Stable Algorithms for Link Analysis

Andrew Y. Ng
Computer Science Division
U.C. Berkeley
Berkeley, CA 94720
ang@cs.berkeley.edu

Alice X. Zheng
Computer Science Division
U.C. Berkeley
Berkeley, CA 94720
alicez@cs.berkeley.edu

Michael I. Jordan
CS Div. & Dept. of Stat.
U.C. Berkeley
Berkeley, CA 94720
jordan@cs.berkeley.edu

ABSTRACT

The Kleinberg HITS and the Google PageRank algorithms are eigenvector methods for identifying "authoritative" or "influential" articles, given hyperlink or citation information. That such algorithms should give reliable or consistent answers is surely a desideratum, and in [10], we analyzed when they can be expected to give stable rankings under small perturbations to the linkage patterns. In this paper, we extend the analysis and show how it gives insight into ways of designing stable link analysis methods. This in turn motivates two new algorithms, whose performance we study empirically using citation data and web hyperlink data.

1. INTRODUCTION

From its origins in bibliometric analysis [11], the analysis of cross-referencing patterns—"link analysis"—has come to play an important role in modern information retrieval. Link analysis algorithms have been successfully applied to web hyperlink data to identify authoritative information sources, and to academic citation data to identify influential papers [8, 3]. In particular, together with classical IR ranking techniques, link analysis provides the basis for some of today's Internet search engines.

An important feature of collections such as the World Wide Web is their dynamic nature. References can be changed, become inaccessible, or be missed by a search engine. If link analysis is to provide a robust notion of authoritativeness in such a setting, it is natural to ask that it also be robust in the sense of being *stable* to perturbations of the link structure. Indeed, it seems unlikely that a highly unstable search engine—say one that completely changes its results from day to day—would be trusted by its users to be always returning all the relevant articles. In the setting of academic citations, stability also means (for example) that a few authors writing a relatively small number of papers should rarely cause us to completely change our minds about what articles in a community had been seminal. This issue of stability seems to have received little attention in the link analysis literature, and is the principal focus of our paper.

Two popular algorithms, in particular the Kleinberg HITS algorithm [8] and the Google PageRank algorithm [3], are eigenvector-based methods; they essentially compute principal eigenvectors of particular matrices related to the adjacency graph to determine "authority." Understanding the robustness of link analysis algorithms therefore involves an analysis of the stability of these eigenvector calculations.

Using ideas from matrix perturbation theory and Markov chain theory, in [10] we formally characterized conditions under which HITS and PageRank are stable. In this paper, we briefly summarize the results derived in [10], and show how they give insight into ways of designing stable link analysis algorithms. This then motivates two new algorithms: Randomized HITS, which merges the hubs-and-authorities notion from HITS with a stabilizing "reset" mechanism from PageRank (see also [14]); and Subspace HITS, which provides a principled way of combining multiple eigenvectors from HITS to yield aggregate authority scores. These new algorithms are also demonstrated empirically to produce good results on both academic citation and web query data.

We also explore the issue of the "diversity" of the results returned by these algorithms. This leads into a discussion of the relationship between Latent Semantic Indexing (LSI) [6] and HITS.

2. AN EXAMPLE

We begin with an example. The *Cora* database [9] is a collection containing citation information from several thousand academic papers in various areas of computer science. We ran the HITS and PageRank algorithms on the subset of the *Cora* database consisting of all its Machine Learning papers, and examined the list of papers that they considered "authoritative." To evaluate the stability of the algorithms, we also constructed five perturbed versions of the databases, each of which contained a randomly selected 70% subset of the papers. ("Since *Cora* obtained its database via a web crawl, what if, by chance or mishap, it had only retrieved 70% of these papers?") If a paper is truly authoritative, we might hope that it is still identifiable as such with only a subset of the citation data.

The results from HITS are shown below. The leftmost column is the HITS authority ranking obtained by analyzing the full set of Machine Learning papers; the five rightmost columns report the ranks in runs on the perturbed databases. We see substantial variation across the different runs:

HITS results on Cora ML papers:

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

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA..

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	-	-	-

One might think that this variability is intrinsic to the problem and hence unavoidable, but this is not the case, as shown by the results from the PageRank algorithm, which are much more stable:

PageRank results on Cora ML papers ($\epsilon = 0.2$):

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 stable. In fact, in certain web query experiments, PageRank displays its own perturbation pattern. The issue to which we direct our attention is more subtle. In order to understand the stability of either algorithm, we need to consider issues such as the relationships between multiple eigenvectors and invariant subspaces, and the effects of a universal resetting probability. Stability is certainly an important desideratum in algorithms that identify authoritative or relevant articles, hence these issues will play an important role in the two new algorithms that we will present in Section 5. The following are results on the same data using the new algorithms.

Randomized HITS results on Cora ML papers ($\epsilon = 0.2$):

1	“Genetic Algorithms in Search, Optimization...”, Goldberg	1	1	1	1	1
2	“Learning internal representations by error...”, Rumelhart+al	2	2	2	2	2
3	“Probabilistic Reasoning in Intelligent Systems”, Pearl	3	3	3	3	3
4	“Adaptation in Natural and Artificial Systems”, Holland	4	4	5	4	4
5	“Classification and Regression Trees”, Breiman+al	5	5	6	6	5
6	“Genetic Programming: On the Programming of...”, John+al	6	6	4	5	6
7	“Pattern classification and scene analysis”, Duda+Hart	8	7	7	8	10
8	“Maximum likelihood from incomplete data via...”, Dempster+al	7	8	8	9	7
9	“Learning to Predict by the Method of Temporal...”, Sutton	9	9	9	7	8
10	“Introduction to the theory of neural computation”, Hertz+al	10	10	10	10	9

Subspace HITS results on Cora ML papers ($k = 20$, $f(\lambda) = \lambda^2$):

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

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 constructs 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 [8] 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:

$$a^{(t+1)} = A^T h^{(t)} = (A^T A) a^{(t)} \quad (1)$$

$$h^{(t+1)} = A a^{(t+1)} = (A A^T) h^{(t)}. \quad (2)$$

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 [7], and so (under mild conditions) a^* and h^* are the principal eigenvectors of $A^T A$ and AA^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 [3] 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 right 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

$$(\epsilon U + (1 - \epsilon)M)^T p = p. \quad (3)$$

[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, (e.g. [1] reports on results for both algorithms in the former setting,) and this distinction does not affect the outcome of our analysis.

[2][8] discusses several other heuristics regarding issues such as intra-domain references, which are ignored here for simplicity (but are used in our experiments). See also Bharat and Henzinger [2] 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 [12] 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.

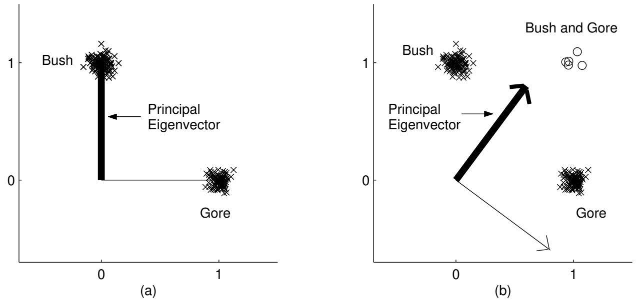

Figure 1: Jittered scatterplot of hyperlink graph.

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 pages linking to `http://www.algore.com/`, and another 103 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.

4.1 Stability of HITS

It is possible to give fairly precise characterizations of when HITS will be sensitive to small perturbations. HITS uses the principal eigenvector of $S = A^T A$ to determine authorities. It turns out that the algorithm's stability to small perturbations is determined by the *eigengap* of S, which is defined to be the difference between the biggest and the second biggest eigenvalues. (Recall that, if $A^T Ax = \lambda x$, then x is an eigenvector of the matrix $A^T A$, and λ the corresponding eigenvalue.) In the sequel, we use a tilde to denote perturbed quantities. (For instance, $\tilde{S}$ denotes a perturbed version of S.) Our stability results are summarized in the following two theorems:

THEOREM 1. *Let $S = A^T A$ 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:*

$$||a^* - \tilde{a}^*||_2 \leq \varepsilon \tag{4}$$

[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.

THEOREM 2. *Suppose S is a symmetric matrix with eigengap δ. Then there exists a $O(\delta)$ perturbation[5] to S that causes a large ($\Omega(1)$) change in the principal eigenvector.*

Proofs of these Theorems are given in [10]. So, if the eigengap is big, HITS will be insensitive to small perturbations. If it is small then there may be a small perturbation that can dramatically change its results. Specifically, if the eigengap is small, then a small perturbation may cause the principal eigenvector and the secondary eigenvectors to swap places. This causes the "flipping" phenomenon discussed in Section 6.

4.2 Stability of PageRank

For PageRank, we have the following stability result:

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:*

$$||\tilde{p} - p||_1 \leq \frac{2\sum_{j=1}^{k} p_{i_j}}{\epsilon} \tag{5}$$

Thus, assuming ϵ is not too close to 0, this means that if 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.

The full proof of Theorem 3 is given in [10]. But since the intuitions from PageRank will shortly be used to design changes to HITS to improve its stability, we sketch the proof of Theorem 3 in the remainder of this section. (This may be safely skipped on a first reading.)

Imagine the PageRank random surfer starting from a randomly chosen web page, and following hyperlinks at random to take a random walk of $T = 1/\epsilon$ steps on the graph. We take the chance of a web page being visited on a typical step under this random walk procedure to be the PageRank score of that page.[6] Now, suppose that someone changes pages $i_1, i_2, \ldots, i_k$; in particular, these k pages may now link to completely different pages than they had previously. What is the chance that our random surfer will even notice that these pages have been changed? When taking a random walk on the new graph, some fraction of the time the user will visit some of these perturbed pages, and some fraction of the time the random walk will only visit the unperturbed pages. If we had been taking the random walk on the original (unperturbed) graph, then on a typical step, the chance of visiting the perturbed page i_j is p_{i_j}—this was, after all, the definition of a page's PageRank score. So the chance of visiting *any* of the k perturbed pages on a typical time step is at most $\sum_{j=1}^{k} p_{i_j}$. But the user is taking T steps altogether, so the chance of visiting any of the perturbed pages at some point on the walk is at most $T\sum_{j=1}^{k} p_{i_j} = \sum_{j=1}^{k} p_{i_j}/\epsilon$.

This means that, with chance $1 - \sum_{j=1}^{k} p_{i_j}/\epsilon$, a random T-step walk taken in the original (unperturbed) graph "would have been exactly the same" random walk as if we had instead run it on the perturbed graph. More precisely, the distribution over T-step random walks in both graphs differ from each other by at most about $2\sum_{j=1}^{k} p_{i_j}/\epsilon$ (in "variational distance"), and hence the PageRank

[5] More formally, there exists a perturbed version of S, denoted $\tilde{S}$, so that $||S - \tilde{S}||_F = O(\delta)$, where $||\cdot||_F$ is the Frobenius norm.

[6] Under the "real" PageRank algorithm, the chance of the user "resetting" or "quitting" is ϵ on each step, which means that number of steps until the next reset follows a geometric distribution with mean $1/\epsilon$. To simplify our discussion, we have imagined here that the reset occurs after exactly $T = 1/\epsilon$ steps.

scores (which are just the distributions induced by these walks on web pages) in both graphs must also be similar.

The formal version of this proof uses uses a coupled Markov chains argument, and is given in [10]. But to summarize, so long as the perturbed pages have small total PageRank score, then the chance of them being visited and hence affecting an T-step random walk is small and only on the order $\sum_{j=1}^{k} p_{i_j}/\epsilon$. Note that as the reset probability ϵ is decreased, the length T of the random walks becomes large, and the chance of visiting the perturbed pages increases again. Thus, we should expect PageRank with small ϵ to become more sensitive to perturbations; this is a hypothesis that will later be explored in our experiments.

5. TWO NEW ALGORITHMS

5.1 Randomized HITS

The preceding discussion suggests a natural way of designing a random-walk based algorithm that is similar in spirit to HITS (and finds both hubs as well as authorities), and that, like PageRank, is stable to small perturbations. It is as follows:

Let there be a random surfer who is able to follow hyperlinks in both the forward and in the backward directions. More precisely, the surfer starts from a randomly chosen page, and visits a new web page at every time step. Every time step, he tosses a coin with bias ϵ, and if the coin lands heads, he jumps to a new web page chosen uniformly at random. If the coin lands tails, then he checks if it is an odd time step or an even time step. If it is an odd time step, then he follows a randomly chosen out-link from the current page; if it is an even time step, then he traverses a random in-link of the current page. Thus, the random surfer alternately follows links in the forwards and in the backwards directions, and occasionally "resets" and jumps to a page chosen uniformly at random.

This process defines a random walk on web pages, and the stationary distribution on odd time steps is defined to be the authority weights. (Informally, let t be a very large odd number, chosen large enough that the random walk has converged to its stationary distribution by t steps. The authority weight of a page is the chance that the surfer visits that page on time step t.) Similarly, the stationary distribution on even time steps is defined to be the hub weights. Mathematically, these quantities can also be written:

$$\begin{aligned} a^{(t+1)} &= \epsilon\vec{1} + (1-\epsilon)A_{\text{row}}^T h^{(t)} \\ h^{(t+1)} &= \epsilon\vec{1} + (1-\epsilon)A_{\text{col}} a^{(t+1)} \end{aligned}$$

where $\vec{1}$ is the vector of all ones, A_{row} is the same as A with its rows normalized to sum to 1, and A_{col} is A with its columns normalized to sum to 1. Note the similarity of these to the original HITS update rules (Equations 1 and 2). It is straightforward to show that iterating these equations will cause $a^{(t)}$ and $h^{(t)}$ to converge to the odd-step and the even-step stationary distributions.[7] We refer to this method as "Randomized HITS." By analogy to the proof of the stability of the PageRank algorithm [10] reviewed in the previous section, it is straightforward to establish related conditions under which Randomized HITS is insensitive to small perturbations.

5.2 Subspace HITS

There is a second way of improving the stability of HITS. Sometimes, individual eigenvectors of a matrix may not be stable, but *subspaces* spanned by eigenvectors may be. For example, it is possible that the subspace spanned by (say) the first two eigenvectors can be stable, even though the two eigenvectors may rotate freely within this subspace. (Recall, for instance, the example in Figure 1; under our perturbation, our the first two eigenvectors changed significantly, but the subspace they span was not changed at all.)

More generally, if the eigengap between the k-th and $k+1$-st eigenvalues is large, then the subspace spanned by the first k eigenvectors will be stable [15]. Thus, one might consider refraining from examining individual eigenvectors, but instead treating them as a *basis* for a subspace to obtain authority scores.[8] But more generally, we may even want to allow the case of $k = n$—so that we use all the eigenvectors—but *weight* them appropriately, so that the ones corresponding to the larger eigenvalues are given more importance. Consider the following procedure for calculating authority scores, where $f(\cdot)$ is a non-negative, monotonically increasing function that we will specify later:

1. Find the first k eigenvectors $x_1, \ldots, x_k$ of $S = A^T A$ (or AA^T for hub weights), and their corresponding eigenvalues $\lambda_1, \ldots, \lambda_k$.[9]
2. Let e_j be the j-th basis vector (whose j-th element is 1, and all other elements 0). Calculate the authority scores $a_j = \sum_{i=1}^{k} f(\lambda_i)(e_j^T x_i)^2$. (This is the square of the length of the projection of e_j onto the subspace spanned by $x_1, \ldots, x_k$, where the projection in the x_i direction is "weighted" by $f(\lambda_i)$.)

There are many choices for f: If we take $f(\lambda) = 1$ when $\lambda \geq \lambda_{\max}$ and $f(\lambda) = 0$ otherwise, we get back the original HITS algorithm; taking $k = n$ and $f(\lambda) = \lambda$ corresponds to simple citation counting; if we take $f(\lambda) = 1$, the authority weight of a page becomes $\sum_{i=1}^{k} x_{ij}^2$. In the last case, the authority weights depend only on the *subspace* spanned by the k eigenvectors, but not on the eigenvectors themselves (see, e.g., [7]). This new method thus gives a simple yet principled way of automatically combining multiple eigenvectors into a single measure of authoritativeness for each page. We call this second method Subspace HITS. In general, subspaces are more stable than individual eigenvectors (see [15]), therefore it is reasonable to expect that Subspace HITS will do better than HITS in certain cases. More specifically, if we use all the eigenvectors ($k = n$), then we have the following, strong, stability guarantee that requires no assumptions on the eigenvalues:

THEOREM 4. *Let f be Lipschitz continuous with Lipschitz constant L,[10] and let $k = n$. Let the co-citation matrix be perturbed according to $\tilde{S} = S + E$, where $||E||_F = \epsilon$ (E symmetric). Then the change in the vector of authority scores is bounded as follows:*

$$||a - \tilde{a}||_2 \leq L\epsilon \tag{6}$$

The proof of the theorem is given in the Appendix. Note that, for computational reasons, it will frequently be more practical to use $k < n$ eigenvectors as an approximation to using the full set (which is reasonable since this corresponds to dropping the eigenvectors with the smallest weight). In subsequent experiments, we will take $k = 20$ and $f(\lambda) = \lambda^2$.

[7] Technically, this calculates not the stationary distribution, but n times the stationary distribution, where n is the number of pages.

[8] Note that Kleinberg also discusses using multiple eigenvectors, but proposes asking the user to interpret individual eigenvectors rather than combining the eigenvector projections into a single authority score as we do here.

[9] In the case of repeated eigenvalues, we assume the eigenvectors are chosen orthogonal to each other.

[10] Formally, this means that, for all x, y, we have that $|f(x) - f(y)| \leq L|x - y|$. This is certainly satisfied if f has a first derivative bounded by L. Note that, even if f does not have uniformly bounded derivatives over the entire real line, the theorem will also hold so long as the derivatives are bounded within the applicable domain.

6. EXPERIMENTS

We now present experimental results comparing HITS, PageRank, Randomized HITS, and Subspace HITS. We use both academic paper citation data from the Cora database and web page linkage data built from actual queries. Section 6.1 focuses on the stability of the four algorithms, and Section 6.2 discusses the "diversity" of pages returned in the web query experiments.

6.1 Stability Results

Our first experiment used all of the Artificial Intelligence (AI) papers in Cora. Our results largely replicated those of [5], with HITS returning several Genetic Algorithms (GA) papers as the top-ranked ones. These results, however, were very sensitive to perturbation, and indeed under perturbation we often found that HITS omitted the GA papers and returned as top-ranked documents seminal papers from broader AI areas. In contrast, PageRank almost always returned general AI papers as the top ranked ones, and the results were very stable to perturbation.

HITS results on all Cora AI papers:

1	"Genetic Algorithms in Search, Optimization...", Goldberg	4	1	1	4	24
2	"Adaptation in Natural and Artificial Systems", Holland	7	2	2	5	460
3	"Genetic programming: On the programming of...", Koza	19	3	3	11	484
4	"Analysis of the behavior of a class of (GA)", De Jong	92	4	4	94	566
5	"Uniform crossover in genetic algorithms", Syswerda	389	5	5	418	793
6	"Artificial Intelligence through simulated evolution", Fogel+al	170	8	9	211	655
7	"A survey of evolution strategies", Back+al	408	7	8	723	788
8	"Optimization of control parameters for (GA)", Grefenstette	505	6	6	887	934
9	"The GENITOR algorithm and selection pressure", Whitley	362	9	7	284	756
10	"Genetic Algorithms + Data Structures = ...", Michalewicz	340	11	10	292	643
11	"Genetic Programming II: Automatic Discovery", Koza	-	10	-	-	-
2060	"Learning internal representations by error...", Rumelhart+al	2	-	-	3	-
2061	"Learning to Predict by temporal...", Sutton	5	-	-	6	-
2063	"Classification and Regression Trees", Breiman+al	1	-	-	1	1
2065	"Pattern classification and scene analysis", Duda+Hart	3	-	-	2	2
2087	"UCI repository of machine learning databases", Murphy+al	8	-	-	7	3
2100	"Irrelevant features and subset selection", John+al	10	-	-	9	4
2110	"Very simple classification rules perform well", Holte	-	-	-	10	5
2111	"Probabilistic Reasoning in Intelligent Systems", Pearl	6	-	-	8	-
2130	"The CN2 induction algorithm", Clark+al	9	-	-	-	-
2134	"Learning Boolean Concepts...", Almuallim+Dietterich	-	-	-	-	7
2139	"The MONK's problems: A performance...", Thrun	-	-	-	-	6
2189	"C4.5: Programs for Machine Learning", Quinlan	-	-	-	-	8
2263	"Multi-interval discretization of continuous...", Fayyad+al	-	-	-	-	9
2304	"A conservation law for generalization performance", Schaffer	-	-	-	-	10

PageRank results on all Cora AI papers ($\epsilon = 0.2$):

1	"Classification and Regression Trees", Breiman+al	1	2	3	2	2
2	"Genetic Algorithms in search, optimization...", Goldberg	3	1	2	1	1
3	"Probabilistic Reasoning in Intelligent Systems", Pearl	2	3	1	3	4
4	"Learning internal representations by error...", Rumelhart+al	4	4	4	4	3
5	"Adaptation in natural and artificial systems", Holland	5	5	5	5	5
6	"Pattern classification and scene analysis", Duda+Hart	6	7	6	6	6
7	"A robust layered control system for a mobile...", Brook+al	7	6	7	7	10
8	"Genetic Programming: On the programming of...", Koza	10	9	9	8	7
9	"Learning to predict by the methods of temporal...", Sutton	8	8	8	9	9
10	"Maximum likelihood from incomplete data via...", Dempster+al	9	10	10	10	8

HITS seemed to be unstable primarily in its flipping between Genetic Algorithms (GA) papers and other papers. To attempt to create a slightly more favorable environment for HITS, we repeated the above experiment, but keeping only the AI papers that were not GA papers. We obtained:

HITS results on subset of Cora AI papers (1st eigenvector):

1	"Classification and Regression Trees", Breiman+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...", 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...", 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 MDS Principle", Quinlan	-	8	-	7	8
13	"Multi-interval discretization of continuous...", Fayyad+Irani	-	-	-	-	10
14	"Learning relations by pathfinding", Ricchards+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 computaion", Hertz+al	-	-	8	-	-
49	"Explanation-based generalization: a unifying view", Mitchell	-	-	10	-	-
282	"A robust layered control system for a mobile robot", Brooks	-	-	9	-	-

HITS results on subset of Cora AI papers (2nd eigenvector):

1	"Learning to predict by the methods of temporal...", Sutton	1	1	4	4	6
2	"Learning internal representations by error...", Rumelhart+al	45	2	856	2	4
3	"Probabilistic Reasoning in Intelligent Systems", Pearl	109	3	33	1	5
4	"A robust layered control system for a mobile...", Brook+al	2	4	1	6	1
5	"STRIPS: A New Approach to the Application of...", Fikes+al	4	5	2	8	2
6	"Learning to act using real-time dynamic programming", Barto+al	5	6	9	11	75
7	"Neuronlike elements that can solve difficult...", Barto+al	7	7	34	17	81
8	"Integrated Architectures for Learning, Planning...", Sutton	6	11	13	14	74
9	"Explanation-based generalization: a unify T. M. Mitchell,	37	8	25	7	27
10	"Practical issues in temporal difference learning", Tesauro	10	9	36	29	78
12	"Maximum likelihood from incomplete data via...", Dempster+al	-	-	-	9	-
13	"Automatic programming of behavior-based robots", Mahadevan+al	9	-	-	-	-
14	"An implementation of a theory of activity", Agre+al	-	-	6	-	-
16	"Pattern classification and scene analysis", Duda+Hart	-	-	5	5	-
18	"SOAR: An architecture of general intelligence", Laird+al	-	-	-	-	3
19	"Introduction to the theory of neural computation", Hertz+al	-	-	-	10	-
28	"Reactive Reasoning and Planning", Georgeff+al	-	-	8	-	8
32	"Classification and Regression Trees", Breiman+al	3	10	3	3	-
4233	"UCI repository of machine learning databases", Murphy+al	8	-	7	-	-

We see that, apart from the top 2-3 ranked papers, the results are rather unstable. For example, Pearl's book is originally ranked 8-th, but drops to rank 459 on the second trial. Similarly, Brooks' paper has jumped up from rank 282 to rank 9 on trial three. The problem persists in the second eigenvector. However, this variability is not intrinsic to the problem, as shown by PageRank's results:

PageRank results on subset of Cora AI papers ($\epsilon = 0.2$):

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 is a drop from 10 to 14—these results are much more stable than for HITS. Closer examination of the HITS' authority weights reveals that its jumps in rankings are not merely small fluctuations in authority weights, but are indeed large changes. The PageRank scores, on the other hand, tend to remain fairly stable.

We next present some results using Randomized HITS and Subspace HITS. Both are more stable than HITS, though Subspace HITS seems to perform a little worse than PageRank.

Randomized HITS results on subset of Cora AI papers ($\epsilon = 0.2$):

1	"Learning internal representations by error...", Rumelhart+al	1	3	3	2	1
2	"Probabilistic Reasoning in Intelligent Systems", Pearl	4	1	1	1	2
3	"Classification and Regression Trees", Breiman+al	2	2	2	3	4
4	"Pattern classification and scene analysis", Duda+Hart	3	4	4	4	3
5	"Maximum likelihood from incomplete data via...", Dempster+al	5	6	6	6	5
6	"A robust layered control system for a mobile robot", Brook+al	6	5	5	5	6
7	"Numerical Recipes in C", Press+al	7	7	7	7	7
8	"Learning to Predict by the Method of Temporal...", Sutton	8	8	8	8	8
9	"STRIPS: A New Approach to ... Theorem Proving", Fikes+al	9	10	10	10	15
10	"Introduction To The Theory Of Neural Computation", Hertz+al	11	11	9	9	9
11	"Stochastic relaxation, gibbs distributions, ...", Geman+al	10	9	-	-	-
12	"Introduction to Algorithms", Cormen+al	-	-	-	-	10

Subspace HITS results on subset of Cora AI papers ($k = 20$, $f(\lambda) = \lambda^2$):

1	"Probabilistic Reasoning in Intelligent Systems", Pearl	4	1	1	1	4
2	"Classification and Regression Trees", Breiman+al	2	2	2	3	3
3	"Learning internal representations by error...", Rumelhart+al	1	3	3	2	1

4	“A robust layered control system for a mobile...”, Brooks	3	4	4	4	2
5	“Pattern classification and scene analysis”, Duda+hart	5	9	5	7	7
6	“Maximum likelihood from incomplete data via...”, Dempster+al	6	7	7	6	6
7	“Learning to predict by the method of emporal...”, Sutton	8	6	6	5	5
8	“STRIPS: A new approach to ... Theorem Priving...”, Fikes+al	7	5	8	8	9
9	“Explanation-based generalization: a unifying view”, Mitchell+al	9	8	9	9	8
10	“Learnability and the VC dimension”, Blumer+al	50	12	210	11	10
11	“Explanation-Based learning: An alternative view”, DeJong+al	-	10	10	10	-
12	“UCI repository of machine learning databases”, Murphy+Aha	10	-	-	-	-

To compare the four algorithms more extensively, we performed tests on 50 web queries on various topics (constructed by examining actual search engine queries). Kleinberg [8] describes a method for obtaining a collection of web pages on which to run HITS. We use exactly the method described there, and perturb the web page collection in a natural way.[11]

Some examples of query results are shown in the following four tables. (A “*” indicates a page originally in the top 10, but deleted by the perturbation.) Notice that, in the table below, with the exception of one trial, all of the original top 10 documents were “flipped” with something originally lower in rank. As discussed in Sections 4 and 6.2, this flipping phenomenon can arise from perturbations to the principal eigenvalue, which in turn causes the principal eigenvector to swap places with other eigenvectors.

HITS results on query “neural networks”: [long URLs truncated]

1	http://www.neci.nec.com/	*	1	*	*	*
2	http://researchindex.org/	*	2	*	*	*
3	http://citeseer.nj.nec.com/cs/	*	3	*	*	*
4	http://citeseer.nj.nec.com/terms.html	*	4	*	*	*
5	http://citeseer.nj.nec.com/yao93review.html	*	5	*	*	*
6	http://citeseer.nj.nec.com/17901.html	*	6	*	*	*
7	http://citeseer.nj.nec.com/yao91optimizat	*	7	*	*	*
8	http://citeseer.nj.nec.com/yao91simulated	*	8	*	*	*
9	http://citeseer.nj.nec.com/yao93evolution	*	9	*	*	*
10	http://citeseer.nj.nec.com/yao99evolving	*	10	*	*	*
12	http://www.ieee.org/	1	-	1	1	-
13	http://www.cs.washington.edu/research/jai	8	-	5	-	-
14	ftp://ftp.sas.com/pub/neural/FAQ.html	4	-	4	4	-
35	http://www.ieee.org/nnc/	2	-	2	2	-
36	http://www.okstate.edu/elec-engr/faculty/	3	-	3	3	-
37	http://www.icsi.berkeley.edu/˜jagota/NCS/	5	-	-	5	-
38	http://www.elsevier.nl/	6	-	-	-	-
39	http://www.inns.org/	7	-	6	6	-
40	http://www.ai.univie.ac.at/oefai/nn/nngro	10	-	-	7	-
41	http://synapse2.eng.wayne.edu/tpage3.html	9	-	7	-	-
44	http://www.emsl.pnl.gov:2080/docs/cie/neu	-	-	-	10	-
48	http://www.classify.org/safesurf/	-	-	8	8	-
49	http://www.weburbia.com/safe/ratings.htm	-	-	9	9	-
50	http://www.nd.com/	-	-	10	-	-
64	http://www.kcl.ac.uk/neuronet/	-	-	-	-	6
86	http://www.mitgmbh.de/	-	-	-	-	5
195	http://www.kcl.ac.uk/	-	-	-	-	7
230	http://www.kcl.ac.uk/neuronet/about/exec-	-	-	-	-	8
231	http://www.kcl.ac.uk/neuronet/about/map/	-	-	-	-	9
232	http://www.kcl.ac.uk/neuronet/about/roadm	-	-	-	-	10
381	http://www.ubcom.net/	-	-	-	-	1
382	http://www.brd.net/brd-cgi/sendemail/send	-	-	-	-	2
383	http://www.amazon.de/exec/obidos/redirect	-	-	-	-	3
384	http://amazon.de/exec/obidos/ASIN/3528064	-	-	-	-	4

The PageRank algorithm does not exhibit “flipping.” The following table presents PageRank results on the same query. With a few exceptions, the rankings are more stable under perturbation.

PageRank results on query “neural networks” ($\epsilon = 0.2$):

1	http://www.neci.nec.com/	*	1	*	*	*
2	http://researchindex.org/	*	2	*	*	*
3	http://www.ieee.org/	1	4	1	1	1
4	http://mathworld.wolfram.com/	5	3	*	3	5
5	http://www.wolfram.com/	*	5	246	*	*
6	http://www.cmu.edu/	2	6	2	2	4
7	http://192.38.71.109/htdig/search_thor.ht	3	*	3	*	2
8	http://www.unibo.it/	4	9	4	*	3
9	http://citeseer.nj.nec.com/cs/	*	7	*	*	*
10	http://citeseer.nj.nec.com/terms.html	*	8	*	*	*
11	http://www.okstate.edu/elec-engr/faculty/	6	10	5	4	-
13	http://www.ubcom.net/	7	-	6	5	6
14	http://www.brd.net/brd-cgi/sendemail/send	8	-	7	6	7
15	http://dmoz.org/about.html	9	-	8	7	-
16	http://ads.admonitor.net/clicktrack.cgi?F	10	-	9	8	-
17	http://www.deis.unibo.it/	-	-	10	-	8
18	http://www.cs.cmu.edu/	-	-	-	9	9
22	http://www.epfl.ch/	-	-	-	-	10
24	http://www.mathworks.com/	-	-	-	10	-

Results for Randomized HITS and Subspace HITS are listed below. Similar to the Cora results, both Randomized HITS and Subspace HITS seem to have comparable performance to PageRank.

Randomized HITS results on query “neural networks” ($\epsilon = 0.2$):

1	http://www.neci.nec.com/	*	1	*	*	*
2	http://researchindex.org/	*	2	*	*	*
3	http://www.ieee.org/	1	3	1	1	1
4	http://www.cmu.edu/	2	4	2	2	6
5	http://192.38.71.109/htdig/search_thor.ht	3	*	3	*	2
6	http://www.ubcom.net/	4	5	4	3	3
7	http://www.brd.net/brd*cgi/sendemail/send	5	6	5	4	4
8	http://dmoz.org/about.html	6	7	6	5	*
9	http://ads.admonitor.net/clicktrack.cgi?F	7	8	7	6	*
10	http://www.ieee.org/nnc/	8	9	8	8	*
11	http://www.unibo.it/	9	10	9	-	5
12	http://www.cs.cmu.edu/	10	-	10	7	8
13	http://www.deis.unibo.it/	-	-	-	-	7
14	ftp://ftp.sas.com/pub/neural/FAQ.html	-	-	-	9	-
15	http://www.erudit.de/erudit/index.htm	-	-	-	-	9
16	http://www.iau.dtu.dk/˜jj/address.html	-	-	-	-	10
17	http://www.okstate.edu/elec-engr/faculty/	-	-	-	10	-

Subspace HITS results on query “neural networks” ($k = 20$, $f(\lambda) = \lambda^2$):

1	http://www.neci.nec.com/	*	1	*	*	*
2	http://researchindex.org/	*	2	*	*	*
3	http://www.ieee.org/	1	3	1	1	1
4	http://www.cmu.edu/	2	4	2	2	5
5	http://www.ubcom.net/	3	5	3	3	2
6	http://www.brd.net/brd-cgi/sendemail/send	4	6	4	4	3
7	http://dmoz.org/about.html	5	7	5	5	*
8	http://ads.admonitor.net/clicktrack.cgi?F	6	8	6	6	*
9	http://www.ieee.org/nnc/	7	9	7	7	*
10	http://www.unibo.it/	8	10	8	*	4
11	http://www.deis.unibo.it/	9	-	9	-	6
12	ftp://ftp.sas.com/pub/neural/FAQ.html	-	-	-	9	-
13	http://mathworld.wolfram.com/	-	-	-	8	10
14	http://www.erudit.de/erudit/index.htm	10	-	10	-	8
15	http://www.iau.dtu.dk/˜jj/address.html	-	-	-	-	9
16	http://www.slac.stanford.edu/˜rhatcher/	-	-	-	10	-
20	http://www.cs.cmu.edu/	-	-	-	-	7

Note that rankings for all four algorithms appear less stable than those from the Cora dataset. This is largely an artifact of the way we perturb the web datasets. In the Cora experiments, removing a paper from the dataset does not remove any of the papers cited by or citing it. In the web query case, each deleted root page also removes its surrounding link structure. (See footnote 11.) This type of perturbation more accurately models the scenario of search engines missing certain pages, but is also a much harsher form of perturbation in which a large component of the graph can be removed all at once. Under different perturbation models, we also obtain results closer to the Cora ones (for instance, in which there was much less “flipping” of HITS’ eigenvectors.)

To get a better idea of the flipping pattern of eigenvectors, we count the number of top ten pages which drop in ranking drastically in each trial. (Because our collections are perturbed by deletions rather than insertions, large *rises* in ranking are rare for all four algorithms. Therefore large *drops* in ranking are more interesting.) There are 50 total web queries, and 5 trials are performed for each query. The histogram counts in Figure 2 represent the number of

[11] Kleinberg [8] first uses a text-based 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, for example to ignore intra-domain references) to define the web-graph on which HITS operates. Our perturbations are 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.

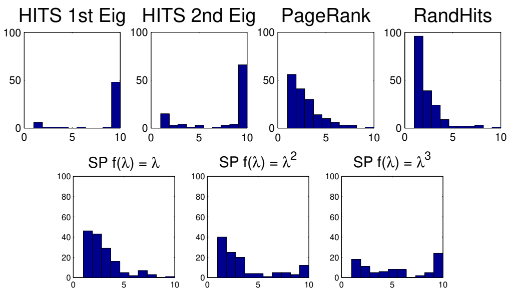

Figure 2: "Flip" count histograms for the four algorithms, showing how frequently c pages originally in the top 10 fall out of the top 20 after perturbation, for $c = 1, \ldots, 10$.

ϵ	HITS	Subspace HITS	PageRank	Rand HITS
0.2	21.20%	16.56%	17.00%	14.08%
0.1	21.20%	16.56%	18.40%	13.88%
0.05	21.20%	16.56%	19.96%	13.92%
0.02	21.20%	16.56%	20.52%	14.16%
0.01	21.20%	16.56%	19.72%	14.40%

Table 1: Expected percentage of rank drops with decreasing ϵ.

top ten pages that drop below rank 20, i.e., if 8 out of the top 10 pages drop below rank 20 in a certain trial for PageRank, then bin 8 in the PageRank histogram gets incremented by 1. We call these plots "flip histograms." In Figure 2, we exclude the 0 bin in order to highlight the trials where something goes wrong. From the figure, we can see that HITS is much more likely than PageRank and Randomized HITS to have large numbers of pages simultaneously drop to low rankings, though PageRank and Randomized HITS are also more likely to have at least one such drop on any given trial. More specifically, out of 250 trials, 8 to 10 pages drop together in 49 HITS trials; this number increases to 73 for the second eigenvector of HITS. The number of trials on which this occurs is 4 for both PageRank and Randomized HITS. For Subspace HITS, when $f(\lambda) = \lambda$ or $f(\lambda) = \lambda^2$, the histogram is similar to PageRank and Randomized HITS. But as we change f to $f(\lambda) = \lambda^3$ (so that it grows faster with λ), it behaves increasingly similarly to HITS. This is no surprise given that, as the degree of f increases, Subspace HITS gives increasing weight to the principal eigenvector.

Note that a flip histogram is in fact the empirical distribution of the number of rank drops in a trial. In Table 1, we show the effect of the reset probability ϵ on the expectation of the percentage of pages that suffer rank drops under this empirical distribution. For PageRank, rank drop expectations increase as ϵ increases, whereas for Randomized HITS, the expectation only fluctuates slightly. Since HITS and Subspace HITS do not involve resetting, their expectations are not affected.

6.2 Multiple Connected Components

A closer examination of the web query results draws our attention to the question of the "diversity" of the pages returned. Ignoring the stability issue for now, we notice that the algorithms return pages with a different "range" in the number of domains. The tables below present results from HITS and PageRank on the query "SQL tutorial." All the top ten pages returned by HITS are from the same site, and are therefore not very useful even had the rankings been stable.[12] PageRank, on the other hand, returns a wider variety of web pages. The question of variety is probably not as crucial as stability, but the example leads us to consider the different algorithms' behavior when there are multiple "connected components" in the linkage graph.

HITS results on query "sql tutorial":

1	http://www.internet.com/	1	1	36	14	1
2	http://www.internet.com/sections/download	2	2	37	15	2
3	http://www.internet.com/sections/internat	3	3	38	16	3
4	http://www.internet.com/sections/isp.html	4	4	39	17	4
5	http://www.internet.com/sections/it.html	5	5	40	18	5
6	http://www.internet.com/sections/marketin	6	6	41	19	6
7	http://www.internet.com/sections/news.htm	7	7	42	20	7
8	http://www.internet.com/sections/resource	8	8	43	21	8
9	http://www.internet.com/sections/stocks.h	9	9	44	22	9
10	http://www.internet.com/sections/webdev.h	10	10	45	26	10
2033	http://welcome.hp.com/country/us/eng/term	-	-	1	1	-
2034	http://welcome.hp.com/country/us/eng/welc	-	-	2	2	-
2035	http://welcome.hp.com/country/us/eng/howt	-	-	3	3	-
2036	http://welcome.hp.com/country/us/eng/priv	-	-	4	4	-
2037	http://welcome.hp.com/country/us/eng/prod	-	-	5	5	-
2038	http://welcome.hp.com/country/us/eng/solu	-	-	6	6	-
2039	http://welcome.hp.com/country/us/eng/supp	-	-	7	7	-
2040	http://welcome.hp.com/country/us/eng/cont	-	-	9	9	-
2041	http://www.hp.com/go/search-us-eng/	-	-	8	8	-
2042	http://www.hp.com/go/smartfriend/	-	-	10	10	-

PageRank results on query "sql tutorial" ($\epsilon = 0.2$):

1	http://search.intraware.com/search.html	1	1	1	1	1
2	http://jazz.external.hp.com/	*	1572	1155	2	2
3	http://www.hp.com/go/search-us-eng/	*	2	3	3	3
4	http://www.yahoo.com/	2	4	2	5	4
5	http://jump.altavista.com/ff.go	*	3	4	4	*
6	http://www.sun.com/	8	10	18	16	22
7	http://www.altavista.com/	109	5	5	6	132
8	http://www.goto.com/	3	15	6	22	5
9	http://welcome.hp.com/country/us/eng/term	*	20	22	7	6
10	http://welcome.hp.com/country/us/eng/welc	*	21	23	8	7
11	http://www.webring.org/cgi-bin/webring?ri	4	-	7	-	8
12	http://www.webring.org/cgi-bin/webring?ri	5	-	8	-	9
13	http://www.webring.org/cgi-bin/webring?ri	6	-	9	-	10
14	http://u.extreme-dm.com/?login=astentec	7	-	10	-	-
15	http://v1.nedstatbasic.net/stats?AAlCJgnm	9	-	-	-	-
16	http://welcome.hp.com/country/us/eng/howt	-	-	-	10	-
23	http://www.wiwi.uni-frankfurt.de/	10	-	-	-	-
42	http://www.sqlcourse.com/	-	6	-	9	-
50	http://www.webreference.com/	-	7	-	-	-
54	http://ecommerce.internet.com/	-	8	-	-	-
59	http://www.sqlcourse2.com/	-	9	-	-	-

A connected component of a graph is a subset of nodes 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. If a graph has multiple connected components, then the principal eigenvector of $A^T A$ will have non-zero values only for nodes from the "largest" connected component (more formally the component with the largest eigenvalue). (See, e.g. [4].) Therefore, unless the eigenvalue is shared among several connected components (which, as explained in Section 7, would render the eigenvector unstable), each HITS eigenvector would only represent one part of the graph. On the other hand, an algorithm with a universal resetting probability obtains its principal eigenvector by concatenating (and possibly weighting) the results from all of the connected components.[13] Therefore, PageRank and Randomized HITS return results from a wider variety of domains, while HITS concentrates its

[12] Pages from the same site tend to be heavily linked to each other, and therefore are known to "trap" authority and hub scores. Avoiding "rank traps" is one of the original design motivations behind PageRank [3]. Kleinberg [8] suggests avoiding this problem by ignoring intra-domain links to prune down this kind of false authority propagation. We have indeed followed Kleinberg's design in our experiments and pruned all links pointing to pages on the same domain, though multiple links from another domain are still allowed, and may account for the inclusion of pages from the same site in our results.

[13] For example, if a graph has two connected components C_1 and C_2, then PageRank run on the entire graph gives the same result as running it separately on C_1 and on C_2, and then concatenating

ϵ	HITS	Subspace HITS	PageRank	Rand HITS
0.2	8.59	8.97	9.38	9.49
0.1	8.59	8.97	9.32	9.48
0.05	8.59	8.97	9.26	9.48
0.02	8.59	8.97	9.23	9.44
0.01	8.59	8.97	9.24	9.41

Table 2: Average number of different sites from top 10 pages.

strength on the most popular cluster. Subspace HITS attempts to broaden its scope by using multiple eigenvectors, so that different connected components can be included. Of course, the previous discussion assumes the graph contains more than one connected component. If the entire graph is connected, then HITS would not have to worry about any of the above problems. However, empirical data from the web query experiments do indeed suggest the presence of multiple connected components in the graph. For example, consider the query results on "neural networks" presented in the last section. Without any perturbations, the top two ranked pages for all four algorithms are `www.neci.nec.com` and `researchindex.org`, both of which are removed in trials 1, 3, 4, and 5. All four algorithms then return `www.ieee.org` as the top ranked page in those trials. This is in fact the top ranked page in the second eigenvector of HITS. All the top ranked pages in the original principal eigenvector of HITS are from the same removed "cluster," whereas PageRank, Randomized HITS, and Subspace HITS also include pages from other clusters, and as a result suffer less impact from the cluster's removal.

Table 2 lists the average number of different sites/domains represented in the top 10 pages for each algorithm. Note the effect of decreasing the reset probability ϵ: PageRank's average decreases as ϵ decreases, indicating that the "diversity" of results is decreasing; Randomized HITS' diversity also decreases, but not as much. HITS and Subspace HITS are not affected as neither involves resetting.

7. LSI AND HITS

In this section we descibe an interesting connection between HITS and Latent Semantic Indexing (LSI) [6] that provides additional insight into our results (see also [5, 13]). In LSI a collection of documents is represented as a matrix A, where A_{ij} is 1 if document i contains the jth-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 right 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 row 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 retrieval.

Consider constructing the following citation graph from a set of documents. Let there be one node for each document and one for each word. Let each document node link to all the words that appear in it, and let $\hat{A}$ be the adjacency matrix of this graph. If we apply HITS to this graph, we find that only the document-nodes have non-zero hub weights (since the word-nodes do not link to anything) and only the word-nodes have non-zero authority weights. Moreover, the vector of HITS word authority weights is exactly x, the first right singular vector found by LSI. In this setup, an authoritative word is a commonly used word, and connected components in the linkage graph indicate distinct document topics.

This connection allows us to transfer insight from experiments on LSI to our understanding of HITS. In this vein, we carry out an experiment on the *Associated Press* (AP) portion of the TREC volume 1 corpus. In order to obtain non-trivial word authorities the words were stemmed and a stop list was used. We used a vocabulary size of 1500 words and 2000 articles, that had an average length of 72 words. We first ran LSI/HITS on the entire set of articles, keeping the top 20 authority vectors, i.e. the first 20 right singular vectors. Then we randomly removed 30% of the articles, and recalculated the authority weights.

Original vector		Perturbed vectors			
1	offici	1	against	1	news
2	kill	2	todai	2	member
3	against	3	talk	3	govern
4	death	4	mondai	4	negoti
5	year	5	report	5	talk
6	govern	6	member	6	peopl
7	israel	7	critic	7	agenc
8	west	8	london	8	report
9	soviet	9	publish	9	israel
10	british	10	american	10	offic
11	isra	11	kill	11	chairman
12	palestinian	12	district	12	meet
13	di	13	peopl	13	isra
14	member	14	fridai	14	u
15	capit	15	british	15	palestinian
16	bank	16	west	16	began
17	communist	17	book	17	newspap
18	london	18	communist	18	condition
19	arab	19	union	19	presid
20	associ	20	death	20	american

Table 3: Authoritative words from AP articles.

The results show sets of authoritative words which are unions of authoritative words from different topics. As discussed in Section 6.2, linear combinations of topics/connected components can occur when the subcomponents have roughly equal eigenvalues. Because of the small eigengap, linear combinations of eigenvectors are much more unstable under perturbation. This is demonstrated in Table 3, where words from the 13th original eigenvector originally contains words from both the Israeli-Palestinian middle-east conflict and British politics, but is split into eigenvectors 13 and 14 under perturbation. In perturbed eigenvector 13, the main topic is British politics, and words *palestinian* and *israel* drop to ranks 59 and 77, respectively; in perturbed eigenvector 14, the main topic is middle-east conflicts, with the word *british* dropping to rank 197.

This example also illustrates the danger of defining "semantic directions" for individual eigenvectors. In the presence of multiple topics/connected components, eigenvectors could be a linear combination of the authoritative nodes from different topics, and the combination can be very sensitive to small perturbations. It is much safer to look at a subspace, as is done in LSI and Subspace HITS.

8. SUMMARY

In this paper, we analyzed the stability of HITS and PageRank to small perturbations of a document collection. Based on our finding that the stability of PageRank stems from its usage of a "reset" to the uniform distribution, we proposed Randomized HITS,

the results (weighting C_1's nodes' values by $|C_1|/(|C_1|+|C_2|)$, and similarly for C_2, where $|C_i|$ is the number of web pages in connected component C_i).

which retains the hubs-and-authorities notion of influential articles of HITS, but which is more stable to perturbation. We also presented a second algorithm, Subspace HITS, that is motivated by the observation that subspaces spanned by a few eigenvectors may sometimes be stable even when individual eigenvectors are not. This algorithm may also be viewed as a principled way of combining multiple HITS eigenvectors. We also reported on the empirical performance of the four algorithms, and explored the issue of "diversity" of the results returned by the algorithms, focusing on the setting of web graphs with multiple connected components. Finally, connections between LSI and HITS were discussed.

9. REFERENCES

[1] B. Amento, L. G. Terveen, and W. C. Hill. Does "authority" mean quality? Predicting expert quality ratings of web documents. In *Proc. 23rd Annual Intl. ACM SIGIR Conference*, pages 296–303. ACM, 2000.

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

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

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

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

[6] 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.

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

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

[9] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the contruction of Internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000.

[10] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors, and stability. In *Proc. 17th International Joint Conference on Artificial Intelligence*, 2001.

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

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Unpublished Manuscript, 1998.

[13] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proc. SIGMODS/PODS*, 1998.

[14] Davood Rafiei and Alberto Mendelzon. What is this Page Known for? Computing Web Page Reputations. In *Proc. WWW9 Conference*, 2000.

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

Appendix: Proof of Stability of Subspace HITS

Let f be a Lipschitz continuous function with Lipschitz constant L. Given a co-citation matrix $S = A^TA \in \mathbb{R}^{n\times n}$, we first take the singular value decomposition of S to obtain $S = U\Sigma U^T$, where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$. Define $f(\Sigma)$ to be the diagonal matrix whose (i,i)-element is $f(\Sigma_{ii})$. We calculate $T = Uf(\Sigma)U^T$, and define the authority of page i to be $a_i = e_i^T T e_i$ (where e_i is the i-th basis vector). Note that the diagonal of T contains all the authority scores. The reader may easily verify that this is equivalent to our previous definition of $a_i = \sum_j f(\sigma_j)(u_j^T e_i)^2$ where u_j is the j-th column of U.

Let $\tilde{S} = S + E$, and let $\tilde{U}$, $\tilde{u}_i$, $\tilde{\Sigma}$, $\tilde{\sigma}_i$, $\tilde{T}$, and $\tilde{a}_i$ be all the corresponding perturbed quantities. Let $\epsilon = ||E||_F$. By the invariance of Frobenius norm to unitary transforms:

$$||\tilde{U}^T SU - \tilde{U}^T\tilde{S}U||_F = ||\tilde{U}^T(S-\tilde{S})U||_F = ||E||_F = \epsilon. \quad (7)$$

Now, the (i,j)-element of the matrix $\tilde{U}^TSU - \tilde{U}^T\tilde{S}U$ is:

$$\begin{aligned} \tilde{u}_i^T S u_j - \tilde{u}_i^T\tilde{S}u_j &= \tilde{u}_i^T U\Sigma U^T u_j - \tilde{u}_i^T\tilde{U}\tilde{\Sigma}\tilde{U}^T u_j && (8)\\ &= \tilde{u}_i^T U\sigma_j e_j - e_i^T\tilde{\sigma}_i\tilde{U}^T u_j && (9)\\ &= (\sigma_j - \tilde{\sigma}_i)\tilde{u}_i^T u_j. && (10)\end{aligned}$$

Substituting this back into Equation (7), we have that

$$\textstyle\sum_{i=1}^n\sum_{j=1}^n(\sigma_j - \tilde{\sigma}_i)^2(\tilde{u}_i^T u_j)^2 = \epsilon^2. \quad (11)$$

Now, treating the columns of U and $\tilde{U}$ as two bases for $\mathbb{R}^n$, for each $k = 1, \ldots, n$, we let $\alpha_j^{(k)}$ and $\tilde{\alpha}_i^{(k)}$ be the basis expansions of e_k in these two bases. i.e., $\alpha_j^{(k)} = e_k^T u_j$, and $\tilde{\alpha}_i^{(k)} = e_k^T\tilde{u}_i$, so that $e_k = \sum_{j=1}^n \alpha_j^{(k)} u_j$ and $e_k = \sum_{i=1}^n\tilde{\alpha}_i^{(k)}\tilde{u}_i$. Note also that

$$\textstyle\sum_{j=1}^n \alpha_j^{(k)}\alpha_j^{(l)} = 1\{k=l\}, \quad \sum_{i=1}^n\tilde{\alpha}_i^{(k)}\tilde{\alpha}_i^{(l)} = 1\{k=l\} \quad (12)$$

LEMMA 5. *Let* $G : \{1,\ldots,n\}\times\{1,\ldots,n\} \mapsto \mathbb{R}$ *satisfy* $\sum_{i=1}^n\sum_{j=1}^n(G(i,j))^2 \le \tau^2$. *Then*

$$\textstyle\sum_{k=1}^n\left(\sum_{i=1}^n\sum_{j=1}^n\tilde{\alpha}_i^{(k)}\alpha_j^{(k)}G(i,j)\right)^2 \le \tau^2. \quad (13)$$

Proof. View G, whose domain has n^2 elements, as a vector in $\mathbb{R}^{n^2}$. Also define $H^{(k)}$, given by $H^{(k)}(i,j) = \tilde{\alpha}_i^{(k)}\alpha_j^{(k)}$ to be another vector in this space. Using Equation (12), it is easy to show that $\sum_{i=1}^n\sum_{j=1}^n H^{(k)}(i,j)H^{(l)}(i,j) = 1\{k=l\}$, so that the vectors $H^{(k)}$ in fact form an orthonormal basis for an n dimensional subspace of $\mathbb{R}^{n^2}$. In $\mathbb{R}^{n^2}$, the precondition in the Lemma is exactly that "$||G||_2 \le \tau$." Moreover, the inner product of G and $H^{(k)}$ is $\langle H^{(k)}, G\rangle = \sum_{i=1}^n\sum_{j=1}^n\tilde{\alpha}_i^{(k)}\alpha_j^{(k)}G(i,j)$. Since the projection of a vector onto some subspace can be no longer than the original vector, we have $\sum_{k=1}^n\left(\langle H^{(k)}, G\rangle\right)^2 \le ||G||_2^2 \le \tau^2$, which proves the lemma. □

Proof of Theorem 4. We have the following:

$$\begin{aligned} ||a-\tilde{a}||_2^2 &= \textstyle\sum_{k=1}^n(a_k-\tilde{a}_k)^2 = \sum_{k=1}^n(e_k^TTe_k - e_k^T\tilde{T}e_k)^2 && (14)\\ &= \textstyle\sum_{k=1}^n\left(\sum_{i=1}^n\sum_{j=1}^n\tilde{\alpha}_i^{(k)}\alpha_j^{(k)}(f(\sigma_j)-f(\tilde{\sigma}_i))\tilde{u}_i^Tu_j\right)^2 && (15)\end{aligned}$$

Define $G(i,j) = (f(\sigma_j) - f(\tilde{\sigma}_i))\tilde{u}_i^Tu_j$. Using the Lipschitz condition on f, we have:

$$\begin{aligned}\textstyle\sum_{i=1}^n\sum_{j=1}^n(G(i,j))^2 &= \textstyle\sum_{i=1}^n\sum_{j=1}^n(f(\sigma_j)-f(\tilde{\sigma}_i))^2(\tilde{u}_i^Tu_j)^2\\ &\le \textstyle\sum_{i=1}^n\sum_{j=1}^n L^2|\sigma_j-\tilde{\sigma}_i|^2(\tilde{u}_i^Tu_j)^2\\ &= L^2\epsilon^2,\end{aligned}$$

where the last step used Equation (11). Applying Lemma 5 with $\tau^2 = L^2\epsilon^2$, we therefore conclude that

$$||a-\tilde{a}||_2^2 \le \textstyle\sum_{k=1}^n\left(\sum_{i=1}^n\sum_{j=1}^n\tilde{\alpha}_i^{(k)}\alpha_j^{(k)}G(i,j)\right)^2 \le L^2\epsilon^2 \;\square$$