Perturbation of the hyper-linked environment

Hyun Chul Lee and Allan Borodin

Department of Computer Science
University of Toronto
Toronto, Ontario, M5S3G4
{leehyun,bor}@cs.toronto.edu

Abstract. After the seminal paper of Kleinberg[1] and the introduction of PageRank[2], there has been a surge of research activity in the area of web mining using link analysis algorithms. Subsequent to the first generation of algorithms, a significant amount of improvements and variations appeared. However, the issue of stability has received little attention in spite of its practical and theoretical implications. For instance, the issue of "link spamming" is closely related to stability: is it possible to boost up the rank of a page by adding/removing few nodes to/from it? In this paper, we study the stability aspect of various link analysis algorithms concluding that some algorithms are more robust than others. Also, we show that those unstable algorithms may become stable when they are properly "randomized".

keywords Search, web graph, link structure, stability, PageRank, HITS, SALSA

1 Introduction

The use of link analysis algorithms for different web mining purposes became quite popular after the first introduction of algorithms to identify authoritative sources in the web[1, 2]. Different attempts [6, 7, 3, 9, 10, 12–14] to improve these algorithms were taken. A simple evaluation of query results using human judgement is normally employed to measure the performance of algorithms. A. Ng. et al.[4] and A. Borodin et al.[3] take a slightly different path from other papers: A. Ng. et al. study the stability aspect of some link analysis algorithms like PageRank and HITS, providing some insight into ways of designing stable link analysis methods. A. Borodin et al. introduce some formal definitions of stability and rank stability along with the analysis of some algorithms.

Stability is an important feature to consider in a such highly dynamic environment as World Wide Web. The World Wide Web is continuously evolving, so if a link analysis is to provide a robust notion of authoritativeness of pages, then it is natural to ask for a link analysis algorithm to be stable under small perturbations on the web topology. Intuitively, a small change of the web topology should not affect the overall link structure, and a proper definition of stability should reflect this intuition properly. The stability issue also has some practical implications such as that of "link spamming", i.e. a good link analysis algorithm should be robust to any malicious attempt of web designers to promote the rank of their pages by adding/removing few links to/from them.

The current link analysis algorithms can be classified into two categories. The first class of algorithms are *algebraic methods* such as HITS[1],PageRank [2], SALSA[6] and various hybrid algorithms of the first two[3, 9, 11]. These methods essentially compute principal eigenvectors of particular matrices related to the adjacency matrix of a certain web graph to identify the most relevant pages on that web graph. The second class of algorithms are *probabilistic methods* such as PHITS[10] and Bayesian [3] algorithms. These algorithms, using some probabilistic assumptions and techniques, estimate the rank of pages on a specific topic. *Algebraic methods* are the most popular ones, thus in this paper we only concentrate on the analysis of *algebraic methods*. More specifically, after introducing our revised definition of stability, we show the following results regarding the stability of algebraic methods: 1) PageRank is stable on the class of all directed graphs. 2) SALSA is stable on the class of authority connected graphs but not stable on the class of all directed graphs. 3) HITS is not stable on the class of authority connected graphs. Finally, we introduce randomized versions of HITS and SALSA showing stability for these algorithms.

2 Overview of algorithms

We begin by reviewing some algebraic link analysis algorithms, the reader familiar with this material may wish to skip ahead to Section 3.

2.1 HITS

Created by Kleinberg[1], HITS is the first link analysis algorithm used for web mining. In contrast to PageRank, it was never implemented in a commercial search engine until a new search engine Teoma[1] integrated a variation of HITS[2] as part of its ranking system. First, this algorithm constructs a *Root Set* of pages consisting of a short list of webpages returned by the search engine. Later, this *Root Set* is augmented by pages that are pointed to by pages in the *Root Set*, and also by pages that point to pages in the *Root Set* to form a larger set called *Base Set*, which makes HITS a query dependent method. With the *Base set*, HITS forms the adjacency matrix A where $A_{ij} = 1$ if there is a link from i to j and 0 otherwise. Next, it assigns to each page i an authority weight a_i and a hub weight h_i, then the equations $a_i^{(t+1)} = \sum_{j \to i} h_j^{(t)}$ and $h_i^{(t)} = \sum_{i \to k} a_k^{(t)}$ are iterated until $a_i^{(t)}$ and $h_i^{(t)}$ converge to the fixed points a_i^* and h_i^* respectively (with the vectors renormalized to unit length at each iteration). Also, it is easily seen than the fixed points a^* and h^* are principal eigenvectors of A^tA and AA^t respectively. The authority value of a page i is taken to be a_i^*, and the hub value of page i is taken to be h_i^* in a similar manner.

2.2 PageRank

The popularity of PageRank is due to the commercial success search engine Google [3] created by Brin and Page[2]. PageRank simulates a random surfer who jumps to a randomly chosen web page with probability ϵ, and follows one of the forward-links on the current page with probability $1-\epsilon$. This process defines a markov chain on the web pages. The transition probability matrix of this markov chain is given by $(\epsilon U + (1-\epsilon)A_{row})$ where A_{row} is constructed by renormalizing each row of the adjacency matrix A to sum to 1 [4] and U is the transition matrix of uniform transition probabilities. The vector p that represents PageRank scores of pages is then defined to be the stationary distribution of this markov chain. PageRank does not make distinction between hub values and authority values, rather it assigns a single value(PageRank) to each page. In this paper, the PageRank score p_i of page i is taken to be both authority and hub values of the page for the sake of our analysis.

2.3 SALSA

As an alternative algorithm to HITS(an algorithm to avoid "topic-drift"), SALSA is proposed by Lempel and Moran[6]. SALSA performs two random walks on web pages; a random walk by following a backward-link and then a forward-link alternately, and another one by following a forward-link and then a backward-link alternately. The authority weights are defined to be the stationary distribution of the former random walk, and the hub weights are defined to be the stationary distribution of the latter random walk. Thus, SALSA assigns separate hub and authority scores to each page. The transition probability matrices of the markov chains for the authorities and hubs are given by $\tilde{A} = A_{col}^t A_{row}$, $\tilde{H} = A_{row} A_{col}^t$, where A_{col} is constructed by renormalizing each column of the adjacency matrix A to sum to 1, and A_{row} is constructed by renormalizing each row of the adjacency matrix A to sum to 1. One attractive aspect of SALSA is that its stationarity distributions have explicit forms [6].

[1] http://www.teoma.com

[2] teoma vs. Google, Round Two, Siliconvalley.internet.com, April 2,2002

[3] http://www.google.com

[4] It is not clear from the original definition how to deal with the situation where the current page has no forward-link from it. In this paper, we use the simplest approach, i.e. when a page has no forward-link(a row of A has all zero entries), then the corresponding row of A_{row} is constructed to have all entries equal to 1/n.

3 Definitions and notations

In this section, we introduce some basic definitions and notations used throughout the rest of paper. Given G=(V,E) a directed graph representing a set of pages and their interconnecting links, we define the *co-citation graph* of G as an undirected graph $G_a = (V', E')$ such that V'=V and E'={(p,q)| if there exists a node r that links to both p and q }. A directed graph G=(V,E) is called *authority connected* if its co-citation graph is connected. The edge distance d_e between two graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ is defined as $d_e(G_1, G_2) = |(E_1 \cup E_2) \backslash (E_1 \cap E_2)|$. We define a link analysis algorithm T as a pair of functions that map a directed graph G of size N to a N-dimensional vector [5]. We call the vector $a^T(G)$ the authority weight vector of algorithm T on graph G and $h^T(G)$ the hub weight vector of algorithm T on graph G. The value of the entry $a_i^T(G)$ of vector $a^T(G)$ denotes the authority weight assigned by the algorithm T to the page i. Similarly, the value of the entry $h_i^T(G)$ of vector $h^T(G)$ denotes the hub weight assigned by the algorithm T to the page i. If the algorithm T does not make distinction between hub and authority values, then we treat the single weight of page as both hub and authority weights. If it is clear in the context, then we simply use $\boldsymbol{a}$ instead of $a^T(G)$ to denote the authority vector of algorithm on graph G, and a_i instead of $a_i^T(G)$ to denote the authority weight assigned by the algorithm T to the page *i*. Similar approach is used for the hub vector. Given a graph G, we can view a perturbation on graph G, as an operation ∂ on graph G, that adds and/or removes links to produce a new graph $G' = \partial$ G. We denote by $\tilde{a}^T(G) = a^T(\partial G)$ the new authority vector of the perturbed graph ∂G, and by $\widetilde{a}_i^T$ its respective new authority weight assigned by the algorithm T to page i.

Let BP and FP denote the set of pages whose backward-links are perturbed and the set of pages whose forward-links are perturbed respectively. Let BU denote the set of pages whose backward-links remain unperturbed even after the perturbation, and let FU be the set of pages whose forward-links remain unperturbed even after the perturbation. Let $\tilde{\mathcal{G}}$ be the set of all directed graphs, let $\mathcal{G}^{\mathcal{N}}$ be the class of all directed graphs of size N, let $\mathcal{G}_{\mathcal{AC}}$ be the class of all authority connected graphs, and let $\mathcal{G}_{\mathcal{AC}}^{\mathcal{N}}$ be the class of authority connected graphs of size N. Therefore, $\mathcal{G}_{\mathcal{AC}} \subset \tilde{\mathcal{G}}$, $\mathcal{G}^{\mathcal{N}} \subset \tilde{\mathcal{G}}$ and $\mathcal{G}_{\mathcal{AC}}^{\mathcal{N}} \subset \mathcal{G}_{\mathcal{AC}}$ hold. It is our particular interest to study the stability issues of link analysis algorithms on the class $\mathcal{G}_{\mathcal{AC}}$ because an authority connected graph can be viewed as representation of topical web graphs (set of pages that pertain to the same topic).

Before introducing our definition of stability, the original definition of stability will be introduced, so that the reader who is not familiar with [3] can understand the motivation driving a new definition.

Definition 1. *We say that an algorithm T is L_1-stable if for every fixed K, we have*

$$\lim_{N \to \infty} \max_{G_1 \in \mathcal{G}^{\mathcal{N}}, d_e(G, \partial G) \leq K} \min_{\gamma_1, \gamma_2 \geq 1} ||\gamma_1 a^T(G) - \gamma_2 a^T(\partial G)||_1 = 0$$

Based on this definition, A. Borodin et al. show 1) HITS is not stable on $\mathcal{G}^{\mathcal{N}}$. 2) SALSA is not stable on $\mathcal{G}_{\mathcal{AC}}^{\mathcal{N}}$ but it is stable on $\mathcal{G}^{\mathcal{N}}$.

We think this definition of stability is not sufficiently robust to reflect the realistic stability of link analysis algorithms, i.e. *the impact of perturbation depends on both the number and the weights of perturbed nodes, but rather, the definition only considers the number of perturbed links.* Thus, motivated by [4] which bounds the magnitude of perturbation for PageRank by a linear function of the aggregated PageRank scores of all perturbed pages, we define our notion of stability.

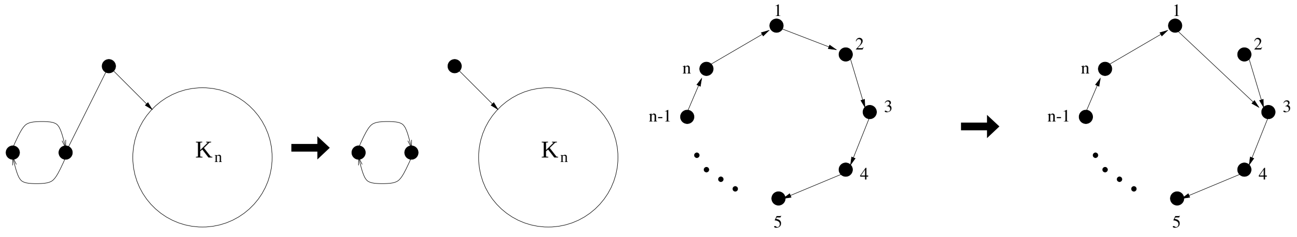
Definition 2. *Let c_i be the number of backward-links of page i that are perturbed. We say that an algorithm T is stable on $S \subseteq \tilde{\mathcal{G}}$ if we have a fixed constant value k such that for any $G \in S$ and $\partial G \in S$*

$$||a^T(G) - a^T(\partial G)||_1 \leq k \left(\sum_{i \in BP} c_i a_i + \sum_{j \in FP} h_j \right)$$

holds.

The intuitive idea behind this definition is as follows. Each time we add/remove a link there are two pages

[5] Notice that if the algorithm T does not make distinction between hubs and authorities, then T will be defined as single mapping



a. Stability of SALSA **b.** Stability of HITS

Fig. 1. Stability of SALSA and HITS

involved with this action, namely a page(call it j) whose forward-link is perturbed and another page(call it i) whose backward-link is perturbed. Roughly speaking, the "cost" of this addition/removal is $a_i + h_j$. Both a_i and h_j will contribute to the magnitude of perturbation, but the contribution of h_j would not be as considerable as that of a_i since the authority weight of a page is mainly due to backward-links rather than forward-links. Thus, a_i is more heavily weighted than h_j in our definition. Although it is also possible to have a definition in terms of the eigengap of particular matrix related to the link analysis algorithm, it presents some difficulties. For some link analysis algorithms like probabilistic ones, there is no natural way of formalizing eigengap since its role in the algorithm is obscure.

4 Results

In this section, we present our results regarding the stability of Pagerank, SALSA, and HITS.

4.1 Stability of PageRank

It is proven in [4] that $||\tilde{p} - p||_1 \leq 2/\epsilon \cdot \left(\sum_{i \in P} p_i\right)$, where P denotes the set of perturbed nodes. Slightly adapting this result to our definiton of stability, the following proposition is obtained.

Proposition 1. *PageRank is stable on the class of all directed graphs $\tilde{\mathcal{G}}$. Specifically, given a graph G=(V,E) $\in \tilde{\mathcal{G}}$, G is perturbed producing a new graph ∂G. Let p be the original PageRank score, then the new PageRank score $\tilde{p}$ satisfies:*

$$||\tilde{p} - p||_1 \leq \frac{2(1-\epsilon)}{\epsilon} \cdot \left(\sum_{i \in FP} p_i \right)$$

Note that our proposition only focuses on the set FP rather than the entire set of perturbed nodes.

4.2 Stability/Instability of SALSA

Proposition 2 (Appendix). *Let G, G' $\in \mathcal{G}_{\mathcal{AC}}$, s the original SALSA authority vector, and c_i the number of perturbed backward-links of page i, then the new SALSA authority vector $\tilde{s}$ obtained after the perturbation satisfies:*

$$||s - \tilde{s}||_1 \leq 2 \left(\sum_{i \in BP} \frac{c_i}{w} \right)$$

where w denotes the number of links (edges) in G. Moreover, if we only perturb those pages whose $|B(i)| > 0$ (†), then

$$||s - \tilde{s}||_1 \leq 2 \left(\sum_{i \in BP} c_i s_i \right)$$

Note that proposition 2 states that SALSA is stable on the class of authority connected graphs $\mathcal{G}_{AC}$ under the assumption (†).

Proposition 3. *SALSA is not stable on the class of all directed graphs $\tilde{\mathcal{G}}$*

Proof: Consider a graph that consists of complete graphs C_1 and C_2 of size 2 and n respectively (see Figure 1a). Also, there exists an extra hub h that points to two authority nodes p and q of the component C_1 and C_2 respectively. Now, we perturb the graph removing the link from hub h to authority p. Then, we observe that for the node s $\in C_1 \setminus p$, we have $a_s = 1/(n(n-1)+4)$, $\widetilde{a_s} = (2/(n+2))(1/2) = 1/(n+2)$, and for p, we have $a_p = \frac{2}{n(n-1)+4}$, $\widetilde{a_p} = (2/(n+2))(1/2) = 1/(n+2)$. Moreover, $h_h = 2/(n(n-1)+4)$. Thus,$|a_p - \widetilde{a_p}| + |a_s - \widetilde{a_s}| = (2n^2-5n+2)/((n+2)(n(n-1)+4)) > (n-6)/4 \cdot (2/(n(n-1)+4) + 2/(n(n-/1)+4)) = (n-6)/4 \cdot (a_p + h_h)$. Consequently $||a - \tilde{a}||_1 > (n-6)/4 \cdot (a_p + h_h)$ which proves the proposition.

4.3 Instability of HITS

To illustrate the high sensitivity of HITS to the topology of graph, we start with an example.

Example 1. Consider a graph G=(V,E) that consists of n nodes that form a cycle (See Figure 1b). More precisely, G has links $\{1 \rightarrow 2, 2 \rightarrow 3, \dots, n-1 \rightarrow n, n \rightarrow 1\}$. Next, G is perturbed by removing $1 \rightarrow 2$ and adding $1 \rightarrow 3$. The weight is evenly distributed among all nodes in G, i.e. for all i $\in$ V, we have $a_i = 1/n$, $h_i = 1/n$. On the other hand, we have $\tilde{a}_3 = 1$ and $\tilde{a}_i = 0$ for the rest of nodes. Moreover, we have $\tilde{h}_1 = 1/2$, $\tilde{h}_2 = 1/2$ and $\tilde{h}_i = 0$ for the rest of nodes. Hence, $||a - \tilde{a}||_1 = (n-1)/n + 1 - 1/n > 1 = n/3(a_2 + a_3 + h_1)$.

Note that this example shows that HITS fails to be stable even under small perturbation of a connected graph [6]. Therefore, the following proposition is not surprising.

Proposition 4. *HITS is not stable on the class of authority connected graphs $\mathcal{G}_{AC}$*

Proof: Consider the graphs G and ∂G that consist of 2n+1 nodes. Let $\mathcal{A} = \{1, \dots, n\}$ denote the first n nodes, let $\mathcal{B} = \{n+1, \dots, 2n\}$ denote the next n nodes, and let s denote the last node. Both graphs contain the links $\{s \rightarrow i | i \in \mathcal{B}\}$ and $\{i \rightarrow j | i \in \mathcal{A}, j \in \mathcal{B}\}$. The perturbed graph ∂G additionally contains links $\{j \rightarrow i | j \in \mathcal{B}, i \in \mathcal{A}\}$ and $\{s \rightarrow i : i \in \mathcal{A}\}$. G and ∂G are authority connected graphs. For all i $\in \mathcal{A}$, we have $a_i = 0$, $h_i = 1/(n+1)$, $\tilde{a}_i = 1/(2n)$. For all j $\in \mathcal{B}$, we have $a_j = 1/n$, $h_j = 0$, $\tilde{a}_j = 1/(2n)$. Finally, we have $a_s = 0$, $h_s = 1/(n+1)$, $\tilde{a}_s = 0$. Therefore, $||a - \tilde{a}|| \geq \sum_{i \in A} ||a_i - \tilde{a}_i|| = 1/2 > n/2(\sum_{i \in A} 2 \cdot a_i + \sum_{j \in B \cup \{s\}} h_j)$ proving instability.

Example 1 and Proposition 4 show some extreme scenarios where HITS fails to be stable. Apparently, addition/removal of even small number of links may alternate substantially the whole weight distribution under HITS, and the experimental study about the stability of HITS appears in Section 6.

5 Improvement of algorithms

In the previous section, some limitations of SALSA and HITS in terms of stability were shown. In this section, we explore how the randomization of the algorithms can eliminate their instability.

5.1 Randomized HITS

The first version of randomized HITS is introduced in [9] under the name of *two-level reputation rank*. Also, a slightly different version is proposed by A. Ng et al.[7] This randomization of HITS consists of the following random surfer model: the random surfer picks uniformly a random page with probability ϵ and follows a link with probability $1 - \epsilon$. If he decides to follow a link then he checks if it is odd time step or even time step. If it is odd time step, then he follows uniformly at random a forward-link. If it is even time step, then he

[6] It is not answered in [3] whether HITS is stable or not when the perturbed graph remains connected after the perturbation.

follows uniformly at random a backward-link. Note that this process defines a random walk on pages which is similiar in spirit to HITS. The stationary distribution on odd time steps is defined to be the authority weights of pages and the stationary distribution on even time steps is defined to be the hub weights of pages. Formally, the authority weights and hub weights of pages are calculated by updating the following equations:

$$a^{(t+1)} = \epsilon \cdot U + (1-\epsilon) \cdot A^t_{row} h^{(t)} \quad , \quad h^{(t+1)} = \epsilon \cdot U + (1-\epsilon) \cdot A_{col} a^{(t+1)} \tag{1}$$

where each entry of U is 1/n, A_{row} is the same as the adjacency matrix of the graph A with its rows normalized to sum to 1, A_{col} is the the same as the adjancency matrix of the graph A with its rows normalized to sum to 1. [7] The equations in (1) are iterated until they converge to the fixed points a^* and h^*. The convergence of these iterations is proved in [9]. We refer this version of HITS as *randomized HITS* or simply *RHITS*. Under *RHITS* each node is treated as both authority and hub. Next, we investigate stability aspect of *RHITS*.

Proposition 5. *RHITS is stable on the class of all directed graphs $\tilde{\mathcal{G}}$. Specifically, given a graph G=(V,E) $\in \tilde{\mathcal{G}}$,the graph G is perturbed producing a new graph ∂G, then we have*

$$||\tilde{a} - a||_1 \leq \frac{2(1-\epsilon)}{\epsilon} \cdot \left(\sum_{j \in FP} h_i + \frac{1}{2-\epsilon} \sum_{i \in BP} a_i \right)$$

By analogy to the proof of Pagerank, it is not very hard to show the stability.

5.2 Randomized SALSA

In a similar manner as that of HITS, it is possible to overcome the limitation of SALSA by randomizing the algorithm. Let call this algorithm *Randomized SALSA* or simply *RSALSA*. Let be two random surfers; The first random surfer picks uniformly a random page with probability ϵ, and it follows a backward-link then a foward link with probability $1-\epsilon$. This random surfer model defines the random walk on the authority nodes. The second random surfer picks uniformly a random page with probability ϵ, and it follows a forward-link then a backward-link with probability $1-\epsilon$, defining a random walk on the hub nodes. More precisely, the markov chain for the authorities and hubs have following transition probabilities

$$P_a(i,j) = \frac{\epsilon}{n} + (1-\epsilon) \sum_{\{k: k \in B(i) \cap B(j)\}} \frac{1}{|B(i)|} \frac{1}{|F(k)|} \ , \ P_h(i,j) = \frac{\epsilon}{n} + (1-\epsilon) \sum_{\{k: k \in F(i) \cap F(j)\}} \frac{1}{|F(i)|} \frac{1}{|B(k)|}$$

The convergence of markov chains to unique distributions are guaranteed from the fact that a markov chain that has transition probabilities P(x,y) of the form $P(x,y) = \epsilon\mu(y) + (1-\epsilon)Q(x,y)$ for some distributions μ and Q is *uniformly ergodic* [16]. Hence, the powers of transition probabilities converge geometrically to the unique distributions. Similar to *RHITS*, *RSALSA* treats each page as both authority and hub.

Proposition 6 (Appendix). *RSALSA is stable on the class of all directed graphs $\tilde{\mathcal{G}}$. Specifically, given a graph G=(V,E) $\in \tilde{\mathcal{G}}$ representing a web subgraph,the graph G is perturbed producing a new graph ∂G, then we have*

$$||\tilde{a} - a||_1 \leq \frac{4(1-\epsilon)}{\epsilon} \cdot \left(\sum_{i \in BP} a_i \right)$$

[7] When a row of A_{row} has all zero entries, then the corresponding row of A_{row} is constructed to have all entries equal to 1/n. Similarly, if a col of A_{col} has all zero entries, then the corresponding col of A_{col} is constructed to have all entries equal to 1/n.

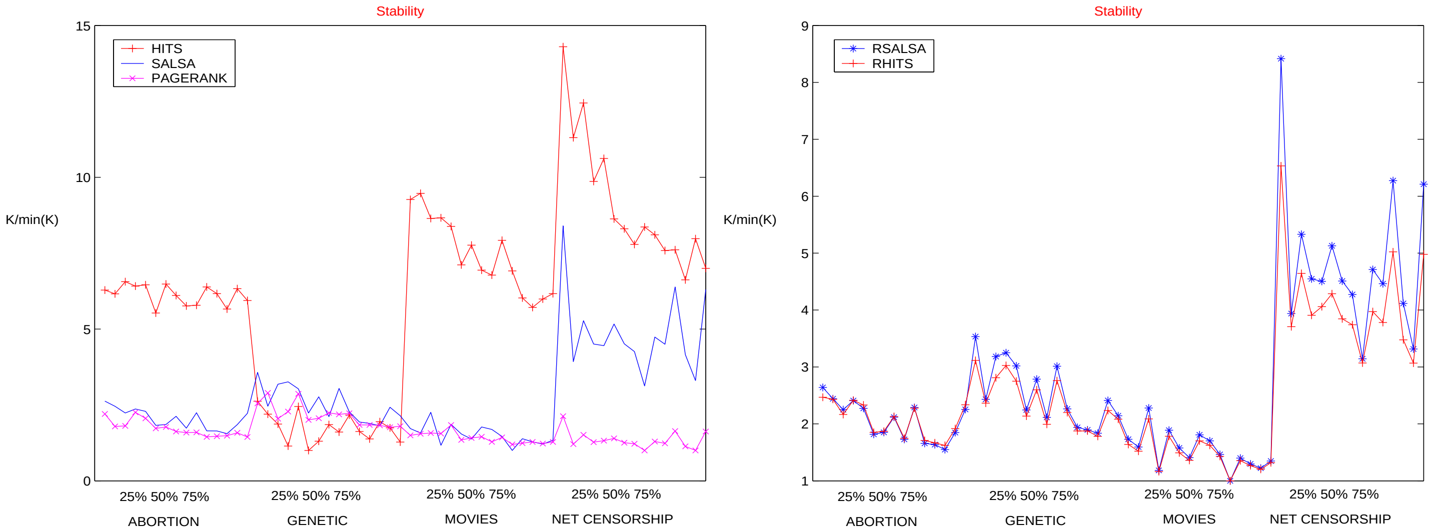


Fig. 2. Sensitivity Analysis

6 Experimental Results

Although the study of stability from the previous sections gives some useful theoretical insight about the robustness of algebraic link analysis algorithms, the notion of stability introduced in this paper is a *worst-case* notion. Hence, the theoretical analysis from previous sections will be complemented with some experimental studies to evaluate the robustness of algorithms in practice. Also, we study the performance of *RHITS* and *RSALSA* relative to some queries. From this study we show that *RHITS* can be, for instance, a way to overcome the limitation of HITS while being robust to perturbations since both *RHITS* and *RSALSA* outperform HITS specially on those queries in which HITS fails because of "Topic Drift" [6]. Stability results will be presented in Section 6.1 while the performance of algorithms relative to various queries will be presented in Section 6.2

6.1 Stability result

Given four sets of web pages produced as results of queries on "Genetic", "Abortion", "Movies" and "Net Censorship", we randomly perturbed each set by removing 25%, 50% and 75% of pages from each set of web pages. We ran five cycles of perturbation on each set to construct fifteen datasets in total for each topic. In order to measure the stability, we compared the magnitude of perturbation $||a - \tilde{a}||_1$ to the weights of perturbed pages. More precisely, we defined our sensitivity measure of link analysis algorithm T as $K = ||a - \tilde{a}||_1 / (\sum_{i \in BP} c_i a_i + \sum_{j \in FP} h_j)$ where c_i denotes the number of backward-links that are perturbed. Recall from our definition of stability that when the algorithm is not stable then K would be unbounded. Thus, volatile K values would be a possible indication of instability of the algorithm. In fact, HITS seems to present this kind of behavior as shown later on. We computed K for each dataset respect to all algebraic algorithms considered in this paper[8]. We divided each K by the smallest K on each series of datasets which is represented as the Y axis in Figure 2. For instance, All of K values computed from fifteen datasets on the query "Abortion" were divided by the smallest K,which is denoted by min(K), out of fifteen K values. One can observe from Figure 2 the high sensitivity of HITS from its volatile K/min(K) values on different queries. On the other hand, the stability of PageRank is remarkable showing a stable behavior regardless of the dataset. Finally, the sensitivity of SALSA, *RSALSA* and *RHITS* are shown to be between HITS and PageRank.

[8] Notice that when $\epsilon \to 1$, the stability of PageRank, *RHITS* and *RSALSA* is increased as the algorithms are reduced into simple uniform random jumps. Thus, the value of $\epsilon = 0.1$ was chosen to minimize the influence of ϵ on the stability of algorithms even though $0.1 < \epsilon < 0.2$ is the most widely used value of ϵ for PageRank

	HITS	SALSA	RHITS	RSALSA	PageRank
	1165	717	717	717	1984
	1193	962	962	962	1983
	1184	1769	1769	1769	2375
	1188	719	719	719	1985
	1191	925	0	925	2382
	1189	0	925	0	46
	1187	666	1461	666	2501
	1192	718	718	718	717
	1190	1325	666	1325	1139
	1948	2262	2	2262	368
precision	0.1	1	1	1	0.3

	HITS	SALSA	RHITS	RSALSA	PageRank
HITS	10	0	0	0	0
SALSA	0	10	8	10	0
RHITS	0	8	10	8	0
RSALSA	0	10	8	10	0
Pagerank	0	0	0	0	10

Index	URL	Title	Index	URL	Title
(1165)	DimeClicks.com -...	http://www5.dime clicks.com	(368)	Current Events - Law	http://law.miningco.com
(1193)	HitBox.com - ...	http://rd1.hitbox.com/	(1769)	Priests for Life Index	http://www.priestsforlife.org
(1184-92)	Amazon.com-...	http://www.amazon.com/	(0)	Abortion Clinics OnLine	http://www.gynpages.com
(1948)	Politics1: Hot Politics...	http://www.politics1.com/	(925)	Pregnancy Centers Online	http://www.pregnancy centers.org
(962)	ProlifeInfo	http://www.proli fe.org/ultimate	(1461)	Planned Parenthood Federation	http://www.planned parenthood.org
(1769)	Priests for Life Index	http://www.priests forlife.org	(666)	RoevWade.org	http://www.roevwade.org
(719)	Abortion and reprod. Res.	http://www.naral.org	(2)	The Abortion Rights Activist	http://www.cais.com /agm/main
(925)	Pregnancy Centers Online	http://www.pregnanc ycenters.org	(1984)	The John Birch Society	http://www.jbs.org
(0)	Abortion Clinics OnLine	http://www.gynpages.com	(1983)	American Opinion Book Service	http://www.aobs-store.com
(666)	RoevWade.org	http://www.roevwade.org	(2375)	About	http://home.about.com
(718)	Human Life International	http://www.hli.org	(1985)	TRIMonline	http://www.trimonline.org
(1325)	Feminists For Life of A.	http://www.serve.com /fem4life	(2382)	AllExperts.com	http://www.allexperts.com
(2262)	The Ultimate Pro-Life Resources	http://www.prolifeinfo.org	(46)	Project Rachel,..	http://manaco.simplenet.com/
(717)	National Right to Life Organization	http://www.nrlc.org	(2501)	The March For Life Fund	http://www.marchforlife.org
(1139)	Simple Catholicism	http://www.geocities.com/			

Table 1. Top 10 pages on "Abortion" (ϵ=0.1, Base Set Size=2293)

6.2 Performance Evaluation

In this section, we report results of series of experiments that we conducted to evaluate the ranking quality of algebraic link analysis algorithms considered in this paper. We ran each algorithm on four different queries using the same datasets as those of [3]. For the sake of brevity, we only present top 10 pages on the query "Abortion" (Table 1). The full set of experimental results can be found at the web page http://www.cs.toronto.edu/ leehyun/experiment.htm. The "Tightly Knit Community(TKC)" [6] effect for HITS is clearly observed with this particular query since its returned pages contain many irrelevant pages from "Amazon.com" in its top 10 pages. All top 10 pages produced by *RSALSA* and *RHITS*, in contrast, are relevant to the topic "Abortion". We can neglect the apparent similarity of ranking output between SALSA and RSALSA since a more careful study of low ranked pages revealed substantial difference between these algorithms.

7 Conclusions

We studied the stability aspect of different algebraic link analysis algorithms. We gave a new definition of stability motivated by the definition of stability given in [3] and some bounds for $||a - \tilde{a}||_1$ found in [4]. In this paper, we showed that PageRank is stable, HITS is not stable, and SALSA is stable under certain circumstances according to our new definition of stability. Also, we reexamined *Randomized HITS* introduced in [9, 7] showing that the algorithm is stable. Also, we proposed *Randomized SALSA* as a way to overcome the limitation of SALSA. Finally, stability of link analysis algorithms were analyzed in practice. Our work leads toward some practical and theoretical open questions to be attacked for future work. Above all, more detailed studies about the real stability aspect of link analysis algorithms will be required. Also, it would be interesting to extend/refine the notion of similarity between two link analysis algorithms in [3], and apply this revised notion to study the similarity among link analysis algorithms.

References

1. J. Kleinberg, "Authoritive sources in a hyperlinked environment", *Journal of the ACM*, 46 1999.
2. S. Brin and L. Page, "The natomy of a large-scale hypertextual(Web) search engine", *Proc. 7th International World Wide Web Conference*, 1998.
3. A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas, "Finding authorities and hubs from link structures on the world wide web",*Proc. 10th International WWW conference*,2001.
4. A.Ng, A.Zheng and M.Jordan, "Link Analysis, Eigenvectors and Stability", *Proc. 7th International Conference on Artificial Intelligence*, 2001.
5. R. Lempel and S. Moran, "Rank-Stabiltiy and Rank-Similarity of Web Link-Based Ranking Algorithms", Technion CS Department technical report, CS-2001-22, 2001.
6. R. Lempel and S. Moran, The stochastic approach for link-structure analysis(SALSA) and the TKC effect. *Proc. 9th International World Wide Web Conference*, May 2000.
7. A.Ng, A.Zheng and M.Jordan, "Stable Algorithms for Link Analysis". *Proc. 24th ACM-SIGIR Conference on research and development in Information Retrieval* 415-429, May 2001.
8. T. Lindvall, Lectures on the coupling method, wiley series in probability and mathematica statistics, 1992.
9. D. Rafiei and A. Mendelzon, "What is this page known for? Computing web page reputations", *Proc. the 9th International World Wide Web Conference*, Amsterdam, Netherlands, 2000.
10. D. Cohn and H. Chang, "Probabilistically Identifying Authoritative Documents", *Proc. 17th International Conference on Machine Learning*, 2000.
11. K. Bharat and M. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment", *Proc. of 21st International Conference on Research and Development in Information Retrieval (SIGIR 1998).*
12. S. Chakrabarti, M. Joshi and V. Tawde, "Enhanced topic distillation using text, markup tags, and hyperlinks", *Proc. 24th ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2001.
13. D. Achilioptas, A. Fiat, A. Karlin, and F. McSherry, "Web search through hub synthesis", *Proc. 42nd Foundation of Computer Science*, Las Vegas, Nevada, 2001.
14. Y. Azar, A. Fiat, A. Karlin, F. McSherry, J. Saia, "Spectral analysis of data", *Proc. 33rd Symposium on Theory of Computing*, Hersonissos, Crete, Greece, 2001.
15. T. Haveliwala, "Efficient Computation of PageRank", Technical report, Stanford University Database Group, 1999.
16. S. Meyn and R. Tweedie, "Markov chains and stochastic stability", Springer, 1993.

8 Appendix

Proof of proposition 2

We start the proof with the introduction of some additional notation and definitions. Extending notations from Section 2, let $|B(i)|$ denote the number of backward-links of page i, and let $|\tilde{B}(i)|$ denote the number of backward-links of page i after the perturbation. Denote by w the sum of all backward-links, and by $\tilde{w}$ the sum of all backward-links after the perturbation. Note that a page is perturbed by the addtion/removal of backwar-links to it. Thus, depending on the number of added/removed links, total number of backward-links to the page is either increased or decreased or remains the same. Let X denote the set of pages whose number of backward-links remains the same even after the perturbation. Let IBP denote the set of pages whose number of backward-links is increased. Moreover, for each node i ∈ IBP, let $K_i = |\tilde{B}(i)| - |B(i)|$ be its increase in the number of backward-links. Similarly, let DBP denote the set of pages whose number of backward-links is decreased. For each node i ∈ DBP, let $C_i = |B(i)| - |\tilde{B}(i)|$ be its decrease in the number of backward-links. From [6],

$$\sum_{i \in BP} \frac{|B(i)|}{w} + \sum_{i \in X} \frac{|B(i)|}{w} = \sum_{i \in V} \frac{|B(i)|}{w} = \sum_{i \in V} \frac{|\tilde{B}(i)|}{\widetilde{w}} = \sum_{i \in BP} \frac{|\tilde{B}(i)|}{\widetilde{w}} + \sum_{i \in X} \frac{|\tilde{B}(i)|}{\widetilde{w}} (\ddagger) \tag{2}$$

holds on the class of authority connected graphs $\mathcal{G}_{\mathcal{AC}}$. We have

$$||a - \widetilde{a}||_1 = \sum_{i \in V} \left| \frac{|B(i)|}{w} - \frac{|\tilde{B}(i)|}{\widetilde{w}} \right| = \sum_{i \in IBP} \left| \frac{|B(i)|}{w} - \frac{|\tilde{B}(i)|}{\widetilde{w}} \right| + \sum_{i \in X} |B(i)| \left| \frac{1}{w} - \frac{1}{\widetilde{w}} \right| + \sum_{i \in DBP} \left| \frac{|B(i)|}{w} - \frac{|\tilde{B}(i)|}{\widetilde{w}} \right|$$

To evaluate $||a - \tilde{a}||_1$, we consider two separate cases.

- **Case 1** ($\widetilde{w} \geq w$): since $1/\tilde{w} \leq 1/w$, the expression becomes $\sum_{i \in IBP} ||\tilde{B}(i)|/\widetilde{w} - |B(i)|/w| + \sum_{i \in X} |B(i)|(1/\widetilde{w} - 1/w) + \sum_{i \in DBP}(|B(i)|/w - |\tilde{B}(i)|/\widetilde{w})$. From (2), we know $\sum_{i \in X} |B(i)|(1/w - 1/\widetilde{w}) = \sum_{i \in BP}(|\tilde{B}(i)|/\tilde{w} - |B(i)|/w)$. Thus, the expression is simplified into $||a - \widetilde{a}||_1 = \sum_{i \in IBP} ||B(i)|/w - |\tilde{B(i)}|/\widetilde{w}| + \sum_{i \in IBP}(|\tilde{B}(i)|/\widetilde{w} - |B(i)|/w)$. We split up the set IBP into two subsets IBP^+ and IBP^-. Let IBP^+ be the set of nodes in IBP such that $|B(i)|/w > |\tilde{B}(i)|/\tilde{w}$, and let IBP^- be the set of nodes in IBP such that $|B(i)|/w < |\tilde{B}(i)|/\tilde{w}$. Therefore, the expression is further simplified into $\sum_{i \in IBP^+}(|B(i)|/w - |\tilde{B}(i)|/\widetilde{w}) + \sum_{i \in IBP^-}(|\tilde{B}(i)|/\widetilde{w} - |B(i)|/w) + \sum_{i \in IBP}(|\tilde{B}(i)|/\widetilde{w} - |B(i)|/w) = 2\sum_{i \in IBP^-}(|\tilde{B}(i)|/\widetilde{w} - |B(i)|/w) \leq 2\sum_{i \in IBP^-}(|B(i)| + K_i - |B(i)|)/w = 2\sum_{i \in IBP^-} K_i/w \leq 2\sum_{i \in BP} c_i/w$.
- **Case 2** ($\widetilde{w} \leq w$) is analogous to that of case 1, so its proof is skipped and the first part of proposition follows.

Furthermore, if for all $i \in BP$, $|B(i)| > 0$ is held, then $2\sum_{i \in BP} c_i/w \leq 2\sum_{i \in BP} c_i \cdot |B(i)|/w \leq 2\sum_{i \in BP} c_i \cdot a_i$ which proves the rest of proposition 2.

Proof for proposition 6

The proof is based on the coupling method (see [8] for details). First, some notations are introduced. Given i $\in$ FP, we denote by W(i) the set of those nodes that became pointed to by i just after the perturbation(added nodes), by Z(i) the set of nodes that are not pointed any more after the perturbation(removed nodes), and finally by N(i) the set of nodes that remain pointed even after the perturbation (unperturbed nodes). Note $|F(i)| = |N(i)| + |Z(i)|$ and $|\widetilde{F}(i)| = |N(i)| + |W(i)|$. Now, we construct coupled markov chains $\{(X_t, Y_t) : t \geq 0\}$ over pairs of web pages. $X_0 = Y_0$ is drawn according to the probability vector **a** of $RSALSA$ on graph G. On step t, we "reset" both chains with probability ϵ, in which case we reset both X_t and Y_t to the same page chosen uniformly at random. If "no reset" occurs at time t with probability $1 - \epsilon$,and $X_{t-1} = Y_{t-1}$,$X_{t-1} \in$ FP, then normal SALSA steps are performed independently on each graph to choose X_t and Y_t, i.e. follow uniformly at random a backward-link, then a forward-link on G to select X_t. Similarly, follow uniformly at random a backward-link,then a forward-link on G' to select Y_t. If "no reset" occurs at time t, and $X_{t-1} = Y_{t-1}$, $X_{t-1} \in BU$, then the selection of X_t and Y_t is made in two steps: 1) follow uniformly at random a common backward-link that point to $X_{t-1} = Y_{t-1}$, say i. 2) if i $\in$ FU, then follow uniformly at random a forward-link setting $X_t = Y_t$. Otherwise, if i $\in$ FP, then we consider several subcases to select X_t and Y_t according to the following rules. 1) (set selection): in this step, we select pair of sets before the actual selection of nodes for X_t and Y_t. The pair N(i) $\subset G$ and N(i) $\subset G'$ are chosen with probability $|N(i)|^2/|F(i)||\widetilde{F}(i)|$. The pair N(i) $\subset \mathcal{G}$ and W(i) $\subset \tilde{\mathcal{G}}$ are chosen with probability $|N(i)||W(i)|/|F(i)||\widetilde{F}(i)|$. The pair Z(i) $\subset G$ and N(i) $\subset G'$ are chosen with probability $|Z(i)||N(i)|/|F(i)||\widetilde{F}(i)|$. The pair Z(i) $\subset G$ and W(i) $\subset G'$ are chosen with probability $|Z(i)||W(i)|/|F(i)||\widetilde{F}(i)|$. 2) (node selection): if N(i) $\subset G$ and N(i) $\subset G'$ are selected, then we choose uniformly at random a node l $\in$N(i) setting $X_t = Y_t = l$. In other cases, X_t will be selected uniformly at random either from Z(i) if Z(i) $\in G$ was selected in the first step or from N(i) if N(i) $\in G$ was selected in the first step. Similarly, Y_t will be selected uniformly at random either from W(i) if W(i) $\in G'$ was selected in the previous step or from N(i) $\in G'$ if N(i) was selected in the first step.

By this construction, two coupled markov chains X_t and Y_t are created with (a, $\tilde{a}$) as their asymptotic distributions. With little bit of work, it is possible to show that $P(X_{t+1} \neq Y_{t+1}, X_t = Y_t, X_t \in BU|\text{no reset at t+1}) \leq \sum_{i \in BP} c_i a_i$ (§) where c_i denotes the number of perturbed backward-links that are pointing to i. Then, we have

$$P(X_{t+1} \neq Y_{t+1}) = (1-\epsilon) \cdot P(X_{t+1} \neq Y_{t+1}, X_t \neq Y_t|\text{nr(t+1)}) + (1-\epsilon) \cdot P(X_{t+1} \neq Y_{t+1}, X_t = Y_t|\text{nr(t+1)})$$
$$\leq (1-\epsilon) \cdot ((P(X_t \neq Y_t|\text{nr(t+1)} + P(X_{t+1} \neq Y_{t+1}, X_t = Y_t, X_t \in BU|\text{nr(t+1)})) + P(X_{t+1} \neq Y_{t+1}, X_t = Y_t$$
$$, X_t \in BP|\text{nr(t+1)})) \leq (1-\epsilon) \cdot (P(X_t \neq Y_t) + P(X_{t+1} \neq Y_{t+1}, X_t = Y_t, X_t \in BU|\text{nr(t+1)})$$
$$+ P(X_t \in BP|nr(t+1)))$$

Using (§), we have $P(X_t \neq Y_t) \leq (1-\epsilon) \cdot (P(X_t \neq Y_t) + \sum_{i \in BP} c_i a_i + \sum_{i \in BP} a_i) \leq (1-\epsilon) \cdot (P(X_t \neq Y_t) + 2 \cdot \sum_{i \in BP} c_i a_i)$. From $P(X_0 \neq Y_0) = 0$, $P(X_\infty \neq Y_\infty) \leq \frac{2 \cdot (1-\epsilon)}{\epsilon} \sum_{i \in BP} c_i a_i$ is obtained, and applying the coupling lemma the proposition follows.