ANALYSIS OF WEB SEARCH ALGORITHM HITS*

DAWEI HONG
Department of Computer Science
Rutgers University at Camden
Camden, NJ 08102 USA
dhong@camden.rutgers.edu

SHUSHUANG MAN
Department of Mathematics and Computer Science
Southwest Minnesota State University
Marshall, MN 56258 USA
mans@southwestmsu.edu

Received 23 July 2003
Accepted 19 January 2004
Communicated by Oscar H. Ibarra

ABSTRACT
One of the well known Web search algorithms is HITS by Kleinberg [9]. We analyze the stability of HITS, when and how much outputs of HITS depend on initial values chosen by the algorithm. More importantly, we proposed a model for a type of hyperlink structures, which have been frequently observed on the Web, and we prove that in the model a crucial technical assumption made in HITS is satisfied, and accordingly HITS works well.

Keywords: Analysis of algorithms, Web search algorithms.

1. Introduction

It is observed that on the Web for a topic there are highly likely two sets of pages: An authoritative page set contains pages focused on the topic, and a hub page set contains pages linking to relevant pages on the topic. As a user enters a query on a topic (which consists of a couple of keywords), HITS applies two steps:

Step 1 With help from a text-based search engine, it finds a base set \( S \) of pages for the topic, and then constructs a directed graph \( G_S \) according to the hyperlinks among the pages in \( S \).

Step 2 It searches in \( G_S \) (which is a hyperlink structure), and returns a set of authoritative pages on the topic.

*Supported in part by NSF grants CCR 0310793 and 0310571.
We shall discuss how the base set is obtained and its effects on the performance of HITS in the last section. We refer reader to [9] for a complete view of HITS.

Each page in $S$ is a node of $G_S$ and vice versa. Let us denote the pages in $S$ by $v_i$, $i = 1, \ldots, n$. Let $A$ be the adjacency matrix of $G_S$. That is, $A$ is a 0-1 matrix whose $(i,j)$th entry is 1 if there is a link\footnote{The link here is understood as transverse [9], i.e., the two pages, $v_i$ and $v_j$, have different domain names.} from $v_i$ to $v_j$; otherwise, 0. By linear algebra we know that the symmetric matrix $A^T A$ is positive semi-definite,\footnote{Here, we have for any real $n$-vector $x$, $x^T (A^T A) x = (A x)^T (A x) = \|A x\|_2^2 \geq 0.$} and hence, all its eigenvalues are real and nonnegative (cf. [5]). Let us denote the $n$ eigenvalues of $A^T A$ by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. Let us denote an eigenvector of the largest eigenvalue $\lambda_1$ by

$$x^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_n^* \end{bmatrix}$$

Through this paper we shall use lower case letter for a vector, and use superscript * to indicate it is an eigenvector. and use subscript $i$ for the $i$th component of a vector. Also, we shall denote by $U_{\lambda_1}$ the eigenspace of the largest eigenvalue $\lambda_1$, and by $U_{\lambda_1}^\perp$ its complement.

In theory, HITS finds a set of authoritative pages according to $x^*$ as follows: With a predetermined parameter (cut-off) $c$ it chooses $c$ pages, $v_{i_1}, \ldots, v_{i_c}$, such that $x_{i_1}^*, \ldots, x_{i_c}^*$ are the $c$ largest coordinates in $x^*$. In practice, HITS employs a power iteration from linear algebra (cf. [6]): It takes an initial vector

$$x^{(0)} = \begin{bmatrix} x_{i_1}^{(0)} \\ \vdots \\ x_{i_c}^{(0)} \end{bmatrix}$$

whose components are all positive, and then computes for $t = 1, 2, \ldots$

$$x^{(t)} \leftarrow A^T A x^{(t-1)}$$

where right after each iteration $x^{(t)}$ is normalized such that $\sqrt{\sum_{i=1}^n (x_i^{(t)})^2} = 1$ for all $t \geq 1$. HITS runs this iteration until it becomes stable at $t = M$, and takes $x^{(M)}$ as $x^*$. A technical assumption is made at this point [9]:

**Assumption:** The multiplicity of the largest eigenvalue $\lambda_1$ equals one, i.e., $\lambda_1 > \lambda_2$.

In this paper, we address and answer the following three questions:

Q1 If a fixed text-based search engine is used in the Step 1 and if the assumption above holds, then to what extent the authoritative page set computed by HITS is independent on choices of initial vectors in the Step 2?

Q2 Without the assumption above, will the authoritative page set computed by HITS be dependent on choices of initial vectors in the Step 2?
Q3 What is a type of hyperlink structures (if any) such that if the directed graph $G_S$ obtained in the Step 1 has such a structure then the assumption above will be satisfied automatically? Moreover, how often hyperlink structures of the type are formed on the Web?

We explain our idea behind those three questions. HITS is designed for search on the Web. The Web has its intrinsic topological structures usually called hyperlink structures. A number of articles on those structures have recently been published (for example, see [1], [2], [4] and references therein). HITS is a link-analysis based algorithm. Given a query, HITS analyzes the hyperlink structure around the topic, and then computes a set of authoritative pages. Hyperlink structures are combinatorial. The method used by HITS to analyze such combinatorial structures is algebraic. In [9] and [10] quite a few experimental results showed that HITS works surprisingly well. This made us think on the Web there must be some type of hyperlink structures that support HITS. We couldn’t find an answer neither in [9] nor in anywhere else. A natural requirement (which is also crucial as we shall see later) on such a type of hyperlink structures is that HITS ought to determine a unique set of authoritative pages for a given query. Thus, our investigation begins with the stability of HITS. This was how the first two questions, Q1 and Q2, were formulated. A major goal was to discover a type of hyperlink structures such that (i) HITS is stable, and (ii) hyperlink structures of that type are frequently seen on the Web. This was the motivation behind Q3.

In [14] Ng et al analyzed the stability of HITS. They showed that if the eigengap (the difference between the largest and second largest eigenvalues) is small then HITS may be unstable under small perturbations. Also, the authors run HITS on the Cora database [13] with perturbations to demonstrate how unstable HITS may be. The Cora database collects citation information from thousands of research articles in computer science. It is known that citation networks have topological structures similar to what the Web does (cf. [2]). Our result Lemma 2 shows that in the worst case where the largest and second largest eigenvalues are equal, the instability of HITS can be severe: For a given query, depending on initial values the outputs of HITS can be arbitrary. Based upon their analysis [14] Ng et al proposed two improved versions, randomized and subspace HITS, and proved that the two are stable under small perturbations [15]. But in both [14] and [15] the authors didn’t investigate under what type of hyperlink structures HITS can work well as shown in experiments [9] and [10].

In [12] Kumar et al investigated hyperlink structures in communities on the Web. The authors systematically enumerated over 100,000 of those communities. They found that “Web communities are characterized by dense directed bipartite subgraphs” (p. 1484 [12]). Also, the authors pointed out that HITS works very well over such hyperlink structures (p. 1487 [12]). But they didn’t give any analytic reason why it is so. We define a type of hyperlink structures (Definition 1) which are more general than directed bipartite subgraphs, and prove that (Theorem 2)

---

*A community on the Web, as defined in [12], is a group of individuals who share a common interest together with their Web pages most popular amongst them.*
under any hyperlink structure of the type HITS can work well.

This paper is organized as follows. Section 2 is devoted to Q1 and Q2. We show that the stability of HITS heavily relies on the assumption mentioned above: With the technical assumption HITS is stable; while, without it the algorithm may be unstable. Section 3 is for Q3. We define a type of hyperlink (d-mature) structures, and prove that under that type of hyperlink structures HITS is stable. Also, we discuss how often d-mature structures are formed on the Web. Some remarks are in the last section.

2. Stability of HITS

Since a central part of HITS is a power iteration and the iteration takes an initial vector, an immediate question is about the stability of HITS: How much do outputs of HITS depend on choices of initial vectors? It was mentioned in [10] that “power iteration will converge to the principal eigenvector for any ‘non-degenerated’ choice of initial vector - in our case, for example, for any vector all of whose entries are positive.” We know that “non-degenerated” here means that a vector is not in the complement \( U_{\lambda_1}^\perp \). But it is not obvious at all why any vector all of whose entries are positive is not in the complement. We could not find an answer in either [9] or [10] or elsewhere. More seriously, we would question: Is the assumption crucial, i. e., what may happen without it? Those are two aspects of the stability of HITS, which we phrased as the first two questions in Section 1. We answer the two questions by two lemmas below.

**Lemma 1** If a text-based search engine used in the Step 1 is fixed and if the assumption (Section 1) holds, then we have there is a unique eigenvector \( x^* \in U_{\lambda_1} \) all of whose components are nonnegative, such that for any initial vector \( x^{(0)} \) with all components positive, iteration (1) converges to \( x^* \).

**Proof.** Using a fixed text-based search engine, in the Step 1 we obtain a fixed \( G_S \), and hence, a fixed adjacency matrix \( A \) for a given query. Thus, our only concern is if the iteration (1) converges to the same eigenvector for every initial vector with all its components positive.

By the assumption we have \( \dim(U_{\lambda_1}) = 1 \) and \( \dim(U_{\lambda_1}^\perp) = n - 1 \). Hence, we have \( U_{\lambda_1} = \{ ax^* : a \in R \} \), where \( x^* \) is an eigenvector of \( \lambda_1 \). By linear algebra (cf. [5]) we know that for any \( x^{(0)} \notin U_{\lambda_1}^\perp \), the iteration (1) converges to an eigenvector \( x^* \in U_{\lambda_1} \).

Let us consider the linear span \( L \) of all vectors, whose components are all positive

\[
L = \left\{ \sum_{i=1}^{n} a_i x^{(l)} : a_i \in R, \ x^{(l)} \in R^n \text{ and } x^{(l)} \text{ have their all components } > 0 \right\}
\]  

(2)

Clearly, we have \( \dim(L) = n \). Since \( \dim(U_{\lambda_1}^\perp) = n - 1 \), there must be at least one vector with all components positive, which is not in \( U_{\lambda_1}^\perp \). Now, we take that vector as the initial vector \( x^{(0)} \). Then in the iteration (1) we have the limit vector \( x^* \) with all components nonnegative, since all entries of \( A^T A \) are nonnegative. Taking the vector \( x^* \) we can write \( U_{\lambda_1} \) as \( \{ ax^* : a \in R \} \). This means that any vector in
$U_{\lambda_1}$ has either all its components nonnegative, or all its components nonpositive. Therefore, a vector with all components positive cannot be perpendicular to any non-zero vector in $U_{\lambda_1}$, and hence, cannot be in $U_{\lambda_1}^\perp$. Therefore, taking any vector with all components positive as the initial we have the iteration (1) converge to an eigenvector $x^* \in U_{\lambda_1}$ such that all components of $x^*$ are nonnegative. Since $	ext{dim}(U_{\lambda_1}) = 1$ and $\|x^*\|_2 = 1$, the $x^*$ is unique. □

Lemma 1 gives a positive answer to the question Q1 (Section 1). Recall that HITS chooses the $c$ authoritative pages according to the $c$ largest coordinates of $x^*$. The lemma shows that the assumption (Section 1) ensures the uniqueness of the choice of the $c$ authoritative pages. The next lemma shows that without the assumption we may face a unpleasant situation, in which the choice of the $c$ authoritative pages depends on the initial vector $x^{(0)}$ even if all components of $x^{(0)}$ are positive.

Suppose that the directed graph $G_S$ obtained in Step 2 of the HITS consists of two disjoint components. One component has $m_1$ nodes, and the other has $m_2$ nodes. We have $n = m_1 + m_2$. The adjacency matrix for the first component is $B_{m_1 \times m_1}$, and for the second one is $C_{m_2 \times m_2}$. Suppose that $B^TB$ and $C^TC$ have the same largest eigenvalue $\lambda_1$, and the multiplicity of $\lambda_1$ in both $B^TB$ and $C^TC$ is one. Without loss of generality we may write the adjacency matrix of $G_S$ as

$$A_{n \times n} = \begin{bmatrix} B_{m_1 \times m_1} & 0 \\ 0 & C_{m_2 \times m_2} \end{bmatrix}$$

We have

$$A^TA = \begin{bmatrix} B^TB & 0 \\ 0 & C^TC \end{bmatrix}$$

The largest eigenvalue of $A^TA$ has multiplicity two ($\lambda_1 = \lambda_2 = \kappa_1$). This means that the technical assumption (Section 1) made in HITS is not satisfied.

For $B^TB$ and $C^TC$ we let

$$\begin{bmatrix} x_1^* \\ \vdots \\ x_{m_1}^* \end{bmatrix}_{m_1 \times 1} \quad \text{and} \quad \begin{bmatrix} y_1^* \\ \vdots \\ y_{m_2}^* \end{bmatrix}_{m_2 \times 1}$$

be two eigenvector of the largest eigenvalue $\kappa_1$, respectively. Since $B^TB$ and $C^TC$ are two separate blocks, we can apply Lemma 1 respectively to $B^TB$ and $C^TC$. Then we have all components of the two eigenvectors above are nonnegative. Now, for the matrix $A^TA$ we have that given any pair of $\alpha, \beta > 0$ with $\alpha^2 + \beta^2 = 1$,

$$\begin{bmatrix} \alpha x_1^* \\ \vdots \\ \alpha x_{m_1}^* \\ \beta y_1^* \\ \vdots \\ \beta y_{m_2}^* \end{bmatrix}_{n \times 1} = \alpha \begin{bmatrix} x_1^* \\ \vdots \\ x_{m_1}^* \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1} + \beta \begin{bmatrix} 0 \\ \vdots \\ 0 \\ y_1^* \\ \vdots \\ y_{m_2}^* \end{bmatrix}_{n \times 1}$$

(3)
is an eigenvector of the largest eigenvalue $\lambda_1 = \lambda_2 = \kappa_1$.

Lemma 2 For any given $\alpha, \beta > 0$ with $\alpha^2 + \beta^2 = 1$, there is a vector $x^{(0)}$, all whose components are positive, so that taking $x^{(0)}$ as the initial we have the iteration (1) converge to (3).

Proof. We construct such a vector $x^{(0)}$. We first notice that if an eigenvector has all its components nonnegative then the sum of all those components must be positive. Thus, we have $\sum_{i=1}^{m_1} x_i^*, \sum_{i=1}^{m_2} y_i^* > 0$. Let $a = \sum_{i=1}^{m_1} x_i^*$ and $b = \sum_{i=1}^{m_2} y_i^*$. We have $a, b > 0$. Now, we let the first $m_1$ components of $x^{(0)}$ be $\frac{a}{a}$, and let the last $m_2$ components of $x^{(0)}$ be $\frac{b}{b}$. All components $x_i^{(0)}$, $i = 1, \ldots, n$, are positive.

We show that taking this $x^{(0)}$ as initial the iteration (1) converges to (3). We consider the following two eigenvectors of the largest eigenvalue $\lambda_1$:

$$u_1 = \begin{bmatrix} x_1^* \\ \vdots \\ x_{m_1}^* \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1} \text{ with } \|u_1\|_2 = 1,$$

$$u_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ y_1^* \\ \vdots \\ y_{m_2}^* \end{bmatrix}_{n \times 1} \text{ with } \|u_2\|_2 = 1.$$

Then we have that $\{u_1, u_2\}$ is an orthogonal basis for the eigenspace $U_{\lambda_1}$. Therefore, decomposing $x^{(0)}$ in $U_{\lambda_1} \oplus U_{\lambda_1}^\perp$, we have for $u \in U_{\lambda_1}^\perp$

$$x^{(0)} = \langle x^{(0)}, u_1 \rangle u_1 + \langle x^{(0)}, u_2 \rangle u_2 + u$$

where $\langle \cdot, \cdot \rangle$ stands for the inner product. Since $\langle x^{(0)}, u_1 \rangle = \alpha$ and $\langle x^{(0)}, u_2 \rangle = \beta$, we have

$$x^{(1)} = c_1 \left( A^T A \right) x^{(0)} = c_1 \left( (\lambda_1 \cdot \alpha)u_1 + (\lambda_1 \cdot \beta)u_2 + (A^T A) u \right)$$

where $c_1$ is the normalization factor. Thus, for all $t \geq 1$ we have

$$x^{(t)} = \prod_{i=1}^{t} c_i \left( (\lambda_1^t \cdot \alpha)u_1 + (\lambda_1^t \cdot \beta)u_2 + (A^T A)^t u \right)$$

(4)

where $c_i$ is the normalization factor for $x^{(i)}$.

By linear algebra (cf. [5]) we know that for all $t \geq 1$, $(A^T A)^t u \in U_{\lambda_1}^\perp$ (since $u$ is originally in $U_{\lambda_1}^\perp$). Combining this with the facts that (a) $\{u_1, u_2\}$ is an orthogonal basis of $U_{\lambda_1}$ and (b) $\alpha^2 + \beta^2 = 1$, by calculation we have

$$\prod_{i=1}^{t} c_i = \frac{1}{\lambda_1^t \sqrt{1 + \frac{1}{\lambda_1^2} \cdot \| (A^T A)^t u \|_2^2}}.$$  (5)

Recall that the $n$ eigenvalues of $A^T A$ are $\lambda_1 = \lambda_2 > \lambda_3 \geq \ldots \geq \lambda_n \geq 0$. With linear algebra (cf. [5]) we know that

$$\| (A^T A) u \|_2^2 \leq \lambda_2^2 \| u \|_2^2.$$
and hence,
\[
\frac{\| (A^T A)^t u \|_2^2}{\lambda_1^{2t}} \leq \left( \frac{\lambda_3}{\lambda_1} \right)^{2t} \| u \|_2^2.
\]

Putting this, (4) and (5) together, we have \( x^{(t)} \to \alpha u_1 + \beta u_2 \) as \( t \to \infty \). \( \Box \)

Lemma 2 answers the question Q2 (Section 1) showing that without the assumption (Section 1) HITS may be unstable, i.e., depending on initial values for the same query HITS may output arbitrarily (for any \( \alpha, \beta > 0 \) with \( \alpha^2 + \beta^2 = 1 \)). The meaning of Lemma 2 is clear. HITS essentially uses a principal eigenvector\(^d\) to find authoritative pages. Thus, for two disjoint hyperlink structures, which happen to have the same largest eigenvalue, HITS would hardly distinguish pages from the two hyperlink structures. From computational viewpoint, this matter could be serious: Once the two hyperlink structures have their largest eigenvalues close enough (not necessarily equal), the problem may occur. Also, notice that in Lemma 2 there are no restrictions either on \( m_1 \) and \( m_2 \) (the numbers of the pages in the two hyperlink structures), or on the formations of the two hyperlink structures.

But in experiments HITS works surprisingly well in a wide range of cases (cf. [9] and [10]). This shows that HITS does capture some intrinsic feature on the Web. We look at this phenomenon through another angle: The Web supplies some hyperlink structures in which HITS works well. Of what type are those hyperlink structures? We discuss this issue in the next section.

3. Hyperlink structures for HITS

This section, which consists of two subsections, is devoted to answering the question Q3 (Section 1). In the first subsection we define a type of hyperlink structure and discuss how often they are formed on the Web. In the second one we prove that HITS works well under any hyperlink structure of the type.

3.1. A type of hyperlink structures

It has been observed in many experiments that there are some intrinsic structural features among Web pages on a common topic. It would be a long list of references to cover all the reports. Here, we first quote a deep observation made by Kleinberg [9], and then in late of this subsection we introduce experimental results by Kumar et al [12]. Kleinberg observed that \textit{there should be considerable overlap in the sets of pages that point to authorities, and these pages pull together authorities on a common topic}. We propose a model for what has been observed above.

Let us recall the base set \( S \) and the directed graph \( G_S \) obtained in Step 1 of HITS (Section 1). We define two subsets \( S_A \) and \( S_H \) of \( S \),

\[
S_A \overset{\text{def}}{=} \{ v_i \in S : \text{in-degree of } v_i \geq 1 \},
\]

and

\[
S_H \overset{\text{def}}{=} \{ u_i \in S : \text{out-degree of } u_i \geq 1 \}.
\]

\(^d\)A principal eigenvector is an eigenvector of the principal eigenvalue which is the largest eigenvalue of multiplicity one.
Let us call \((S, G_S)\) a hyperlink structure where \(S\) is the set of all pages (as nodes) in the structure and the edges of the directed graph \(G_S\) represent hyperlinks among the pages.

**Definition 1** A hyperlink structure \((S, G_S)\) is said to be \(d\)-mature if

(i) \(S = S_A \cup S_H\) and \(|S_A| = d\), and

(ii) for any \(v_i \neq v_j \in S_A\) there are \(u_1, \ldots, u_k \in S_H\) and \(v_1, \ldots, v_{k+1} \in S_A\) such that \(v_1 = v_i, v_{k+1} = v_j\), and for all \(l = 1, \ldots, k\), \((u_l, v_l)\) and \((u_l, v_{l+1})\) are arcs in \(G_S\).

The figure below illustrates the definition of \(d\)-mature structure. \(v_i\) and \(v_j\) are two pages in \(S_A\), which are potentially authoritative on a topic. \(u_1, \ldots, u_k\) are pages in \(S_H\), which are potentially hubs, pulling \(v_i\) and \(v_j\) together.

![Diagram showing a hyperlink structure with nodes \(v_i, v_2, v_3, \ldots, v_k, v_{k-1}, v_k, v_{k-1}, v_k\) and edges connecting them to \(u_1, u_2, u_3, \ldots, u_{k-2}, u_{k-1}, u_k\).]

This model reflects our viewpoint toward authoritative and hub pages, which is much influenced by Kleinberg’s idea quoted above: It is impossible to define what authoritative and hub pages are. But we can describe how authoritative and hub pages behave. Authoritative pages are pointed by hub pages. Thus, all authoritative pages should be in \(S_A\), and all hub pages should be in \(S_H\). Authorities on the Web are established via common recognitions. To have such common recognitions, pages on one topic should be connected as one component. But such a component may not be simply a connected subgraph of the Web graph. (ii) of the definition above states that such connections can be “remote”: \(v_i\) and \(v_j\) are connected via \(u_1, \ldots, u_k\) and \(v_1, \ldots, v_{k+1}\). We use term “mature” meaning that the \(d\) potential authoritative pages in \(S_A\) have been pulled together by some hub pages in \(S_H\). Notice that our model allows \(S_A \cap S_H \neq \emptyset\). This means that some pages may play “double role” being potentially authority and hub, which has been observed on the Web.

We discuss if a \(d\)-mature structure is practical by investigating the chance for a \(d\)-mature structure being formed on the Web. In terms of graph theory we give an equivalence of Definition 1. Given a hyperlink structure \((S, G_S)\), we construct a undirected graph \(\bar{G}(S)\) such that its nodes are the pages in \(S_A\), and for any \(v_i \neq v_j \in S_A\) there is an edge between them if and only if there exists \(u \in S_H\) so that \((u, v_i)\) and \((u, v_j)\) are arcs in \(G_S\). We have

**Proposition 1** \((S, G_S)\) is \(d\)-mature iff the undirected graph \(\bar{G}(S)\) is connected. □

We recall a classic result in graph theory, which was by Erdős and Rényi (see Theorem 9 in chapter VII of [3]). For our usage we state the result as follows: Constructing a undirected graph \(G\), we begin with \(d\) nodes having no edges among them, and we determine a probability \(0 < p_d < 1\) with which each of \(d(d-1)/2\) edges (those are potentially to be the edges) is added to \(G\).

**Theorem 1** Let \(p_d = \frac{\ln d}{d}\). Then for all \(c > 1\), almost every \(G\) is connected and for all \(c < 1\), almost every \(G\) is disconnected. In particular, for case of \(c > 1\),
the probability of $G$ being connected is greater than $1 - 4\exp(-(c - 1)\ln d)$. □

Kumar et al [12] have observed that on the Web there are several hundred thousand of communities. Systematically enumerating over 100,000 instances the authors concluded that hyperlink structures of communities on the Web are characterized by dense directed bipartite subgraphs. Then in [11] stochastic models for the Web graph were proposed, and it was proved that large number of complete directed bipartite subgraphs can be generated in the models (see Theorem 14 and 15 in [11]).

We focus on how often a dense directed bipartite subgraph is a $d$-mature structure. Recall that (p. 1484 [12]) a directed bipartite graph is defined as a directed graph whose node set can be partitioned into two disjoint subsets such that every direct edge in the graph is from a node in the second subset to a node in the first subset. In [12] the authors didn’t give a mathematical description of how dense is enough to characterize communities on the Web (section 1.2 of the paper). To the best of our knowledge, the results in section 4 of [11] are the only models for the Web graph in which a large number of complete directed bipartite subgraphs are generated. Neither in [11] nor anywhere else, the distribution of densities of dense directed bipartite subgraphs of the Web graph has been mathematically modeled and investigated. But, still, we can put dense directed bipartite subgraph into the frame of $d$-mature structure,

The Web graph is a random graph (cf. [1], [2], [4] and [11]), and thus, so are its dense directed bipartite subgraphs. We use our notation $(S, G_S)$ to denote a dense directed bipartite subgraph. We shall set a probabilistic model for the undirected graph $\tilde{G}(S)$. A probabilistic model for directed bipartite subgraph was proposed in section 1.2 of [12]: "Let $B$ a random bipartite graph with edges directed from a set $L$ of nodes to a set $R$ of nodes, with $m$ random edges each placed between a vertex of $L$ and a vertex of $R$ uniformly at random." We notice that this model can be mathematically interpreted in two different ways. There are $|L| \times |R|$ pairs of vertices each of which may have edge(s). One way to interpret "uniformly at random" is that we consider all patterns in which $m$ edges are placed such that no pairs have multiple edges, and we assign each pattern with equal probability. The other way is that we independently repeat the following procedure $m$ times: With probability $\frac{1}{|L| \times |R|}$ we choose a pair of vertices, and then let the pair have an edge. Here, we adopt the later interpretation. We notice that it allows multiple edges for a pair, which we interpret as one page repeatedly refers the other. Take $(S, G_S)$ as $B$. Then it easy to verify the following: $\tilde{G}(S)$ is a random graph with $d$ nodes, which is generated as follows: With a fixed probability $p$ each of the $\frac{d(d-1)}{2}$ edges is independently added to the graph. Here, the value of $p$ depends on $|L|$, $|R|$ and $m$, and is of secondary importance to us. Carrying this notion we introduce

**A probabilistic model for $\tilde{G}(S)$:** For each $\{v_i, v_j\}$, $1 \leq i < j \leq d$, of the $\frac{d(d-1)}{2}$ pairs of vertices there is probability $p_{i,j}$ with which an edge is added to $\tilde{G}(S)$ independently.

The intuition behind this model is as follows: Each pair of nodes is a pair of pages of $S_A$ which are potentially authoritative on the same topic (since they belong
Corollary 1 Suppose that \( (S, G_S) \) is a directed bipartite graph. Under the probabilistic model above we consider the undirected graph \( \tilde{G}(S) \). If \( \min_{1 \leq i < j \leq d} \{(p_{i,j}) \geq \frac{3\ln d}{d} \}
\) then

(i) \( (S, G_S) \) is \( d \)-mature with probability greater than \( 1 - 4d^{-2} \), and

(ii) the in-density of \( (S, G_S) \) is greater than \( 3(1 - \frac{1}{d}) \ln d - 2\sqrt{(1 - \frac{1}{d}) \ln d} \) with probability not less than \( 1 - d^{-4} \).

Proof. Following the proof of Theorem 1 (pp. 232-234 of [3]) and letting \( c = 3 \) we have with probability greater than \( 1 - 4d^{-2} \), \( \tilde{G}(S) \) is connected. Then (i) follows from Proposition 1.

As for (ii) we recall that the density of a undirected graph is defined as its average degree. Since \( (S, G_S) \) is a directed bipartite graph, its in-density is defined as the total of in-degrees over the number of its nodes having non-zero in-degrees. It is easy to see that the in-density of \( (S, G_S) \) is not less than the density of \( \tilde{G}(S) \). Applying Hoeffding’s inequality [8], by calculation we have for all \( t > 0 \) with probability \( \geq 1 - \exp\left(-\frac{4t^2}{d(d-1)}\right) \), the number of edges in \( \tilde{G}(S) \) is greater than \( \frac{3(d-1)\ln d}{2} - t \), which means the total of degrees in \( \tilde{G}(S) \) is greater than \( 3(d - 1) \ln d - 2t \). Then (ii) follows from letting \( t = \sqrt{d(d-1)\ln d} \).

Consider \( (S, G_S) \) as a directed bipartite subgraph. We have \( S = S_A \cup S_H \), \( S_A \cap S_H = \emptyset \) and \( |S_A| = d \). The corollary above tell us that if for each two \( v_i \neq v_j \in S_A \) with probability \( \geq \frac{3\ln d}{d} \) there is \( u \in S_H \) linking to both \( v_i \) and \( v_j \), then \( (S, G_S) \) is \( d \)-mature with probability \( \geq 1 - 2d^{-2} \), and in the mean time, with probability \( \geq 1 - d^{-4} \) the average number of links pointing to a page in \( S_A \) can be greater than \( 3(1 - \frac{1}{d}) \ln d - 2\sqrt{(1 - \frac{1}{d}) \ln d} \). Notice that there is no restriction on the number of pages in \( S_H \). Let us take \( d = 40 \) as an example. We have \( \frac{3\ln d}{d} < 0.277 \) and \( 1 - 4d^{-2} \geq 0.998 \). This implies that if for any two pages in \( S_A \) with probability \( \geq 0.277 \) there is a page in \( S_H \) linking to the two pages, then with probability \( > 0.998 \) the directed bipartite subgraph is \( d \)-mature, and with probability \( \geq 1 - \frac{1}{2,560,000} \) the in-density of \( (S, G_S) \) can be greater than \( 3(1 - \frac{1}{d}) \ln d - 2\sqrt{(1 - \frac{1}{d}) \ln d} \approx 7 \).

Using the probabilistic model for \( \tilde{G}(S) \) we reasoned why dense directed bipartite subgraphs of the Web graph are likely \( d \)-mature. Several hundred thousand of those subgraphs have been observed [12]. A \( d \)-mature structure \( (S, G(S)) \) may have \( S_A \cap S_H = \emptyset \) while a directed bipartite subgraph must have \( S_A \cap S_H = \emptyset \). It is not hard to verify that all proofs presented in this subsection are applicable to the case when \( (S, G(S)) \) is a \( d \)-mature structure. To this extent, \( d \)-mature structure is a generalization of directed bipartite subgraph. Now, if we can prove that HITS works well under any \( d \)-mature structure, then we have an analytic explanation why HITS works surprisingly well as observed in [9], [10] and [12]. We prove this in next subsection.

3.2. The type is for HITS
We prove that HITS works well under any d-mature structure with \( d > 1 \).

**Theorem 2** For all \( d > 1 \), if \((S, G_S)\) is d-mature then the largest eigenvalue \( \lambda_1 \) of the matrix \( A^T A \) used for the power iteration (1) has multiplicity one, i.e., the technical assumption (Section 1) is satisfied. Consequently, the HITS uniquely returns a set of the authoritative pages.

To prove this theorem we need a technical lemma. Without loss of generality we number the nodes of \( G_S \) in such a way that the \( d \) nodes in \( S_A \) are \( v_1, \ldots, v_d \).

**Lemma 3** For all \( d > 1 \), if \((S, G_S)\) is d-mature then we have \( \lambda_1 > 0 \). Moreover, we have

(i) for any eigenvector \( x^* \in U_{\lambda_1}, x_i^* = 0 \) for all \( i = d + 1, \ldots, n \), and

(ii) there is an eigenvector \( x^* \in U_{\lambda_1} \) such that \( x_i^* > 0 \) for all \( i = 1, \ldots, d \).

**Proof.** Since the matrix \( A^T A \) is positive semi-definite, all its eigenvalues are non-negative. Since \((S, G_S)\) is d-mature and \( d > 1 \), there are at least two entries in the adjacency matrix \( A \), which are 1. Thus, the trace of \( A^T A \) is positive. This means the sum of all eigenvalues of \( A^T A \) (which are all nonnegative) are positive, which implies the largest eigenvalue \( \lambda_1 > 0 \).

Recall that we number the nodes in \( G_S \) in such a way that the \( d \) nodes in \( S_A \) are \( v_1, \ldots, v_d \), i.e., we number the nodes of non-zero in-degrees, by 1, ..., \( d \). Hence, in the adjacency matrix \( A \) we have zero columns from \( d + 1 \) to \( n \). Thus, we have

\[
A^T A = \begin{bmatrix}
H_{d \times d} & 0 \\
0 & 0
\end{bmatrix}
\]

where \( H_{d \times d} \) is the only non-zero block. The representation of \( A^T A \) above implies that the \((d + 1)\)th to \( n\)th components of \( A^T A x^* \) are all zero. Since \( A^T A x^* = \lambda_1 x^* \) and \( \lambda_1 > 0 \), we have \( x_i^* = 0 \) for all \( i = d + 1, \ldots, n \). This completes a proof for (i).

We now turn to prove (ii). First, we show that there is a \( x^* \in U_{\lambda_1} \) such that \( x_i^* \geq 0 \) for all \( i = 1, \ldots, d \). Let us recall the linear span \( L \) ((2) in Section 2). We have \( \dim(L) = n \). On the other hand we have \( \dim(U_{\lambda_1}^\perp) < n \). Thus, \( U_{\lambda_1}^\perp \) cannot contain all vectors whose components are all positive; otherwise, \( U_{\lambda_1}^\perp \) would contain the linear span \( L \), and it is false that a linear space of dimension less than \( n \) contains a linear space of dimension \( n \). Hence, we have a vector \( x^{(0)} \notin U_{\lambda_1}^\perp \) with all its components positive. We take \( x^{(0)} \) as the initial vector for the iteration (1). Since \( x^{(0)} \notin U_{\lambda_1}^\perp \), by linear algebra (cf. [5]) we have the iteration (1) converge to an eigenvector \( x^* \in U_{\lambda_1} \) with \( x_i^* \geq 0 \) for all \( i = 1, \ldots, d \) (since all entries of the matrix \( A^T A \) used in the iteration (1) are nonnegative).

Note that at this point we cannot yet rule out the case of \( \dim(U_{\lambda_1}) > 1 \). But even in that case, as long as \( x^{(0)} \notin U_{\lambda_1}^\perp \), the limit is an eigenvector in \( U_{\lambda_1} \), though different initial vector \( x^{(0)} \) may cause different eigenvector as limit.

Next, we show that for that eigenvector \( x^* \) (obtained above) we have \( x_i^* > 0 \) for all \( i = 1, \ldots, d \). Since \( x_i^* = 0 \) for all \( i = d + 1, \ldots, n \), there must be at least one \( x_{i_0}^* > 0 \) for some \( 1 \leq i_0 \leq d \) (because \( x^* \) is an eigenvector). Without loss of generality we assume that \( i_0 = 1 \). Let us consider \( x_i^* \), \( 2 \leq i \leq d \). Since both \( v_1 \) and \( v_i \) are in \( S_A \) and since \((S, G_S)\) is d-mature, by Definition 1 we have
$v_j, \ldots, v_k \in S_H$ and $v_1, \ldots, v_{k+1} \in S_A$ such that $v_1 = v_i, v_{k+1} = v_i$, and for all $q = 1, \ldots, k$, $(v_j, v_q)$ and $(v_j, v_{q+1})$ are arcs in $G_S$. This implies that in the adjacency matrix $A$ we have the following: All $(j_q, l_q)$th and $(j_q, l_{q+1})$th entries, $q = 1, \ldots, k$, are 1. Using this and the facts that (a) $x_i^* > 0$ and (b) all components of $x^*$ are nonnegative, by matrix multiplication we can see that the $i$th component of $(A^T A)^k x^*$ is positive. On the other hand, since $x^*$ is an eigenvector of the eigenvalue $\lambda_1$, we have $(A^T A)^k x^* = \lambda_1^k x^*$. Because $\lambda_1 > 0$, we have the $i$th component $x_i^*$ of $x^*$ is positive. This shows that for each $i$, $2 \leq i \leq d$, $x_i^* > 0$. A proof for (ii) is completed.

From the proof above, we can deduce the following:

**Corollary 2** For any eigenvector $x^* \in U_{\lambda_1}$, the last $(n-d)$ components are always zero; as for the first $d$ components, if one component is positive and all other $d-1$ are nonnegative, then the first $d$ components must be all positive provided that $(S, G_S)$ is $d$-mature and $d > 1$. □

So far, we see that a $d$-mature ($d > 1$) hyperlink structure $(S, G_S)$ has a desired property: There is an eigenvector $x^*$ of the largest eigenvalue such that the first $d$ components of $x^*$ are positive, which reflect authority weights (since the pages corresponding to those components have non-zero in-degrees), the last $(n-d)$ components are zero, which reflect no authority weights (since pages corresponding to those components have zero in-degree). For a given $(S, G_S)$, the set of authoritative pages (if any) ought be unique. Knowing the result of Lemma 2, we naturally would ask the following question: Is such an eigenvector $x^*$ in (ii) of Lemma 3 unique? Or equivalently, is the multiplicity of the largest eigenvalue $\lambda_1$ one? Theorem 2 confirms this is indeed the case provided that $(S, G_S)$ is $d$-mature and $d > 1$.

**A proof of Theorem 2.** We prove this theorem by contradiction. Suppose that the multiplicity of $\lambda_1$ is greater than one. Then there are two linearly independent vectors $x^*$ and $y^*$ in $U_{\lambda_1}$. By (i) of Lemma 3 we can have

$$
x^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_d^* \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1} \quad \text{and} \quad y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_d^* \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1}
$$

where by (ii) of Lemma 3 we can have $x_i^* > 0$ for all $i = 1, \ldots, d$.

Since $x_i^* > 0$ for all $i = 1, \ldots, d$, we take a small enough $\epsilon > 0$ such that the first $d$ components of $w^* = x^* + \epsilon y^*$ are all positive. Then the linear independence of $x^*$ and $y^*$ leads to the linear independence of $w^*$ and $x^*$. Thus, we have two linearly independent vectors $w^*, x^* \in U_{\lambda_1}$, both of which have the first $d$ components positive and the rest of components zero. Let $\gamma = \min \left\{ \frac{x_i^*}{w_i^*} : i = 1, \ldots, d \right\}$, and let $z^* = x^* - \gamma w^*$. Note that $z^* \in U_{\lambda_1}$. Without loss of generality we assume that $\gamma = \frac{x_1^*}{w_1^*}$. Then we have $z_1^* = 0$. Let us consider the first $d$ components of $z^*$. We have that all $z_i^* \geq 0$, $i = 1, \ldots, d$, (by the definitions of $z^*$ and $\gamma$), and that there
is at least one $z_i^* > 0$, $1 < i \leq d$, (since $x^*$ and $w^*$ are linearly independent). By Corollary 2 this means that all $z_i^* > 0$, $i = 1, \ldots, d$, contradicting $z_1^* = 0$. \hfill \Box

4. Concluding remarks

One issue about analysis of HITS is, as raised in [7], the time complexity of HITS, i.e., bounds on the number of iterations in the Step 2 (Section 1). This issue is related to the text-based engine used in the Step 1 (Section 1). We notice that in [9] there is no specification on the text-based engine. Let us look at the Step 1 closely: How a base set $S$ and the corresponding directed graph $G(S)$ for a given query are obtained. Taking a query by user HITS begins with an external text-based search engine to get pages relevant to the topic in the query. The number of those pages is usually large. With a parameter $q$ (typically set to 200) HITS collects the $q$ highest ranked pages to form the root set. In order to get the base set $S$, HITS expands the root set as follows: For each page in the root set, include any page pointed by that page (with no restriction), and include any page pointing to that page with restriction that at most $b$ such pages are included (where $b$ is a pre-determined parameter). Then the directed graph $G(S)$ is constructed such that its nodes are the pages in $S$ and its arcs are the hyperlinks among pages in $S$. Thus, the hyperlink structure $(S,G(S))$ is essentially by the text-based search engine. Recall the adjacency matrix $A$ for $G(S)$. Let us assume that the assumption (Section 1) holds, and let us denote by $\lambda_1$ and $\lambda_2$ the largest eigenvalue and the second largest eigenvalue of $A^T A$, respectively. Then by linear algebra (cf. [5]) we know that the convergence rate of the power iteration (1) is generally determined by $(\lambda_2/\lambda_1)^t$ where $t$ is the number of iterations taken. Though $0 < \lambda_2/\lambda_1 < 1$, it is an attribute of the matrix $A$, and hence, it depends ultimately on the text-based search engine.

Another issue is about the outcome of HITS. Are authoritative pages by HITS authorities? This is a difficult question, since different groups of individuals do not have the same authorities even on the same topic. In all experimental results published thus far, authoritative pages are authorities, and we believe this is the case for most of queries. But in our experiments we found an interesting example. The query is “card game bridge” whose meaning is not ambiguous. We happen to know the game well (the first author played at a number of professional tournaments). The authorities ought to include WBF (World Bridge Federation) and ACBL (American Contract Bridge League) both of which have Web sites. But HITS (and the google) fails to return both. A few on-line game sites are returned by HITS. This is fine. But among all game sites the Yahoo! Games has much higher authority than the Okbridge. In reality, the former is an entertaining game zone, while the later is for players at professional level. This example shows that authoritative pages computed by HITS may be just simply defined by the Internet culture.

Acknowledgements

The authors are grateful to an anonymous referee whose instructive comments
helped us improve the early version of the paper. Also, the authors thank Jean-Camille Birget for some discussions on the draft of this paper.

References
