To appear in the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01).

Link Analysis, Eigenvectors and Stability

Andrew Y. Ng
Computer Science Division
U.C. Berkeley
Berkeley, CA 94720

Alice X. Zheng
Computer Science Division
U.C. Berkeley
Berkeley, CA 94720

Michael I. Jordan
CS Div. & Dept. of Statistics
U.C. Berkeley
Berkeley, CA 94720

Abstract

The HITS and the PageRank algorithms are eigenvector methods for identifying "authoritative" or "influential" articles, given hyperlink or citation information. That such algorithms should give consistent answers is surely a desideratum, and in this paper, we address the question of when they can be expected to give stable rankings under small perturbations to the hyperlink patterns. Using tools from matrix perturbation theory and Markov chain theory, we provide conditions under which these methods are stable, and give specific examples of instability when these conditions are violated. We also briefly describe a modification to HITS that improves its stability.

1 Introduction

Recent years have seen growing interest in algorithms for identifying "authoritative" or "influential" articles from webpage hyperlink structures or from other citation data. In particular, the HITS algorithm of Kleinberg [1998] and Google's PageRank algorithm [Brin and Page, 1998] have attracted the attention of many researchers (see also [Osareh, 1996] for earlier developments in the bibliometrics literature). Both of these algorithms use eigenvector calculations to assign "authority" weights to articles, and while originally designed in the context of link analysis on the web, both algorithms can be readily applied to citation patterns in academic papers and other citation graphs.

There are several aspects to the evaluation of a link analysis algorithm such as HITS or PageRank. One aspect relates to the specific notion of "authoritativeness" embodied by an algorithm. Thus specific users may have an understanding of what constitutes an authoritative web page or document in a given domain, and the output of HITS or PageRank can be evaluated by such users. While useful, such analyses often have a rather subjective flavor. A more objective criterion—the focus of the current paper—concerns the *stability* of a link analysis algorithm. Does an algorithm return similar results upon a small perturbation of the link structure or the document collection? We view stability as a desirable feature of a link analysis algorithm, above and beyond the particular notion of authoritativeness that the algorithm embodies. If an article is truly authoritative or influential, then surely the addition of a few links or a few citations should not make us change our minds about these sites or articles having been very influential. Moreover, even in the context of a fixed link structure, a dynamic, unreliable infrastructure such as the web may give us different views of the structure on different occasions. Ideally, a link analysis algorithm should be insensitive to such perturbations.

In this paper, we use techniques from matrix perturbation theory and coupled Markov chain theory to characterize the stability of the ranks assigned by HITS and PageRank. Some ways of improving the stability of HITS are also briefly metioned; these algorithmic changes are studied in more detail in [Ng *et al.*, 2001].

2 An Example

Let us begin with an empirical example. The *Cora* database [McCallum *et al.*, 2000] is a collection containing the citation information from several thousand papers in AI.

We ran the HITS and PageRank algorithms on the subset of the *Cora* database consisting of all its Machine Learning papers. To evaluate the stability of the two algorithms, we also constructed a set of five perturbed databases in which 30% of the papers from the base set were randomly deleted. ("Since *Cora* obtained its database via a web crawl, what if, by chance or mishap, it had instead retrieved only 70% of these papers?") If a paper is truly authoritative, we might hope that it would be possible to identify it as such with only a subset of the base set.

The results from HITS are shown in the following table. In this table, the first column reports the rank from HITS on the full set of Machine Learning papers, whereas the five rightmost columns report the ranks in runs on the perturbed databases. We see substantial variation across the different runs:

1	"Genetic algorithms in search, optimization...", Goldberg	1	3	1	1	1
2	"Adaptation in natural and artificial systems", Holland	2	5	3	3	2
3	"Genetic programming: On the programming of...", Koza	3	12	6	6	3
4	"Analysis of the behavior of a class of genetic...", De Jong	4	52	20	23	4
5	"Uniform crossover in genetic algorithms", Syswerda	5	171	119	99	5
6	"Artificial intelligence through simulated...", Fogel	6	135	56	40	8
7	"A survey of evolution strategies", Back+al	10	179	159	100	7
8	"Optimization of control parameters for genetic...", Grefenstette	8	316	141	170	6
9	"The GENITOR algorithm and selection pressure", Whitley	9	257	107	72	9
10	"Genetic algorithms + Data Structures = ...", Michalewicz	13	170	80	69	18
11	"Genetic programming II: Automatic discovey...", Koza	7	-	-	-	10
2060	"Learning internal representations by error...", Rumelhart+al	-	1	2	2	-

2061	“Learning to predict by the method of temporal...”, Sutton	-	9	4	5	-
2063	“Some studies in machine learning using checkers”, Samuel	-	-	10	10	-
2065	“Neuronlike elements that can solve difficult...”, Barto+Sutton	-	-	8	-	-
2066	“Practical issues in TD learning”, Tesauro	-	-	9	9	-
2071	“Pattern classification and scene analysis”, Duda+Hart	-	4	7	7	-
2075	“Classification and regression trees”, Breiman+al	-	2	5	4	-
2117	“UCI repository of machine learning databases”, Murphy+Aha	-	7	-	8	-
2174	“Irrelevant features and the subset selection...”, John+al	-	8	-	-	-
2184	“The CN2 induction algorithm”, Clark+Niblett	-	6	-	-	-
2222	“Probabilistic reasoning in intelligent systems”, Pearl	-	10	-	-	-

Although it might be thought that this variability is intrinsic to the problem, this is not the case, as shown by the results from the PageRank algorithm, which were much more stable:

1	“Genetic Algorithms in Search, Optimization and...”, Goldberg	1	1	1	1	1
2	“Learning internal representations by error...”, Rumelhart+al	2	2	2	2	2
3	“Adaptation in Natural and Artificial Systems”, Holland	3	5	6	4	5
4	“Classification and Regression Trees”, Breiman+al	4	3	5	5	4
5	“Probabilistic Reasoning in Intelligent Systems”, Pearl	5	6	3	6	3
6	“Genetic Programming: On the Programming of ...”, Koza	6	4	4	3	6
7	“Learning to Predict by the Methods of Temporal ...”, Sutton	7	7	7	7	7
8	“Pattern classification and scene analysis”, Duda+Hart	8	8	8	8	9
9	“Maximum likelihood from incomplete data via...”, Dempster+al	10	9	9	11	8
10	“UCI repository of machine learning databases”, Murphy+Aha	9	11	10	9	10
11	“Parallel Distributed Processing”, Rumelhart+McClelland	-	-	-	10	-
12	“Introduction to the Theory of Neural Computation”, Hertz+al	-	10	-	-	-

These results are discussed in more detail in Section 6. It should be stated at the outset, however, that our conclusion is not that HITS is unstable while PageRank is not. The issue is more subtle than that, involving considerations such as the relationships between multiple eigenvectors and invariant subspaces. We do wish to suggest, however, that stability is indeed an issue that needs attention. We now turn to a brief description of HITS and PageRank, followed by our analysis.

3 Overview of HITS and PageRank

Given a collection of web pages or academic papers linking to/citing each other, the HITS and PageRank algorithms each (implicitly) construct a matrix capturing the citation patterns, and determines authorities by computing the principal eigenvector of the matrix.[1]

3.1 HITS algorithm

The HITS algorithm [Kleinberg, 1998] posits that an article has high “authority” weight if it is linked to by many pages with high “hub” weight, and that a page has high hub weight if it links to many authoritative pages. More precisely, given a set of n web pages (say, retrieved in response to a search query), the HITS algorithm first forms the n-by-n adjacency matrix A, whose (i, j)-element is 1 if page i links to page j, and 0 otherwise.[2] It then iterates the following equations:

$$a_i^{(t+1)} = \sum_{j:j \to i} h_j^{(t)}; \; h_i^{(t+1)} = \sum_{j:i \to j} a_j^{(t+1)}$$

(where “$i \to j$” means page i links to page j) to obtain the fixed-points $a^* = \lim_{t \to \infty} a^{(t)}$ and $h^* = \lim_{t \to \infty} h^{(t)}$ (with the vectors renormalized to unit length). The above equations can also be written:

$$\begin{aligned} a^{(t+1)} &= A^T h^{(t)} = (A^T A) a^{(t)} \\ h^{(t+1)} &= A a^{(t+1)} = (A A^T) h^{(t)}. \end{aligned}$$

When the iterations are initialized with the vector of ones $[1, \ldots, 1]^T$, this is the power method of obtaining the principal eigenvector of a matrix [Golub and Van Loan, 1996], and so (under mild conditions) a^* and h^* are the principal eigenvectors of $A^T A$ and $A A^T$ respectively. The “authoritativeness” of page i is then taken to be a_i^*, and likewise for hubs and h^*.

3.2 PageRank algorithm

Given a set of n web pages and the adjacency matrix A (defined previously), PageRank [Brin and Page, 1998] first constructs a probability transition matrix M by renormalizing each row of A to sum to 1. One then imagines a random web surfer who at each time step is at some web page, and decides which page to visit on the next step as follows: with probability $1 - \epsilon$, she randomly picks one of the hyperlinks on the current page, and jumps to the page it links to; with probability ϵ, she “resets” by jumping to a web page picked uniformly and at random from the collection.[3] Here, ϵ is a parameter, typically set to 0.1-0.2. This process defines a Markov chain on the web pages, with transition matrix $\epsilon U + (1 - \epsilon) M$, where U is the transition matrix of uniform transition probabilities ($U_{ij} = 1/n$ for all i, j). The vector of PageRank scores p is then defined to be the stationary distribution of this Markov chain. Equivalently, p is the principal eigenvector of the transition matrix $(\epsilon U + (1 - \epsilon) M)^T$ (see, e.g. Golub and Van Loan, 1996), since by definition the stationary distribution satisfies

$$(\epsilon U + (1 - \epsilon) M)^T p = p. \tag{1}$$

The asymptotic chance of visiting page i, that is, p_i, is then taken to be the “quality” or authoritativeness of page i.

4 Analysis of Algorithms

We begin with a simple example showing how a small addition to a collection of web pages can result in a large change to the eigenvectors returned. Suppose we have a collection of web pages that contains 100 web pages linking to `http://www.algore.com`, and another 103 web pages

[1] It is worth noting that HITS is typically described as running on a small collection of articles (say retrieved in response to a query), while PageRank is described in terms of the entire web. Either algorithm can be run in either setting, however, and this distinction plays no role in our analysis.

[2] Kleinberg [1998] discusses several other heuristics regarding issues such as intra-domain references, which are ignored in this section for simplicity (but are used in our experiments). See also Bharat and Henzinger [1998] for other improvements to HITS. It should be noted that none of these fundamentally change the spirit of the eigenvector calculations underlying HITS.

[3] There are various ways to treat the case of pages with no outlinks (leaf nodes). In this paper we utilize a particularly simple approach—upon reaching such a page, the web surfer picks the next page uniformly at random. This means that if a row of A has all zero entries, then the corresponding row of M is constructed to have all entries equal to $1/n$. The PageRank algorithm described in [Page *et al.*, 1998] utilizes a different reset distribution upon arriving at a leaf node. It is possible to show, however, that every instantiation of our variant of the algorithm is equivalent to an instantiation of the original algorithm on the same graph with a different value of the reset probability.

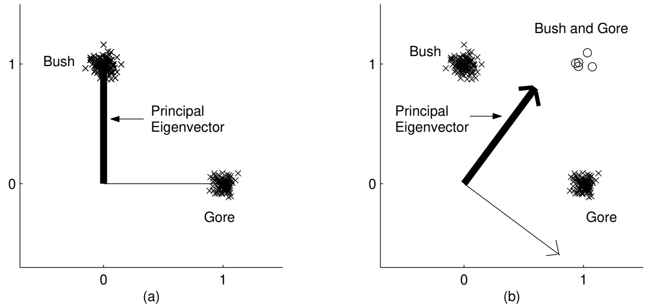

Figure 1: Jittered scatterplot of hyperlink graph.

linking to `http://www.georgewbush.com`. The adjacency matrix A has all zeros except for the two columns corresponding to these two web pages, therefore the principal eigenvector a^* will have non-zero values only for `algore.com` and `georgewbush.com`. Figure 1(a) presents a jittered scatterplot of links to these two web pages, along with the first two eigenvectors. (Only the non-zero portions of the eigenvectors are shown.) Now, suppose five new web pages trickle into our collection, which happen to link to both `algore.com` and `georgewbush.com`. Figure 1(b) shows the new plot, and we see that the eigenvectors have changed dramatically, with the principal eigenvector now near the $45°$ line. Thus, a relatively *small* perturbation to our collection has caused a *large* change to the eigenvectors.[4] If this phenomenon is pervasive, then it needs to be addressed by any algorithm that uses eigenvectors to determine authority. In the next two sections, we give characterizations of whether and when algorithms can be expected to suffer from these problems.

4.1 Analysis of HITS

HITS uses the principal eigenvector of $S = A^TA$ to determine authorities. In this section, we show that the stability of this eigenvector under small perturbations is determined by the *eigengap* of S, which is defined to be the difference between the largest and the second largest eigenvalues.

Here is an example that may shed light on the importance of the eigengap. Figure 2 plots the contours associated with two matrices S_1 and S_2 before (with solid lines) and after (with dashed lines) the same additive perturbation have been made to them.[5] The eigenvalues of the matrices are indicated by the directions of the principal axes of the ellipses. The matrix S_1 shown in Figure 2a has eigengap $\delta_1 \approx 0$, and a small perturbation to S_1 (and hence the ellipse) results in eigenvectors $45°$ away from the original eigenvectors; the matrix S_2 shown in Figure 2b has eigengap $\delta_2 = 2$, and the perturbed eigenvectors are nearly the same as the original eigenvectors. So, we see how, in this example, the size of the eigengap directly affects the stability of the eigenvectors. (Readers fa-

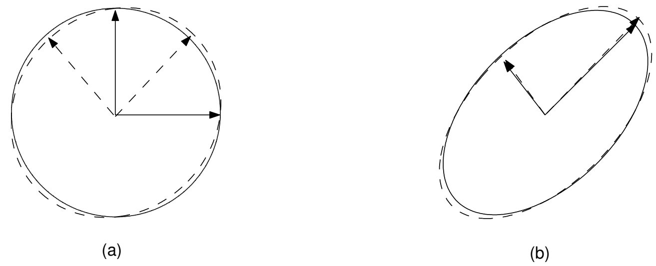

Figure 2: Contours of two matrices with different eigengaps.

miliar with plots of multivariate Gaussians can also think of these as the contours of a Gaussian with small perturbations imposed on the (inverse) covariance matrix.)

In the sequel, we use a tilde to denote perturbed quantities. (For instance, $\tilde{S}$ denotes a perturbed version of S.) We now give our first, positive result, that so long as the eigengap δ is large, then HITS is insensitive to small perturbations.[6]

Theorem 1. *Let $S = A^TA$ be given. Let a^* be the principal eigenvector and δ the eigengap of S. Assume the maximum out-degree of every web page is bounded by d. For any $\varepsilon > 0$, suppose we perturb the web/citation graph by adding or deleting at most k links from one page, where $k < (\sqrt{d+\alpha} - \sqrt{d})^2$, where $\alpha = \varepsilon\delta/(4 + \sqrt{2}\varepsilon)$. Then the perturbed principal eigenvector $\tilde{a}^*$ of the perturbed matrix $\tilde{S}$ satisfies:*

$$||a^* - \tilde{a}^*||_2 \leq \varepsilon \tag{2}$$

So, if the eigengap is big, HITS will be insensitive to small perturbations. This result is proved by showing i) the *direction* of the principal eigenvector does not change too much, and ii) the *magnitudes* of the relevant eigenvalues do not change too much, so the second eigenvector does not "overtake" the first and become the new principal eigenvector.

Proof. Let $||\cdot||_F$ denote the Frobenius norm.[7] We apply Theorem V.2.8 from matrix perturbation theory [Stewart and Sun, 1990]: Suppose $S \in \mathbb{R}^{n\times n}$ is a symmetric matrix with principal eigenvalue λ^* and eigenvector a^*, and eigengap $\delta > 0$. Let E be a symmetric perturbation to S. Then the following inequalities hold for the old principal eigenpair (λ^*, a^*) and *some* new eigenpair $(\tilde{\lambda}, \tilde{a})$.

$$||a^* - \tilde{a}||_2 \leq \frac{4||E||_F}{\delta - \sqrt{2}||E||_F} \tag{3}$$

$$|\lambda^* - \tilde{\lambda}| \leq \sqrt{2}||E||_F \tag{4}$$

(assuming that the denominator in (3) is positive). Let the complementary eigenspace to (λ^*, a^*) be represented by (L_2, X_2), i.e. X_2 is orthonormal, and its columns contain all the eigenvectors of S except a^*; L_2 is diagonal and contains the corresponding eigenvalues, all of which are at least δ less

[4] There is nothing special about the number 5 here; a smaller number also results in relatively large swings of the eigenvectors. Replacing 5 with 1, 2, 3, and 4 causes the principal eigenvector to lie at 73, 63, 58 and 55 degrees, respectively.

[5] More precisely, these are contours of the quadratic form x^TS_ix.

[6] Our analyses also apply directly to hub-weight calculations, simply by reversing link directions and interchanging A and A^T.

[7] The Frobenius norm is defined by $||X||_F = (\sum_i \sum_j (X_{ij})^2)^{1/2}$.

than λ^*; and $SX_2 = X_2 L_2$. A bound similar to Equation (4) holds for L_2:

$$||L_2 - \tilde{L}_2||_F \leq \sqrt{2}||E||_F \tag{5}$$

Let $\tilde{\lambda}_2$ be the largest eigenvalue of $\tilde{L}_2$. Using Corollary IV.3.6 from Stewart and Sun [1990], one can show that Equation (5) implies

$$\tilde{\lambda}_2 \leq \lambda_2 + \sqrt{2}||E||_F \tag{6}$$

If in turn $\sqrt{2}||E||_F < \delta/2$, then Equations (4) and (6) together will ensure that $\tilde{\lambda} > \tilde{\lambda}_2$, i.e. $(\tilde{\lambda}, \tilde{a})$ is the principal eigenpair of $\tilde{S}$.

Since we are adding or deleting links from only one page, let F denote the perturbation to one row of A, so that $\tilde{S} = (A+F)^T(A+F)$. It is straightforward to show $||F^T F||_F \leq k$ and $||A^T F||_F = ||F^T A||_F \leq \sqrt{dk}$. We can thus bound the norm of the perturbation to S:

$$||E||_F = ||\tilde{S} - S||_F \leq k + 2\sqrt{dk} \tag{7}$$

Using Equations (3) and (7) to determine when we may guarantee Equation (2) to hold, we arrive at the bound $k < (\sqrt{d+\alpha} - \sqrt{d})^2$, where $\alpha = \varepsilon\delta/(4 + \sqrt{2}\varepsilon)$. One can easily verify that the same bound on k also ensures $\sqrt{2}||E||_F < \delta/2$ (which also guarantees that the denominator in (3) is positive), hence $\tilde{a}^* = \tilde{a}$ as previously stated. □

Next we give the converse of this result, that if the eigengap is small, then eigenvectors can be sensitive to perturbations.

Theorem 2. *Suppose S is a symmetric matrix with eigengap δ. Then there exists a $O(\delta)$ perturbation[8] to S that causes a large ($\Omega(1)$) change in the principal eigenvector.*

Proof. Since $S = S^T$, it can be diagonalized:

$$S = U \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \Sigma \end{pmatrix} U^T$$

where U is orthogonal, and whose columns are the S's eigenvectors. Let u_i denote the i-th column of U. We pick $\tilde{S} = S + 2\delta u_2 u_2^T$. Since $||u_2||_2 = 1$, the norm of the perturbation is only $||2\delta u_2 u_2^T||_F = 2\delta$. Moreover,

$$\tilde{S} = U \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 + 2\delta & 0 \\ 0 & 0 & \Sigma \end{pmatrix} U^T$$

As $\tilde{\lambda}_2 = \lambda_2 + 2\delta > \lambda_1$, $(\tilde{\lambda}_2, u_2)$ is the new principal eigenpair. But u_2 is orthogonal to u_1, so $||u_2 - u_1||_2 = \Omega(1)$. □

To ground these results and illustrate why Theorem 1 requires a bound d on out-degrees, we give another example of where a small perturbation—adding a single link—can have a large effect. In this example we use the fact that if a graph has multiple connected components, then the principal eigenvalue will have non-zero entries in nodes only from the "largest" connected component (more formally the component with the largest eigenvalue).[9]

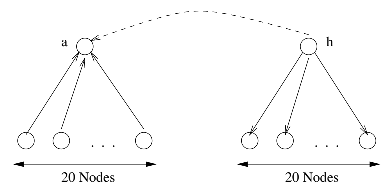

Figure 3: Picture of a web community.

Consider the web/citation-graph shown in Figure 3, which we imagine to be a small subset of a much larger graph. Solid arrows denote the original set of hyperlinks; the dashed arrow represents the link we will add. The original principal eigenvalue for each of the two connected components shown is $\lambda = 20$; with the addition of a single link, it is easy to verify that this jumps to $\tilde{\lambda} = 25$. Suppose that the community shown is part of a larger web/citation graph with multiple subcommunities, and that originally the biggest subcommunity had eigenvalue $20 < \lambda_1 < 25$. By adding one link, the graph shown in Figure 3 becomes the biggest subcommunity, and the principal eigenvector now has positive values only for nodes shown in this figure, and zeros elsewhere.

4.2 Analysis of PageRank

We now analyze the sensitivity of PageRank's authority scores p to perturbations of the web/citation-graph.

Theorem 3. *Let M be given, and let p be the principal right eigenvector of $(\epsilon U + (1-\epsilon)M)^T$. Let articles/pages $i_1, i_2, \ldots, i_k$ be changed in any way, and $\tilde{M}$ be the corresponding (new) transition matrix. Then the new PageRank scores $\tilde{p}$ satisfies:*

$$||\tilde{p} - p||_1 \leq \frac{2\sum_{j=1}^{k} p_{i_j}}{\epsilon} \tag{8}$$

Thus, assuming ϵ is not too close to 0, this shows that so long as the perturbed/modified web pages did not have high overall PageRank scores (as measured with respect to the *unperturbed* PageRank scores p), then the perturbed PageRank scores $\tilde{p}$ will not be far from the original.

Proof. We construct a coupled Markov chain $\{(X_t, Y_t) : t \geq 0\}$ over pairs of web pages/documents as follows. $X_0 = Y_0$ is drawn according to the probability vector p, that is, from the stationary distribution of the PageRank "random surfer" model. The state transitions work as follows: On step t, we decide with probability ϵ to "reset" both chains, in which case we set X_t and Y_t to the *same* page chosen uniformly at random from the collection. If no "reset" occurs, and if $X_{t-1} = Y_{t-1}$ and X_{t-1} is one of the unperturbed pages, then $X_t = Y_t$ is chosen to be a random page linked to by the page X_{t-1}. In all other cases, X_t is chosen to be a random page linked to by page X_{t-1}, and independently of it, Y_t is chosen to be a random page linked to by page Y_{t-1}.

[8] More formally, there exists a perturbed version of S, denoted $\tilde{S}$, so that $||S - \tilde{S}||_F = O(\delta)$.

[9] See, e.g. Chung [1994]. A connected component of a graph is a subset whose elements are connected via length ≥ 1 paths to each other, but not to the rest of the graph. The eigenvalue of a connected component C is the largest eigenvalue of $A_C^T A_C$ (cf. $A^T A$ used by HITS), where A_C, a submatrix of A, is the adjacency matrix of C.

Thus, we now have two "coupled" Markov chains X_t and Y_t, the former using the transition probabilities $(\epsilon U + (1-\epsilon)M)^T$, and latter $(\epsilon U + (1-\epsilon)\tilde{M})^T$, but so that their transitions are "correlated." For instance, the "resets" to both chains always occur in lock-step. But since each chain is following its own state transition distribution, the asymptotic distributions of X_t and Y_t must respectively be p and $\tilde{p}$. Now, let $d_t = P(X_t \neq Y_t)$. Note $d_0 = 0$, since $X_0 = Y_0$ always. Letting $\mathcal{P}$ denote the set of perturbed pages, we have:

$$
\begin{aligned}
d_{t+1} &= P(X_{t+1} \neq Y_{t+1}) \\
&= P(X_{t+1} \neq Y_{t+1} | \text{reset at } t+1) P(\text{reset}) \\
&\quad + P(X_{t+1} \neq Y_{t+1} | \text{no reset at } t+1) P(\text{no reset}) \\
&= 0 \cdot \epsilon + (1-\epsilon) P(X_{t+1} \neq Y_{t+1} | \text{no reset at } t+1) \\
&= (1-\epsilon)[P(X_{t+1} \neq Y_{t+1}, X_t \neq Y_t | \text{no reset at } t+1) \\
&\quad + P(X_{t+1} \neq Y_{t+1}, X_t = Y_t | \text{no reset at } t+1)] \\
&\leq (1-\epsilon)[P(X_t \neq Y_t | \text{no reset at } t+1) \\
&\quad + P(X_{t+1} \neq Y_{t+1}, X_t = Y_t, X_t \in \mathcal{P} | \text{no reset at } t+1)] \\
&\leq (1-\epsilon)(P(X_t \neq Y_t) + P(X_t \in \mathcal{P} | \text{no reset at } t+1)) \\
&\leq (1-\epsilon)(d_t + \textstyle\sum_{i \in \mathcal{P}} p_i)
\end{aligned}
$$

where to derive the first inequality, we used the fact that by construction, the event "$X_{t+1} \neq Y_{t+1}, X_t = Y_t$" is possible only if X_t is one of the perturbed pages. Using the fact that $d_0 = 0$ and by iterating this bound on d_{t+1} in terms of d_t, we obtain an asymptotic upper-bound: $d_\infty \leq (\sum_{i \in \mathcal{P}} p_i)/\epsilon$. Thus, if (X_∞, Y_∞) is drawn from the stationary distribution of the correlated chains—so the marginal distributions of X_∞ and Y_∞ are respectively given by p and $\tilde{p}$—then $P(X_\infty \neq Y_\infty) = d_\infty \leq (\sum_{i \in \mathcal{P}} p_i)/\epsilon$. But if two random variables have only a small d_∞ chance of taking different values, then their distributions must be similar. More precisely, by the Coupling Lemma (e.g., see Aldous, 1983) the *variational distance* $(1/2)\sum_i |p_i - \tilde{p}_i|$ between the distributions must also be bounded by the same quantity d_∞. This shows $||p - \tilde{p}||_1 \leq 2d_\infty$, which concludes the proof. □

5 LSI and HITS

In this section we present an interesting connection between HITS and Latent Semantic Indexing [Deerwester *et al.*, 1990] (LSI) that provides additional insight into our stability results (see also Cohn and Chang, 2000). In LSI a collection of documents is represented as a matrix A, where A_{ij} is 1 if document j contains the i-th word of the vocabulary, and 0 otherwise. LSI computes the left and right singular vectors of A (equivalently, the eigenvectors of AA^T and A^TA). For example, the principal left singular vector, which we denote x, has dimension equal to the vocabulary size, and x_j measures the "strength" of word j's membership along the x-dimension. The informal hope is that synonyms will be grouped into the same singular vectors, so that when a document (represented by a column of A) is projected onto the subspace spanned by the singular vectors, it will automatically be "expanded" to include synonyms of words in the document, leading to improved information retrieval.

Now consider constructing the following citation graph from a set of documents. Let there be a node for each document and for each word. The node of a word links to the document nodes it appears in. Let $\hat{A}$ be the adjacency matrix of this graph. If we apply HITS to this graph, we find only the word-nodes have non-zero hub weights (since none of the document-nodes link to anything) and only the document-nodes have non-zero authority weights. Moreover, the vector of HITS hub weights of the word-nodes is exactly x, the first left singular vector found by LSI.

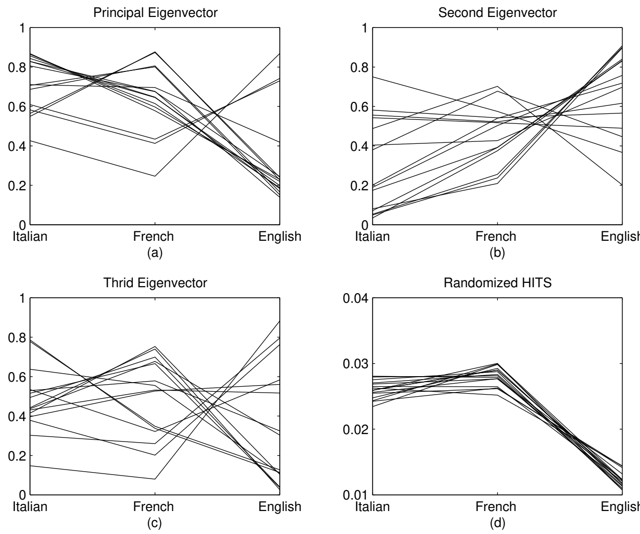

Figure 4: Results on random corpora.

This connection allows us to transfer insight from experiments on LSI to our understanding of HITS. In this vein, we conducted an experiment in which random corpora were generated by sampling from a set of English, French, and Italian documents.[10] Given that these random corpora are combinations of three distinct languages, the solution to Information Retrieval problems such as clustering or synonym-identification are exceedingly simple. The issue that we are interested in, however, is stability. To study stability, we generated 15 such collections and examined the direction of the principal eigenvectors found by HITS.

The principal eigenvector lies in the high dimensional joint-vocabulary space of the three languages. To display our results, we therefore defined English, French, and Italian "directions," and measured the degree to which the eigenvector lies in each these directions.[11] Fifteen independent repetitions of this process were carried out, and the results plotted in Figure 4a. As we see, despite the presence of clear clusters in the corpora, the eigenvectors are highly variable. Moreover, this variability persists in the second and third eigenvectors (Figures 4b,c).

[10] The corpora were generated by taking paragraphs from novels in the three languages. Typical "documents" had 25–150 words, and the vocabulary consisted of the most common 1500 words per language. The collection was also manually "balanced" to equally represent each language.

[11] This was done by picking a vector x_e of unit-norm and whose i-th element is proportional to the frequency of word i in the English collection—thus, x_e should be thought of as the "canonical" English direction—and taking the amount that h^* lies in the English direction to be the absolute magnitude of the dot-product between x_e and h^*, and similarly for French and Italian.

Note that the variability is not an inherent feature of the problem. In Figure 4d, we display a run of a different algorithm (a variant of the HITS algorithm that we briefly describe in Section 7, and is studied in more detail in [Ng *et al.*, 2001]). Here the results are significantly less variable.

6 Further Experiments

In this section we report further results of perturbation experiments on the *Cora* database. We also describe an experiment using web pages.

Recall our methodology in the experiments with the *Cora* database: We choose a subset of papers from the database and generate a set of perturbations to this subset by randomly deleting 30% of the papers. Our first experiment used all of the AI papers in Cora as the base set. Our results largely replicated those of Cohn and Chang [2000]—HITS returned several Genetics Algorithms (GA) papers as the top-ranked ones. With the database perturbed as described, however, these results were very variable, and HITS often returned seminal papers from broader AI areas as its top-ranked documents. Repeating the experiment excluding all the GA papers, HITS did slightly better; the results on five independent trials are shown below:

1	"Classification and Regression Trees", Brieman+al	1	1	1	1	1
2	"Pattern classification and scene analysis", Duda+Hart	2	2	3	2	2
3	"UCI repository of machine learning databases", Murphy+Aha	4	3	7	3	3
4	"Learning internal representations by error...", Rumelhart+al	3	13	2	28	20
5	"Irrelevant Features and the Subset Selection Problem", John+al	7	4	12	4	4
6	"Very simple classification rules perform well on...", Holte	8	5	15	5	5
7	"C4.5: Programs for Machine Learning", Quinlan	11	10	14	10	6
8	"Probabilistic Reasoning in Intelligent Systems", Pearl	6	459	4	462	461
9	"The CN2 induction algorithm", Clark+Niblett	9	54	11	78	105
10	"Learning Boolean Concepts in the ...", Almuallim+Dietterich	14	11	34	9	13
11	"The MONK's problems: A performance comparison...", Thrun	-	9	-	6	7
12	"Inferring decision trees using the MDL Principle", Quinlan	-	8	-	7	8
13	"Multi-interval discretization of continuous...", Fayyad+Irani	-	-	-	-	10
14	"Learning Relations by Pathfinding", Richards+Moon	-	6	-	-	-
15	"A conservation law for generalization performance", Schaffer	-	7	-	8	-
20	"The Feature Selection Problem: Traditional..." Kira+Randall	-	-	-	-	9
21	"Maximum likelihood from incomplete data via..." Dempster+al	10	-	5	-	-
23	"Learning to Predict by the Method of Temporal...", Sutton	5	-	6	-	-
36	"Introduction to the Theory of Neural Computation", Hertz+al	-	-	8	-	-
49	"Explanation-based generalization: a unifying view", Mitchell	-	-	10	-	-
282	"A robust layered control system for a mobile robot", Brooks	-	-	9	-	-

We see that, apart from the top 2-3 ranked papers, the remaining results are still rather unstable. For example, Pearl's book was originally ranked 8th; on the second trial, it dropped to rank 459. Similarly, Brooks' paper was rank 282, and jumped up to rank 9 on trial 3. However, this variability is not intrinsic to the problem, as shown by our PageRank results (all PageRank results in this section were generated with $\epsilon = 0.2$):

1	"Classification and Regression Trees", Breiman+al	1	1	1	1	2
2	"Probabilistic Reasoning in Intelligent Systems", Pearl	3	2	2	2	1
3	"Learning internal representations by error...", Rumelhart+al	2	3	3	3	3
4	"Pattern classification and scene analysis", Duda+Hart	4	4	4	4	4
5	"A robust layered control system for a mobile robot", Brooks	5	6	7	5	5
6	"Maximum likelihood from incomplete data via...' Dempster+al	6	7	6	6	6
7	"Learning to Predict by the Method of Temporal...", Sutton	7	5	5	7	7
8	"UCI repository of machine learning databases", Murphy+Aha	8	9	9	9	11
9	"Numerical Recipes in C", Press+al	10	12	8	11	8
10	"Parallel Distributed Processing", Rumelhart+al	9	14	13	10	9
12	"An implementation of a theory of activity", Agre+Chapmanre	-	8	10	8	-
13	"Introduction to the Theory of Neural Computation", Hertz+al	-	10	-	-	-
22	"A Representation and Library for Objectives in...", Valente+al	-	-	-	-	10

The largest change in a document's rank was a drop from 10 to 12—these results are much more stable than for HITS. Closer examination of the HITS authority weights reviews that its jumps in rankings are indeed due to large changes in authority weights, whereas the PageRank scores tended to remain fairly stable.[12]

We also carried out experiments on web pages. Given a query, Kleinberg [1998] describes a method for obtaining a collection of web pages on which to run HITS. We use exactly the method described there, and perturbed it in a natural way.[13] For the sake of brevity, we only give the results of two experiments here. On the query "mp3 players", HITS' results were as follows (long URLs are truncated):

1	http://www.freecode.com/	82	1	1	1	82
2	http://www.htmlworks.com/	85	2	2	2	83
3	http://www.internettrafficreport.com/	86	3	4	3	85
4	http://slashdot.org/	88	4	5	5	86
5	http://windows.davecentral.com/	87	5	3	4	84
6	http://www.gifworks.com/	84	6	6	6	87
7	http://www.thinkgeek.com/	91	7	7	7	88
8	http://www.animfactory.com/	89	9	8	8	89
9	http://freshmeat.net/	90	8	9	9	90
10	http://subscribe.andover.net/membership.htm	92	10	10	10	91
1385	http://ourstory.about.com/index.htm	1	-	-	-	1
1386	http://home.about.com/index.htm	2	-	-	-	2
1387	http://home.about.com/musicperform/index.htm	3	-	-	-	3
1388	http://home.about.com/teens/index.htm	4	-	-	-	4
1389	http://home.about.com/sports/index.htm	5	-	-	-	5
1390	http://home.about.com/autos/index.htm	6	-	-	-	6
1391	http://home.about.com/style/index.htm	7	-	-	-	7
1392	http://home.about.com/careers/index.htm	8	-	-	-	8
1393	http://home.about.com/citiestowns/index.htm	9	-	-	-	9
1394	http://home.about.com/travel/index.htm	10	-	-	-	10

In contrast, PageRank returned:

1	http://www.team-mp3.com/	*	1	1	1	1
2	http://click.linksynergy.com/fs-bin/click	1	3	2	4	9
3	http://www.elizandra.com/	2	2	3	2	2
4	http://stores.yahoo.com/help.html	4	14	5	10	11
5	http://shopping.yahoo.com/	3	10	4	12	13
6	http://www.netins.net/showcase/phdss/	*	8	6	3	3
7	http://www.thecounter.com/	13	6	9	8	7
8	http://ourstory.about.com/index.htm	5	4	7	5	4
9	http://a-zlist.about.com/index.htm	6	5	10	6	6
10	http://www.netins.net/showcase/phdss/getm	*	9	8	7	5
11	http://software.mp3.com/software/	7	7	-	-	8
12	http://www.winamp.com/	8	-	-	-	-
13	http://www.nullsoft.com/	10	-	-	-	-
14	http://www.consumerspot.com/redirect/1cac	9	-	-	9	10

While PageRank's rankings undergo small changes, HITS' rankings display a mass "flipping" behavior. Similar perturbation patterns to this (and the example below) for PageRank and HITS are observed in fourteen out of nineteen queries. Furthermore, HITS' results displayed such mass "flips" in roughly 20% of the trials, which is in accordance with the 20% removal rate.

Here is another typical web result, this time on the query "italian recipes." Note that "*" means that the page was removed by that trial's perturbation, and therefore has no rank. HITS' results were:

[12] Examination of the second and higher eigenvectors in HITS shows that they, too, can vary substantially from trial to trial.

[13] Kleinberg [1998] first uses a web search engine (www.altavista.com in our case) to retrieve 200 documents to form a "root set," which is then expanded (and further processed) to define the web-graph on which HITS operates. Our perturbations were arrived at by randomly deleting 20% of the root set (i.e. imagining that the web search engine had only returned 80% of the pages it actually did), and then following Kleinberg's procedure.

1	http://ourstory.about.com/index.htm	*	1	1	1	1
2	http://home.about.com/culture/index.htm	*	2	2	2	17
3	http://home.about.com/index.htm	*	3	3	3	25
4	http://home.about.com/food/index.htm	*	4	4	4	2
5	http://home.about.com/science/index.htm	*	5	5	5	3
6	http://home.about.com/shopping/index.htm	*	6	6	6	4
7	http://home.about.com/smallbusiness/index	*	7	7	7	5
8	http://home.about.com/sports/index.htm	*	8	8	8	6
9	http://home.about.com/arts/index.htm	*	9	9	9	7
10	http://home.about.com/style/index.htm	*	10	10	10	8
11	http://home.about.com/autos/index.htm	-	-	-	-	9
12	http://home.about.com/teens/index.htm	-	-	-	-	10
479	http://bestbrandrecipe.com/default.asp	1	-	-	-	-
480	http://myrecipe.com/help/shopping.asp	2	-	-	-	-
481	http://vegetarianrecipe.com/default.asp	3	-	-	-	-
482	http://holidayrecipe.com/default.asp	5	-	-	-	-
483	http://beefrecipe.com/default.asp	4	-	-	-	-
484	http://beveragerecipe.com/default.asp	7	-	-	-	-
485	http://appetizerrecipe.com/default.asp	6	-	-	-	-
486	http://pierecipe.com/default.asp	8	-	-	-	-
487	http://seafoodrecipe.com/default.asp	9	-	-	-	-
488	http://barbequerecipe.com/default.asp	10	-	-	-	-

PageRank, on the other hand, returned:

1	http://ourstory.about.com/index.htm	*	1	1	1	1
2	http://a-zlist.about.com/index.htm	*	2	2	2	2
3	http://www.apple.com/	1	3	3	3	3
4	http://www.tznet.com/isenberg/	2	4	4	13	9
5	http://frontier.userland.com/	3	5	5	4	7
6	http://www.mikrostore.com/	4	6	6	5	*
7	http://www.amazinggiftsonline.com/	5	7	7	6	*
8	http://www.peck.it/peckshop/home.asp?prov	*	8	8	7	4
9	http://geocities.yahoo.com/addons/interac	6	9	9	8	29
10	http://dvs.dolcevita.com/index.html	7	10	*	10	5
11	http://www.dossier.net/	-	-	10	9	6
12	http://www.dolcevita.com/	8	-	-	-	8
14	http://www.q-d.com/	9	-	-	-	10
15	http://www.silesky.com/	10	-	-	-	-

7 Discussion

It is well known in the numerical linear algebra community that a subspace spanned by several (e.g. the first k) eigenvectors may be stable under perturbation, while individual eigenvectors may not [Stewart and Sun, 1990]. Our results—both theoretical and empirical—reflect this general fact.

If the output of an algorithm is a subspace, then the stability considerations that we have discussed may not be a matter of primary concern. Such is the case, for example, for the LSI algorithm, where the goal is generally to project a data set onto a lower-dimensional subspace.

If we wish to interpret specific eigenvectors, however, then the stability issue becomes a matter of more serious concern. This is the situation for the basic HITS algorithm, where primary eigenvectors have been interpreted in terms of a set of "hubs" and "authorities." As we have seen, there are theoretical and empirical reasons for exercising considerable caution in making such interpretations.

Given that the principal eigenvector may not have a reliable interpretation, one can consider variations of the HITS approach that utilize multiple eigenvectors. Indeed, Kleinberg [1998] suggested examining multiple eigenvectors as a way of obtaining authorities within multiple communities. Again, however, it may be problematic to interpret individual eigenvectors, and in fact in our experiments we found significant variability in second and third eigenvectors. An alternative approach may be to automatically combine multiple eigenvectors in a way that explicitly identifies subspaces within the HITS framework. This is explored in [Ng *et al.*, 2001]

The fact that the PageRank algorithm appears to be relatively immune to stability concerns is a matter of considerable interest. It is our belief that the "reset-to-uniform-distribution" aspect of PageRank is a critical feature in this regard. Indeed, one can explore a variation of the HITS algorithm which incorporates such a feature. Suppose that we construct a Markov chain on the web in which, with probability $1-\epsilon$, we randomly follow a hyperlink from the current page in the forward direction (on odd time steps), and we randomly follow a hyperlink in the backwards direction (on even time steps). With probability ϵ, we reset to a uniformly chosen page. The asymptotic web-page visitation distribution on odd steps is defined to be the authority weights, and on even steps the hub weights. As in Theorem 3, we can show this algorithm is insensitive to small perturbations (but unlike PageRank, we obtain hub as well as authority scores). The results of running this algorithm on the "three languages" problem are shown in Figure 4d, where we see that it is indeed significantly more stable than the basic HITS algorithm. This algorithm is also explored in more detail in [Ng *et al.*, 2001].

Acknowledgments

We thank Andrew McCallum for providing the Cora citation data used in our experiments. We also thank Andy Zimdars for helpful comments. This work was supported by ONR MURI N00014-00-1-0637 and NSF grant IIS-9988642.

References

[Aldous, 1983] David Aldous. Random walks on finite groups and rapidly mixing markov chains. In A. Dold and B. Eckmann, editors, *Séminaire de Probabilités XVII 1981/1982*, Lecture Notes in Mathematics, Vol. 986, pages 243–297. Springer-Verlag, 1983.

[Bharat and Henzinger, 1998] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. 21st Annual Intl. ACM SIGIR Conference*, pages 104–111. ACM, 1998.

[Brin and Page, 1998] S. Brin and L. Page. The anatomy of a large-scale hypertextual (Web) search engine. In *The Seventh International World Wide Web Conference*, 1998.

[Chung, 1994] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1994.

[Cohn and Chang, 2000] D. Cohn and H. Chang. Probabilistically identifying authoritative documents. In *Proc. 17th International Conference on Machine Learning*, 2000.

[Deerwester *et al.*, 1990] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[Golub and Van Loan, 1996] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Univ. Press, 1996.

[Kleinberg, 1998] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[McCallum *et al.*, 2000] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the contruction of Internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000.

[Ng *et al.*, 2001] Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. Stable algorithms for link analysis. In *Proc. 24th Annual Intl. ACM SIGIR Conference*. ACM, 2001.

[Osareh, 1996] Farideh Osareh. Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri*, 46:149–158, 1996.

[Page *et al.*, 1998] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Unpublished Manuscript, 1998.

[Stewart and Sun, 1990] G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.