1)].$$

For any $i < j$, the second factor is positive because of (17). Since webpages are indexed according to their indegrees, the first factor is positive for $i \leq m$; otherwise it is nonnegative. Thus the ordering requirement is satisfied. Equations (19), (18) now follow from Theorem 2 directly. ∎

Note that condition (b) of Theorem 3 (cf. (17)) is satisfied if $d_i < (n-1)/2$ for all $i$, which holds for most Web graphs: the indegree of a node is less than half of the total size. Also, indegrees of a Web graph typically follow a power-law distribution [9]: $d_i \propto 1/i^2$. They drop off rapidly. The first few largest indegrees are usually distinct; i.e., condition (a) of Theorem 3 is satisfied.

Given (7), one can also perform a first-order perturbation analysis and obtain eigenvectors very similar to those in (19) (details omitted here).

These principal eigenvectors of $\langle L^T L \rangle$ behave fairly regularly, as illustrated in Figure 3. $\mathbf{u}_1$ is always positive. For $\mathbf{u}_2$, the first node is negative, turning positive from the second node. For $\mathbf{u}_3$, the first two nodes are negative, turning positive from the third node, and so on.

**6. Properties of the HITS Algorithm.** Several interesting results follow directly from Theorem 3:

(1) *Webpage ordering.* The authority ranking is, on average, identical to the ranking according to webpage indegrees. To see this, we have the following corollary.

COROLLARY 4. *Elements of the principal eigenvector* $\mathbf{u}_1$ *are nonincreasing, assuming webpages are indexed such that their indegrees are in nonincreasing order.*

*Proof.* From Theorem 3, we have, for any $i < j$,

$$\mathbf{u}_1(i) - \mathbf{u}_1(j) = \frac{d_i}{\lambda_1 - h_i} - \frac{d_j}{\lambda_1 - h_j} = \frac{(d_i - d_j)[\lambda_1 - d_i d_j/(n-1)]}{(\lambda_1 - h_i)(\lambda_1 - h_j)} \geq 0,$$

because $\lambda_1 - d_i d_j/(n-1) > h_i - d_i d_j/(n-1) = d_i(1 - (d_i + d_j)/(n-1)) > 0$, using (17), and $(\lambda_1 - h_i)(\lambda_1 - h_j)$ is positive. ∎

From this, we conclude that to the extent that the fixed-degree-sequence random graph approximates the web, ranking webpages by their authority scores is the same as ranking by their indegrees. Analogous results hold for hub ranking. These indicate that the duality relationship embedded in mutual reinforcement between hubs and authorities is manifested by their indegrees and outdegrees.

(2) *Uniqueness.* If $d_1$ is larger than $d_2$, then the principal eigenvector of $L^T L$ is unique and is quite different from the second principal eigenvector (see Figure 3).

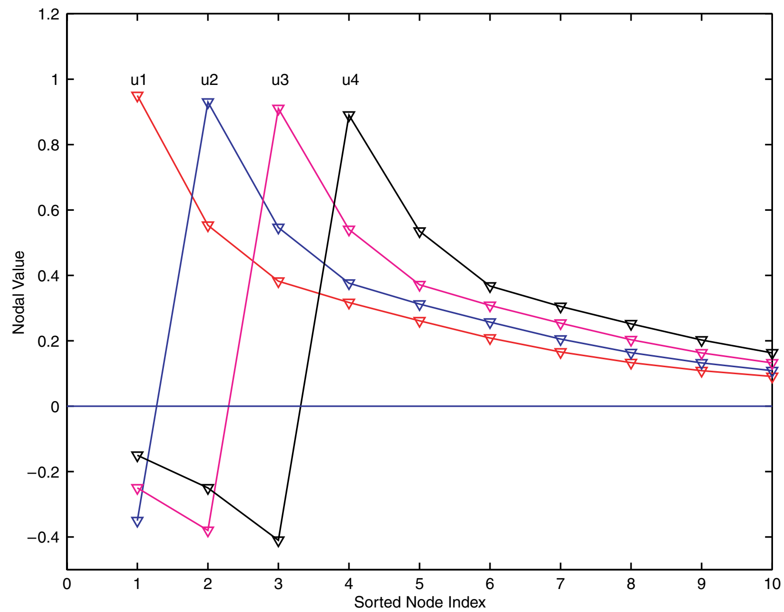


**Fig. 3** *Eigenvectors of* (19).

(3) *Convergence.* The convergence for HITS can be rather fast: (1) The starting vector $\mathbf{x}^{(0)} = (1, \ldots, 1)^T$ has large overlap with principal eigenvector $\mathbf{u}_1$, but little overlap with other principal eigenvectors $\mathbf{u}_k, k = 2, \ldots, m$, because $\mathbf{u}_k$ contains negative nodal values (see Figure 3). (2) In the iterations to compute $\mathbf{u}_1$, the convergence rate depends on $\lambda_2/\lambda_1 \simeq h_1/h_2 \simeq d_1/d_2 \simeq (1/2)^2 = 1/4$, using (18) and the fact that indegrees follow the power-law distribution [9] $d_i \propto 1/i^2$. Thus the iteration converges rapidly. Typically 5–10 iterations are sufficient.

(4) *Web communities.* The HITS algorithm has been used to identify multiple Web communities using different eigenvectors [22, 15]. The principal eigenvector defines a dominant Web community. Each of the other principal eigenvectors $\mathbf{u}_k$ defines two communities, one with nonnegative values $\{i|u_k(i) \geq 0\}$ and the other with negative values $\{i|u_k(i) < 0\}$.

From the pattern of eigenvectors in our solutions (see Figure 3), the positive region of different eigenvectors overlaps substantially. Thus the communities of positive regions nest with each other, as do communities of negative regions. Therefore, we believe that this method to identify multiple communities is less effective. This difficulty is also noticed in practical applications [6]. A number of Web community discovery algorithms are being developed, e.g., trawling to find bipartite cores [23], network maximum flow [14], and graph-clustering [17]. One advantage of these methods is that weak communities (topics) can be separated from dominant communities and thus identified. Without explicit community discovery, webpages of weak topics are typically ranked low by HITS (and by indegree ranking) and are often missed.

## 7. Experimental Results.

**7.1. Internet Archive.** This dataset was supplied by the Internet Archive [19] and was extracted from a crawl performed over 1998–1999. It has 4,906,214 web-

**Table 1** *Authority ranking for Internet Archive.*

| HITS | Indegree | URL |
|---|---|---|
| 1 | 4 | www.yahoo.com |
| 2 | 3 | www.geocities.com |
| 3 | 1 | www.microsoft.com |
| 4 | 6 | members.aol.com |
| 5 | 2 | home.netscape.com |
| 6 | 10 | www.excite.com |
| 7 | 11 | www.lycos.com |
| 8 | 9 | members.tripod.com |
| 9 | 15 | ourworld.compuserve.com |
| 10 | 5 | www.netscape.com |
| 11 | 20 | www.cnn.com |
| 12 | 28 | www.webcom.com |
| 13 | 33 | sunsite.unc.edu |
| 14 | 7 | www.adobe.com |
| 15 | 35 | www.teleport.com |
| 16 | 17 | www.altavista.digital.com |
| 17 | 25 | www.w3.org |
| 18 | 19 | www.infoseek.com |
| 19 | 18 | www.angelfire.com |
| 20 | 21 | www.hotbot.com |
| ... | ... | ...... |
| 111 | 13 | www.linkexchange.com |
| 137 | 14 | ad.linkexchange.com |
| 174 | 17 | member.linkexchange.com |

sites and represents a site-level graph of the Web. The principal eigenvectors were obtained using PARPACK [25] on NERSC's IBM SP computer. Table 1 lists the top 20 authorities, ranked by HITS (first column) and by indegree (second column).

In general, one sees that the HITS ranking and indegree rank are highly correlated, as expected from our analytical results. For these reasons, we consider as *normal* those webpages highly ranked by HITS that also have high indegree. There are two types of webpage that deviate from this general pattern and warrant further examination: (a) those authority webpages highly ranked by HITS, but with relatively smaller indegrees, and (b) those webpages with large indegrees, but ranked low by HITS. These webpages would have been incorrectly ranked if we simply counted indegrees, representing the net improvements brought by the HITS algorithm.

For type (b) webpages, we note that three websites, www.linkexchange.com, ad. linkexchange.com, and member.linkexchange.com, are ranked high by indegree (rank 13, 14, 16, respectively). They are ranked low by HITS (rank 111, 137, 174, respectively). All three sites have very large indegrees, but also very small outdegrees; they are all *sinks*: many sites point to them, but they do not point to anywhere. The mutually reinforcing nature of the HITS algorithm ranked them low, because there are no good hubs pointing to them. These anomalies indicate the effectiveness of the HITS algorithm.

As for type (a) webpages, we mention two websites: (1) sunsite.unc.edu, which is ranked 13 in HITS, but is ranked 33 by indegree. This site holds many software repositories, but few outbound links. Its higher HITS ranking is reasonable because more top sites such as Microsoft point to it. (2) www.teleport.com, which is ranked 15 by HITS, but is ranked 35 by indegree. This site has a large number of outlinks, and more top sites point to it.

**Table 2** *Hub ranking for Internet Archive.*

| HITS | Outdegree | URL |
|---|---|---|
| 1 | 4 | www.yahoo.com.au |
| 2 | 5 | www.yahoo.co.uk |
| 3 | 3 | dir.yahoo.com |
| 4 | 7 | www.yahoo.com.sg |
| 5 | 8 | www.yahoo.ca |
| 6 | 9 | www2.aunz.yahoo.com |
| 7 | 1 | members.aol.com |
| 8 | 2 | www.geocities.com |
| 9 | 6 | members.tripod.com |
| 10 | 10 | ispc.yahoo.co.uk |
| 11 | 11 | y3.yahoo.ca |
| 12 | 12 | y4.yahoo.ca |
| 13 | 13 | www6.yahoo.co.uk |
| 14 | 16 | tv.yahoo.com.au |
| 15 | 17 | www.yahoo.co.nz |
| 16 | 19 | soccer.yahoo.com.au |
| 17 | 18 | www.yahoo.com.my |
| 18 | 21 | www.aunz.yahoo.com |
| 19 | 20 | 203.103.130.22 |
| 20 | 23 | 206.222.66.43 |

Table 2 lists the top hubs, ranked by HITS (first column) and by outdegree (second column). Here one sees very high correlation between the HITS ranking and outdegree ranking, indicating that our approximate analytical results are fairly accurate in this case.

We note, however, that the distinction between hubs and authorities is sometimes blurred. Good examples are members.aol.com, www.geocities.com, and similar sites, which are ranked very high in both authority list and hub list. Although they are not authoritative on any particular subject, careful content selection and organization on these websites make them valuable, almost like authoritative figures. This also happens in the bibliometrics domain, where some good survey papers/books (hubs) become as valuable or important as the original seminal papers (authorities), because these good surveys are written by authoritative people in the field, and they provide additional insights beyond the original seminal papers.

**7.2. Open Directory Project.** This dataset is about the topic *running*, which contains a total of 13,152 webpages. This dataset is a subcategory of a larger category, *fitness*, which is obtained from the Open Directory Project (ODP) www.dmoz.org. Under each category of the ODP, there is a relatively focused topic. The data file from the ODP contains the hierarchical structure of these webpages. We form the linkgraph of subcategory *running* by extracting from the *fitness* linkgraph the document IDs of those webpages under the *running* subcategory.

Table 3 lists the top 20 authorities, ranked either by HITS (first column) or by indegree (second column). Here the correlation between the HITS ranking and the indegree ranking is high. If we organize the ranking results in the first top 10, second top 10, etc., as done by many internet search engines, the matches within the first top 10 and the second top 10 are fairly close.

Table 4 lists the top hubs, ranked by HITS (first column) and by outdegree (second column). For the hub ranking, correlation between the HITS ranking and the

**Table 3** *Authority ranking for running.*

| HITS | Indegree | URL |
|---|---|---|
| 1 | 2 | www.runnersworld.com/ |
| 2 | 5 | sunsite.unc.edu/drears/running/running.html |
| 3 | 4 | www.usatf.org/ |
| 4 | 1 | www.coolrunning.com/ |
| 5 | 6 | www.clark.net/pub/pribut/spsport.html |
| 6 | 8 | www.runningnetwork.com/ |
| 7 | 9 | www.iaaf.org/ |
| 8 | 14 | www.sirius.ca/running.html |
| 9 | 12 | www.wimsey.com/∼dblaikie/ |
| 10 | 15 | www.kicksports.com/ |
| 11 | 7 | www.nyrrc.org/ |
| 12 | 18 | www.usaldr.org/ |
| 13 | 20 | www.halhigdon.com/ |
| 14 | 25 | www.ontherun.com/ |
| 15 | 10 | www.runningroom.com/ |
| 16 | 23 | www.webrunner.com/webrun/running/running.html |
| 17 | 22 | www.doitsports.com/ |
| 18 | 21 | www.arfa.org/ |
| 19 | 19 | www.adidas.com/ |
| 20 | 11 | www.uta.fi/∼csmipe/sport/ |

**Table 4** *Hub ranking for running.*

| HITS | Outdegree | URL |
|---|---|---|
| 1 | 3 | www.fix.net/∼doogie/links.html |
| 2 | 1 | www.gbtc.org/whatelse.html |
| 3 | 4 | www.usateamsports.com/running.htm |
| 4 | 15 | home1.gte.net/gregtrrc/links.htm |
| 5 | 17 | www.afn.org/∼ftc/othlinks.html |
| 6 | 19 | www.grainnet.com/rdraces/websites.html |
| 7 | 14 | www.runner.org/links.htm |
| 8 | 20 | directory.netscape.com/Health/Fitness/Running |
| 9 | 21 | www.dmoz.org/Health/Fitness/Running/ |
| 10 | 20 | directorysearch.mozilla.org/Health/Fitness/Running/ |
| 11 | 15 | dmoz.org/Health/Fitness/Running |
| 12 | 25 | www.cajuncup.com/links.htm |
| 13 | 11 | www.rrm.com/sites.html |
| 14 | 18 | www.doitsports.com/guides/running.html |
| 15 | 20 | www.webcrawler.com/kids_and_family/hobbies/outdoors/running |
| 16 | 20 | magellan.mckinley.com/lifestyle/hobbies_and_recreation/outdoors/... |
| 17 | 28 | www.webfanatix.com/running_resources.htm |
| 18 | 28 | www.webfanatix.com/_vti_bin/shtml.exe/running_resources.htm/map |
| 19 | 25 | www.isp.nwu.edu/∼brianw/running.html |
| 20 | 23 | www.geocities.com/HotSprings/Resort/5457/ |

indegree ranking is not as high as for the authority, but it is still apparent, especially if we look at the top three.

**8. Discussion.** We analyzed the HITS algorithm and obtained the closed-form solutions assuming that Web graphs are fixed degree sequence random graphs. Several important characteristics of the HITS algorithm were explained. One result is that, on average, the HITS authority ranking is the same as the ranking by indegree. Experiments on several Web groups support this result.

Besides HITS, another popular ranking algorithm is PageRank [8], used in the search engine Google. PageRank explores the linkgraph characteristics, but uses a random surf model that can be reviewed as hyperlink weight normalization. (HITS instead focuses on mutual reinforcement between authorities and hubs.) These main features of HITS and PageRank are generalized and combined into a unified framework in which one can show that ranking by PageRank is also highly correlated with ranking by indegree [13].

The key motivation of mutual reinforcement in HITS is that a "good" hub must point to several "good" authorities, while a "good" authority must be pointed to by several "good" hubs. The key motivation of PageRank is that an "informative" webpage must point to and be pointed to by other informative webpages. But for a webpage to become "informative" in the first place, it must have the quality to attract a certain number of inbound links, or votes from other webpages. The dynamics of the Web growth process [23, 4] has a snowball effect that gradually leads to the high correlation between "informativeness" and indegree. Thus mutual reinforcement and the high correlation between HITS ranking and indegree ranking describe different aspects of the Web growth process: one is from a relationship point of view, the other from a statistical point of view.

**Acknowledgment.** We thank the referees for valuable comments and suggestions.

## REFERENCES

[1] D. Achlioptas, A. Fiat, A.R. Karlin, and F. McSherry, *Web search via hub synthesis*, in Proceedings of the Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 2001, pp. 500–509.

[2] W. Aiello, F. Chung, and L. Lu, *A random graph model for massive graphs*, in Proceedings of the ACM Symposium on Theory of Computing, ACM, New York, 2000, pp. 171–180.

[3] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia, *Spectral analysis for data mining*, in Proceedings of the 33rd ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 619–626.

[4] A.-L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509–512.

[5] M.W. Berry, S.T. Dumais, and G. W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Rev., 37 (1995), pp. 573–595.

[6] K. Bharat and M.R. Henzinger, *Improved algorithms for topic distillation in a hyperlinked environment*, in Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR'98), ACM, New York, 1998, pp. 104–111.

[7] A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas, *Finding authorities and hubs from link structures on the world wide web*, in Proceedings of the 10th WWW Conference, ACM, New York, 2001, pp. 415–429.

[8] S. Brin and L. Page, *The anatomy of a large-scale hypertextual Web search engine*, Comput. Networks ISDN Systems, 30 (1998), pp. 107–117.

[9] A. Broder, R Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, *Graph structure in the web*, in Proceedings of the 9th International WWW Conference, Elsevier, Amsterdam, The Netherlands, 2000, pp. 247–256.

[10] S. Chakrabarti, B.E. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, *Automatic resource compilation by analyzing hyperlink structure and associated text*, Comput. Networks ISDN Systems, 30 (1998), pp. 65–74.

[11] D. Cohn and H. Chang, *Learning to probabilistically identify authoritative documents*, in Proceedings of the 17th International Conference on Machine Learning, Morgan-Kaufmann, San Francisco, CA, 2000, pp. 167–174.

[12] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, *Indexing by latent semantic analysis*, J. Amer. Soc. Inform. Sci., 41 (6) (1990), pp. 391–407.

[13] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, *PageRank, HITS and a unified framework for link analysis*, in Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR), ACM, New York, 2002, pp. 353–354.

[14] G. W. Flake, S. Lawrence, and C.L. Giles, *Efficient identification of web communities*, in Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), ACM, New York, 2000, pp. 150–159.

[15] D. Gibson, J. Kleinberg, and P. Raghavan, *Inferring web communities from link topology*, in Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98), ACM, New York, 1998, pp. 225–234.

[16] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University, Baltimore, MD, 1996.

[17] X. He, C. Ding, H. Zha, and H.D. Simon, *Automatic topic identification using webpages clustering*, in Proceedings of the IEEE International Conference on Data Mining, San Jose, CA, 2001, pp. 195–202.

[18] M.R. Henzinger, *Hyperlink analysis for the web*, IEEE Internet Comput., 5 (2001), pp. 45–50.

[19] Internet Archive, http://www.archive.org/.

[20] M. Kessler, *Bibliographic coupling between scientific papers*, Amer. Documentation, 14 (1963), pp. 10–25.

[21] J. Kleinberg and S. Lawrence, *The structure of the web*, Science, 294 (2001), pp. 1849–1850.

[22] J.M. Kleinberg, *Authoritative sources in a hyperlinked environment*, J. ACM, 48 (1999), pp. 604–632.

[23] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, *Extracting large-scale knowledge bases from the web*, in Proceedings of the 25th VLDB Conference, Morgan-Kaufmann, San Francisco, CA, 1999.

[24] R. Lempel and S. Moran, *SALSA: Stochastic approach for link-structure analysis and the TKC effect*, ACM Trans. Inform. Systems, 19 (2001), pp. 131–160.

[25] K.J. Maschhoff and D.C. Sorensen, *A portable implementation of ARPACK for distributed memory parallel computers*, in Proceedings of the Copper Mountain Conference on Iterative Methods, University of Colorado, Boulder, 1996.

[26] A.Y. Ng, A.X. Zheng, and M.I. Jordan, *Stable algorithms for link analysis*, in Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR), ACM, New York, 2001, pp. 258–266.

[27] H. Small, *Co-citation in the scientific literature: A new measure of the relationship between two documents*, J. Amer. Soc. Inform. Sci., 24 (4) (1973), pp. 265–269.