SALSA: The Stochastic Approach for Link-Structure Analysis *

R. Lempel S. Moran†
Department of Computer Science
The Technion,
Haifa 32000, Israel
email: {rlempel,moran}@cs.technion.ac.il

February 2001

ABSTRACT

Today, when searching for information on the WWW, one usually performs a query through a term-based search engine. These engines return, as the query's result, a list of Web pages whose contents matches the query. For broad topic queries, such searches often result in a huge set of retrieved documents, many of which are irrelevant to the user. However, much information is contained in the link-structure of the WWW. Information such as which pages are linked to others can be used to augment search algorithms. In this context, Jon Kleinberg introduced the notion of two distinct types of Web-pages: *hubs* and *authorities*. Kleinberg argued that hubs and authorities exhibit a *mutually reinforcing relationship*: A good hub will point to many authorities, and a good authority will be pointed at by many hubs. In light of this, he devised an algorithm aimed at finding authoritative pages.

We present SALSA - a new stochastic approach for link structure analysis, which examines random walks on graphs derived from the link structure. We show that both SALSA and Kleinberg's Mutual Reinforcement approach

*This paper extends a previous work by the same authors, titled "The Stochastic Approach for Link Structure Analysis (SALSA) and the TKC Effect", which appeared in the Ninth International World Wide Web Conference, Amsterdam, 2000.

†The research of this author was supported by the fund for promoting research in the Technion, and by the Bernard Elkin Chair in Computer Science.

employ the same meta-algorithm. We then prove that SALSA is equivalent to a weighted in-degree analysis of the link-structure of WWW subgraphs, making it computationally more efficient than the Mutual Reinforcement approach.

We compare the results of applying SALSA to the results derived through Kleinberg's approach. These comparisons reveal a topological phenomenon called the *TKC Effect* which, in certain cases, prevents the Mutual Reinforcement approach from identifying meaningful authorities.

Keywords: Information Retrieval; Link Structure Analysis; Hubs and Authorities; Random walks; SALSA.

1 Introduction

Searching the WWW - The Challenge The WWW is a rapidly expanding hyperlinked collection of unstructured information. The lack of structure and the enormous volume of the WWW pose tremendous challenges on the WWW Information Retrieval systems called search engines. These search engines are presented with queries, and return a list of Web-pages which are deemed (by the engine) to pertain to the query.

When considering the difficulties which WWW search engines face, we distinguish between narrow-topic queries and broad-topic queries. This distinction pertains to the presence which the query's topic has on the Web:
Narrow topic queries are queries for which very few resources exist on the Web, and which present a "needle in the haystack" challenge for search engines. An example of such a query is an attempt to locate the lyrics of a specific song, by quoting a line from it ("We all live in a yellow submarine"). Search engines encounter a *recall* challenge when handling such queries - Finding the few resources which pertain to the query.
On the other hand, broad-topic queries pertain to topics for which there is an abundance of information on the Web, sometimes as many as millions of relevant resources (with varying degrees of relevance). The vast majority of users are not interested in retrieving the entire huge set of resources - most users will be quite satisfied with a few *authoritative* results: Web pages which are highly relevant to the topic of the query, significantly more than most other pages. The challenge which search engines face here is one of *precision* - Retrieving only the most relevant resources to the query.

This work focuses on finding authoritative resources which pertain to broad-topic queries.

Term-based search engines Term-based search engines face both classical problems in Information Retrieval, as well as problems specific to the WWW setting, when handling broad-topic queries. The classic problems include the following issues ([Papadimitriou et al. 1998], [Chakrabarti et al. 1999a]):

- Synonymy - Retrieving documents containing the term "car" when given the query "automobile".
- Polysemy/Ambiguity - When given the query "Jordan", should the engine retrieve pages pertaining to the Hashemite Kingdom of Jordan, or pages pertaining to basketball legend Michael Jordan?

- Authorship styles - This is a generalization of the synonymy issue. Two documents, which pertain to the same topic, can sometimes use very different vocabularies and figures of speech when written by different authors (as an example, the styles of two documents, one written in British English and the other in American English, might differ considerably).

In addition to the classical issues in Information Retrieval, there is a Web-specific obstacle which search engines must overcome, called *search engine persuasion* ([Marchiori 1997]). There may be millions of sites pertaining in some manner to broad-topic queries, but most users will only browse through the first ten results returned by their favorite search facility. With the growing economic impact of the WWW, and the growth of e-commerce, it is crucial for businesses to have their sites ranked high by the major search engines. There are quite a few companies who sell this kind of expertise - They design Web sites which are tailored to rank high with specific queries on the major search engines. These companies (which call their business "search engine *optimization/positioning*") research the ranking algorithms and heuristics of term-based engines, and know how many keywords to place (and where) in a Web-page so as to improve the page's ranking (which directly impacts the page's visibility). A less sophisticated technique, used by some site creators, is called *keyword spamming* ([Chakrabarti et al. 1999a]). Here, the authors repeat certain terms (some of which are only remotely connected to their site's context), in order to "lure" search engines into ranking them highly for many queries.

Informative link structure - The answer? The WWW is a hyperlinked collection. In addition to the textual content of the individual pages, the link structure of such collections contains information which can, and should, be tapped when searching for authoritative sources. Consider the significance of a link $p \rightarrow q$: With such a link p suggests, or even recommends, that surfers visiting p follow the link and visit q. This may reflect the fact that pages p and q share a common topic of interest, and that the author of p thinks highly of q's contents. Such a link, called an *informative link*, is p's way to confer authority on q ([Kleinberg 1998]). Note that informative links provide a positive critical assessment of q's contents which originates from outside the control of the author of q (as opposed to assessments based on q's textual content, which is under complete control of q's author). This makes the information extracted from informative links less vulnerable to manipulative techniques such as spamming.

Unfortunately, not all links are informative. There are many kinds of links which confer little or no authority ([Chakrabarti et al. 1999a]), such as intra-domain (inner) links (whose purpose is to provide navigational aid in a complex Web site of some organization) and advertisements/sponsorship links. Another kind of non-informative links are those which result from link-exchange (also called reciprocal links) agreements. These are bidirectional links between two Web pages, whose purpose is to increase the visibility and link popularity of both pages.

As more and more search engines have incorporated link structure analysis into their ranking schemes, many search engine optimization firms have added *link development* services to their other Web site design services ([LinkMe],[Grantastic],[Canyontrace]). These services help customers to find link-exchange partners and to get listed by major directory services, such as Yahoo! ([Yahoo!]).

We stress here that a crucial task which should be completed prior to analyzing the link structure of a given collection, is to filter out as many of the non-informative links as possible.

Related work on link structures Prior to the introduction of hypertext, link structures were studied in the area of bibliometrics, which studies the citation structure of written documents ([Small 1973],[Kessler 1963]). Many works in this area were aimed at finding high-impact papers published in scientific journals ([Garfield 1972]), and at clustering related documents ([Auguston and Minker 1970]).

When hypertext was introduced, it was widely used to present highly structured information (reference books, manuals, etc.) in a flexible computer format which supported browsing. Botafogo et al. ([Botafogo et al. 1992]) provided authors of such hypertexts with tools and metrics (based on the link structure of the hypertexts) to analyze the hierarchical structure of their documents during the authoring phase. Frisse ([Frisse 1988]) proposed a new information retrieval scheme for tree hypertext structures, in which the relevancy of each hypertext node to a given query depends upon the node's textual contents as well as on the relevancy of its descendants.

The advent of the World Wide Web presented many new research directions involving link structure analysis. Some works have studied the Web's link structure, in addition to the textual content of the pages, as means to visualize areas thought to contain good resources ([Carriére and Kazman 1997]). Other works used link structures for categorizing pages and clustering them ([Weiss et al. 1996],[Pirolli et al. 1996]).

Marchiori, in [Marchiori 1997], uses the link-structure of the Web to enhance search results of term-based search engines. This is done by considering the potential *hyper-information* contained in each Web-page: The information that can be found when following hyperlinks which originate in the page.

This work is motivated by the approach introduced by Jon Kleinberg ([Kleinberg 1998]). In an attempt to impose some structure on the chaotic WWW, Kleinberg distinguished between two types of Web-pages which pertain to a certain topic: The first are *authoritative* pages in the sense described previously. The second type are *hub* pages. Hubs are primarily resource lists, linking to many authorities on the topic possibly without directly containing the authoritative information. According to this model, hubs and authorities exhibit a *mutually reinforcing relationship*: Good hubs point to many good authorities, and good authorities are pointed at by many good hubs. In light of the mutually reinforcing relationship, hubs and authorities should form communities, which can be pictured as dense bipartite portions of the Web, where the hubs link densely to the authorities. The most prominent community in a WWW subgraph is called the *principal community* of the collection. Kleinberg suggested an algorithm to identify these communities, which is described in detail in section 2.

Researchers from IBM's Almaden Research Center have implemented Kleinberg's algorithm in various projects. The first was *HITS*, which is described in [Gibson et al. 1998], and offers some enlightening practical remarks. The *ARC* system, described in [Chakrabarti et al. 1998b], augments Kleinberg's link-structure analysis by considering also the anchor text, the text which surrounds the hyperlink in the pointing page. The reasoning behind this is that many times, the pointing page describes the destination page's contents around the hyperlink, and thus the authority conferred by the links can be better assessed. These projects were extended by the *CLEVER* project ([CLEVER]). Researchers from outside IBM, such as Henzinger and Bharat, have also studied Kleinberg's approach and have proposed improvements to it ([Bharat and Henzinger 1998]).

Anchor text was also used by Brin and Page in [Brin and Page 1998]. Another major feature of their work on the *Google* search engine ([Google]) is a link-structure based ranking approach called *PageRank*, which can be interpreted as a stochastic analysis of some random-walk behavior through the entire WWW. See section 4 for more details on *PageRank*.

In [Law et al. 1999], the authors use the links surrounding a small set of same-topic sites to assemble a larger collection of neighboring pages which should contain many authoritative resources on the initial topic. The textual content of the collection is then analyzed in ranking the relevancy of its individual pages.

This work While preserving the theme that Web pages pertaining to a given topic should be split to hubs and authorities, we replace Kleinberg's Mutual Reinforcement approach ([Kleinberg 1998]) by a new stochastic approach (SALSA), in which the coupling between hubs and authorities is less tight. The intuition behind our approach is the following: consider a bipartite graph G, whose two parts correspond to hubs and authorities, where an edge between hub r and authority s means that there is an informative link from r to s. Then, authorities and hubs pertaining to the dominant topic of the pages in G should be highly visible (reachable) from many pages in G. Thus, we will attempt to identify these pages by examining certain random walks in G, under the proviso that such random walks will tend to visit these highly visible pages more frequently than other, less connected pages. We show that in finding the principal communities of hubs and authorities, both Kleinberg's Mutual Reinforcement approach and our Stochastic approach employ the same meta-algorithm on different representations of the input graph. We then compare the results of applying SALSA to the results derived by Kleinberg's approach. Through these comparisons, we isolate a particular topological phenomenon which we call the *Tightly Knit Community (TKC) Effect*. In certain scenarios, this effect hampers the ability of the the Mutual Reinforcement approach to identify meaningful authorities. We demonstrate that SALSA is less vulnerable to the TKC effect, and can find meaningful authorities in collections where the Mutual Reinforcement approach fails to do so.

After demonstrating some results achieved by means of SALSA, we prove that the ranking of pages in the Stochastic approach may be calculated by examining the weighted in/out degrees of the pages in G. This result yields that SALSA is computationally lighter than the Mutual Reinforcement approach. We also discuss the reason for our success with analyzing weighted in/out degrees of pages, which previous work has claimed to be unsatisfactory for identifying authoritative pages.

The rest of the paper is organized as follows: Section 2 recounts Kleinberg's Mutual Reinforcement Approach. In section 3 we view Kleinberg's approach from a higher level, and define a meta-algorithm for link structure analysis. Section 4 presents our new approach, SALSA. In section 5 we compare the two approaches by considering their outputs on the WWW and on artificial topologies. Then, in section 6 we prove the connection between SALSA and weighted in/out degree rankings of pages. Our conclusions and ideas for future work are brought in section 7.

The paper uses basic results from the theories of non-negative matrices and of stochastic processes. The required mathematical background, as well as the proofs of the propositions which appear in Section 5.1, can be found in [Lempel and Moran 2000].

2 Kleinberg's Mutual Reinforcement Approach

The Mutual Reinforcement approach ([Kleinberg 1998]) starts by assembling a collection $\mathcal{C}$ of Web-pages, which should contain communities of hubs and authorities pertaining to a given topic t. It then analyzes the link structure induced by that collection, in order to find the authoritative pages on topic t.

Denote by q a term-based search query to which pages in our topic of interest t are deemed to be relevant. The collection $\mathcal{C}$ is assembled in the following manner:

- A *root set* S of pages is obtained by applying a term based search engine, such as AltaVista [AltaVista], to the query q. This is the only step in which the lexical content of the Web pages is examined.

- From S we derive a *base set* $\mathcal{C}$ which consists of (a) pages in the root set S, (b) pages which point to a page in S and (c) pages which are pointed to by a page in S. In order to obtain (b), we must again use a search engine. Many search engines store linkage information, and support queries such as "which pages point to [a given url]".

The collection $\mathcal{C}$ and its link structure induce the following directed graph G: G's nodes are the pages in $\mathcal{C}$, and for all $i, j \in \mathcal{C}$, the directed edge $i \to j$ appears in G if and only if page i contains a hyperlink to page j. Let W denote the $|\mathcal{C}| \times |\mathcal{C}|$ adjacency matrix of G.

Each page $s \in \mathcal{C}$ is now assigned a pair of weights, a hub-weight $h(s)$ and an authority weight $a(s)$, based on the following two principles:

- The quality of a hub is determined by the quality of the authorities it points at. Specifically, a page's hub weight should be proportional to the sum of the authority weights of the pages it points at.

- "Authority lies in the eyes of the beholder(s)": A page is authoritative only if good hubs deem it as such. Specifically, a page's authority weight is proportional to the sum of the hub-weights of the pages pointing at it.

The top ranking pages, according to both kinds of weights, form the Mutually Reinforcing communities of hubs and authorities. In order to assign such weights, Kleinberg uses the following iterative algorithm:

1. Initialize $a(s) \leftarrow 1, h(s) \leftarrow 1$ for all pages $s \in \mathcal{C}$.

2. Repeat the following three operations until convergence:

   - Update the authority weight of each page s (the $\mathcal{I}$ operation): $a(s) \leftarrow \sum_{\{x|x \text{ points to } s\}} h(x)$
   - Update the hub weight of each page s (the $\mathcal{O}$ operation): $h(s) \leftarrow \sum_{\{x|s \text{ points to } x\}} a(x)$
   - Normalize the authority weights and the hub weights.

Note that applying the $\mathcal{I}$ operation is equivalent to assigning authority weights according to the result of multiplying the vector of all hub weights by the matrix W^T. The $\mathcal{O}$ operation is equivalent to assigning hub weights according to the result of multiplying the vector of all authority weights by the matrix W.

Kleinberg showed that this algorithm converges, and that the resulting authority weights [hub weights] are the coordinates of the normalized principal eigenvector [1] of W^TW [of WW^T]. W^TW and WW^T are well known matrices in the field of bibliometrics:

1. $A \triangleq W^TW$ is the *co-citation matrix* ([Small 1973]) of the collection. $[A]_{i,j}$ is the number of pages which jointly point at (cite) pages i and j. Kleinberg's iterative algorithm converges to authority weights which correspond to the entries of the (unique, normalized) principal eigenvector of A.

[1]The eigenvector which corresponds to the eigenvalue of highest magnitude of the matrix.

2. $H \triangleq WW^T$ is the *bibliographic coupling matrix* ([Kessler 1963]) of the collection. $[H]_{i,j}$ is the number of pages jointly referred to (pointed at) by pages i and j. Kleinberg's iterative algorithm converges to hub weights which correspond to the entries of H's (unique, normalized) principal eigenvector.

3 A Meta-Algorithm for Link Structure Analysis

Examining the Mutual Reinforcement approach from a higher level, we can identify a general framework, or meta-algorithm, for finding hubs and authorities by link-structure analysis. This meta-algorithm is a version of the spectral filtering method, presented in [Chakrabarti et al. 1998a]:

- Given a topic t, construct a page collection $\mathcal{C}$ which should contain many t-hubs and t-authorities, but should not contain many hubs or authorities for any other topic t'. Let $n = |\mathcal{C}|$.

- Derive, from $\mathcal{C}$ and the link structure induced by it, two $n \times n$ association matrices - A *hub matrix* H and an *authority matrix* A. Association matrices are widely used in classification algorithms ([VanRijsbergen 1979]), and will be used here in order to classify the Web pages into communities of hubs/authorities. The association matrices which are used by the meta-algorithm will have the following algebraic property (let M denote such a matrix):

  M will have a unique real positive eigenvalue $\lambda(M)$ of multiplicity 1, such that for any other eigenvalue λ' of M, $\lambda(M) > |\lambda'(M)|$. Denote by $v_{\lambda(M)}$ the (unique) unit eigenvector which corresponds to $\lambda(M)$ whose first non-zero coordinate is positive. $v_{\lambda(M)}$ will actually be a positive vector, and will be referred to as the *principal eigenvector* of M.

- For some user-defined integer $k < n$, the k pages that correspond to the largest coordinates of $v_{\lambda(A)}$ will form the *principal algebraic community of authorities* in $\mathcal{C}$. The *principal algebraic community of hubs* in $\mathcal{C}$ is defined similarly.

For the meta-algorithm to be useful, the algebraic principal communities of hubs and authorities should reflect the true authorities and hubs in $\mathcal{C}$.

The two degrees of freedom which the meta-algorithm allows, are the method for obtaining the collection, and the definition of the association matrices. Given a specific collection, the algebraic communities produced by the meta-algorithm are determined solely by the definition of the association matrices.

4 SALSA: Analyzing a Random Walk on the Web

In this section we introduce *SALSA*, the *Stochastic Approach for Link Structure Analysis*. The approach is based upon the theory of Markov chains, and relies on the stochastic properties of random walks performed on our collection of pages. It follows the meta-algorithm described in section 3, and differs from the Mutual Reinforcement approach in the manner in which the association matrices are defined.

The input to our scheme consists of a collection of pages $\mathcal{C}$ which is built around a topic t in the manner described in section 2. Intuition suggests that authoritative pages on topic t should be visible from many pages in the subgraph induced by $\mathcal{C}$. Thus, a random walk on this subgraph will visit t-authorities with high probability. We combine the theory of random walks with the notion of the two distinct types of Web pages, hubs and authorities, and actually analyze two different Markov chains: A chain of hubs and a chain of authorities. Unlike "conventional" random walks on graphs, state transitions in these chains are generated by traversing <u>two</u> WWW-links in a row, one link forward and one link backwards (or vice versa). Analyzing both chains allows our approach to give each Web page two distinct scores, a hub score and an authority score.

The idea of ranking Web pages using random walks is not new. The search engine *Google* ([Brin and Page 1998],[Google]) incorporates stochastic information into its ranking of pages by assigning each page p a rank of its importance, called *PageRank*. Specifically, the PageRank of a page p is the probability of visiting p in a random walk of the entire Web, where the set of states of the random walk is the set of pages, and each random step is of one of the following two types:

1. From the given state s, choose at random an outgoing link of s, and follow that link to the destination page.

2. Choose a Web page uniformly at random, and jump to it.

PageRank chooses a parameter d, $0 < d < 1$, and each state transition is of the first transition type with probability d, and of the second type with probability $1-d$. The PageRanks obey the following formula (where page p has incoming links from $q_1, \ldots, q_k$):

$$\text{PageRank}(p) = (1-d) + d(\sum_{i=1}^{k} \frac{\text{PageRank}(q_i)}{\text{out degree of } q_i})$$

Since *PageRank* examines a single random walk on the entire WWW, the ranking of Web pages in *Google* is independent of the search query (a global ranking), and no distinction is made between hubs and authorities.

Formal Definition of SALSA Let us build a bipartite undirected graph $\tilde{G} = (V_h, V_a, E)$ from our page collection $\mathcal{C}$ and its link structure:

- $V_h = \{s_h \mid s \in \mathcal{C}$ and $out - degree(s) > 0\}$ (the *hub-side* of $\tilde{G}$).
- $V_a = \{s_a \mid s \in \mathcal{C}$ and $in - degree(s) > 0\}$ (the *authority-side* of $\tilde{G}$).
- $E = \{(s_h, r_a) \mid s \to r \ in \ \mathcal{C}\}$

Each non-isolated page $s \in \mathcal{C}$ is represented in $\tilde{G}$ by one or both of the nodes s_h and s_a. Each WWW-link $s \to r$ is represented by an undirected edge connecting s_h and r_a. Figure 1 shows a construction of a bipartite graph from a given collection:

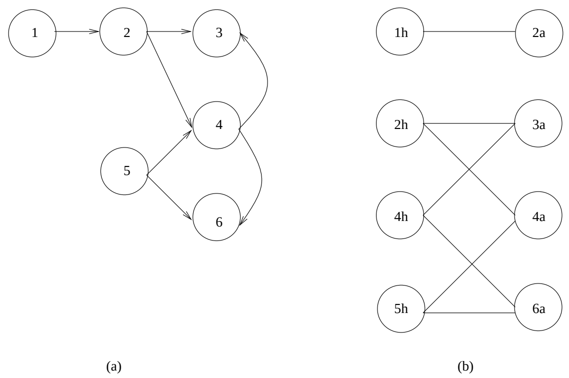

Figure 1: Transforming (a) the collection $\mathcal{C}$ into (b) a bipartite graph $\tilde{G}$

On this bipartite graph we will perform two distinct random walks. Each walk will only visit nodes from one of the two sides of the graph, by traversing paths consisting of two $\tilde{G}$-edges in each step. Since each edge crosses sides of $\tilde{G}$, each walk is confined to just one of the graph's sides, and the two walks will naturally start off from different sides of $\tilde{G}$. Note also that every path of length 2 in $\tilde{G}$ represents a traversal of one WWW link in the proper direction (when passing from the hub-side of $\tilde{G}$ to the authority-side), and a retreat along a WWW link (when crossing in the other direction). Since the hubs and authorities of topic t should be highly visible in $\tilde{G}$ (reachable from many nodes by either a direct edge or by short paths), we may expect that the t-authorities will be amongst the nodes most frequently visited by the random walk on V_a, and that the t-hubs will be amongst the nodes most frequently visited by the random walk on V_h.

We will examine the two different Markov chains which correspond to these random walks: The chain of the visits to the authority side of $\tilde{G}$ (the *authority chain*), and the chain of visits to the hub side of $\tilde{G}$. Analyzing these chains separately naturally distinguishes between the two aspects of each page.

We now define two stochastic matrices, which are the transition matrices of the two Markov chains at interest:

1. *The hub-matrix* $\tilde{H}$, defined as follows :

$$\tilde{h}_{i,j} = \sum_{\{k|(i_h,k_a),(j_h,k_a)\in\tilde{G}\}} \frac{1}{deg(i_h)} \cdot \frac{1}{deg(k_a)}$$

2. *The authority-matrix* $\tilde{A}$, defined as follows :

$$\tilde{a}_{i,j} = \sum_{\{k|(k_h,i_a),(k_h,j_a)\in\tilde{G}\}} \frac{1}{deg(i_a)} \cdot \frac{1}{deg(k_h)}$$

A positive transition probability $\tilde{a}_{i,j} > 0$ implies that a certain page k points to both pages i and j, and hence page j is reachable from page i by two steps: retracting along the link $k \to i$ and then following the link $k \to j$.

Alternatively, the matrices $\tilde{H}$ and $\tilde{A}$ can be defined as follows: Let W be the adjacency matrix of the directed graph defined by $\mathcal{C}$ and its link structure. Denote by W_r the matrix which results by dividing each nonzero entry of W by the sum of the entries in its row, and by W_c the matrix which results by dividing each nonzero element of W by the sum of the entries in its column (Obviously, the sums of rows/columns which contain nonzero elements are greater than zero). Then $\tilde{H}$ consists of the non-zero rows and columns of $W_r W_c^T$, and $\tilde{A}$ consists of the non-zero rows and columns of $W_c^T W_r$. We ignore the rows and columns of $\tilde{A}, \tilde{H}$ which consist entirely of zeros, since (by definition) all the nodes of $\tilde{G}$ have at least one incident edge. The matrices $\tilde{A}$ and $\tilde{H}$ serve as the association matrices required by the meta-algorithm for identifying the authorities and hubs. Recall that the Mutual Reinforcement approach uses the association matrices $A \triangleq W^T W$ and $H \triangleq W W^T$.

We shall assume that $\tilde{G}$ is connected, causing both stochastic matrices $\tilde{A}$ and $\tilde{H}$ to be *irreducible*. This assumption does not form a limiting factor, since when $\tilde{G}$ is not connected, we may use our technique on each connected component separately. Section 6.1 further elaborates on the case when $\tilde{A}$ and $\tilde{H}$ have multiple irreducible components.

Some properties of $\tilde{H}$ and $\tilde{A}$:

- Both matrices are primitive, since the Markov chains which they represent are aperiodic: When visiting any authority(hub), there is a positive probability to revisit it on the next entry to the authority(hub) side of the bipartite graph. Hence, every state (=page) in each of the chains has a self-loop, causing the chains to be aperiodic.

- The adjacency matrix of the support graph of $\tilde{A}$ is symmetric, since $\tilde{a}_{i,j} > 0$ implies $\tilde{a}_{j,i} > 0$. Furthermore, $\tilde{a}_{i,j} > 0 \iff [W^T W]_{i,j} > 0$ (and the same is also true of $\tilde{H}$ and $W W^T$).

Following the framework of the meta-algorithm, the principal community of authorities(hubs) found by the SALSA will be composed of the k pages having the highest entries in the principal eigenvector of $\tilde{A}$ ($\tilde{H}$), for some user defined k. By the Ergodic Theorem ([Gallager 1996]), the principal eigenvector of an irreducible, aperiodic stochastic matrix is actually the stationary distribution of the underlying Markov chain, and its high entries correspond to pages most frequently visited by the (infinite) random walk.

5 Results

In this section we present some combinatorial and experimental results, which compare the Mutual Reinforcement and SALSA approaches. An emphasis is given to the *Tightly Knit Community* effect, which is described in the following subsection.

5.1 The Tightly-Knit Community (TKC) Effect

A tightly-knit community is a small but highly interconnected set of pages. Roughly speaking, the *TKC effect* occurs when such a community scores high in link-analyzing algorithms, even though the pages in the TKC are not authoritative on the topic, or pertain to just one aspect of the topic. Our study indicates that the Mutual Reinforcement approach is vulnerable to this effect, and will sometimes rank the pages of a TKC in unjustified high positions.

In this section we provide a combinatorial construction of an infinite number of topologies in which the TKC effect is demonstrated. For all $k \geq 3$, we will build a collection $\mathcal{C}_k$ which contains two communities: A community C_s, with a small number of hubs and authorities, in which every hub points to all of the authorities; and a much larger community C_l, in which the hubs point only to a portion of the authorities. The topic covered by C_l appears to be the dominant topic of the collection, and is probably of wider interest on the WWW. Since there are many C_l-authoritative pages, the hubs do not link to all of them, whereas the smaller C_s community is densely interconnected. The TKC effect occurs when the pages of C_s are ranked higher than those of C_l, as will happen with the Mutual Reinforcement approach (SALSA ranks the C_l authorities higher).

Formally, for any $k \geq 3$, the collection $\mathcal{C}_k$ has the following structure:

- There are $n \triangleq (k+1)^2$ authorities in the large community, C_l.
- There are $m \triangleq (k+1)$ authorities in the small community, C_s.
- There are $h_l \triangleq \binom{n}{k}$ hubs in the large community. Each such hub covers (links to) a unique subset of k C_l-authorities.
- There are $h_s \triangleq \binom{n-1}{k-1} - n$ hubs in the small community, and each of them links to *all* of the C_s-authorities.
- There are also $n \cdot m$ noisy hubs $g_{1,1}, \ldots, g_{n,m}$. Each such hub $g_{i,j}$ links to the C_l-authority i and to the C_s-authority j.

Indeed, the small, tightly-knit community C_s is highly connected: Its hubs and authorities constitute a complete bipartite graph. The large community, C_l, is sparsely connected: Each hub is linked to less than a square root of the number of authorities.

The ratio between the number of hubs and the number of authorities in both communities is roughly the same: We can see this by examining the following ratio:

$$\frac{h_l}{n} / \frac{h_s}{m}$$

first, we note that $h_s = r \cdot \binom{n-1}{k-1}$ for some $0.5 < r < 1$, since

$$\frac{\binom{n-1}{k-1}}{2} < \binom{n-1}{k-1} - n = h_s < \binom{n-1}{k-1}$$

(the left inequality holds for all $k \geq 3$).

Now:

$$\begin{aligned}
\frac{h_l}{n} / \frac{h_s}{m} &= \frac{\binom{n}{k}}{n} / \frac{\binom{n-1}{k-1} - n}{m} \\
&= \frac{\binom{n}{k}}{\binom{n-1}{k-1} - n} \cdot \frac{m}{n} \\
&= \frac{\binom{n}{k}}{r \cdot \binom{n-1}{k-1}} \cdot \frac{k+1}{(k+1)^2} \quad \text{(for some } 0.5 < r < 1\text{)} \\
&= \frac{n}{r \cdot k} \cdot \frac{1}{k+1} \\
&= \frac{(k+1)^2}{r \cdot k(k+1)} \;=\; \frac{k+1}{k} \cdot \frac{1}{r}
\end{aligned}$$

And since $k \geq 3$, $0.5 < r < 1$ we have:

$$1 \;<\; \frac{k+1}{k} \;<\; \frac{h_l}{n} / \frac{h_s}{m} \;<\; 2\frac{k+1}{k} \;\leq\; \frac{8}{3}$$

Hence, both communities have roughly the same ratio of hubs to authorities. In [Lempel and Moran 2000] we prove that:

Proposition 1 *On the collection $\mathcal{C}_k$, SALSA will rank the C_l-authorities above the C_s-authorities.*

While:

Proposition 2 *On the collection $\mathcal{C}_k$, the Mutual Reinforcement approach will rank the C_s-authorities above the C_l-authorities.*

Thus, in this infinite family of collections, the Mutual Reinforcement approach is affected by the TKC effect (its ranking is biased in favor of tightly knit communities).

We now change the collection $\mathcal{C}_k$ by adding some more pages and links to it. Let A_b be any nonempty proper subset of size b of the C_s-authorities $(1 \leq b < m)$. We add to $\mathcal{C}_k$ a new set of hubs, h_b of size $m+1$, all of which point to all of the authorities in A_b. We call the resulting collection $\tilde{\mathcal{C}}_k$. The resulting principal communities of authorities derived by the two approaches will be:

Proposition 3 *On the collection $\tilde{\mathcal{C}}_k$, SALSA will rank the A_b-authorities first, then the C_l-authorities, and finally the authorities of $C_s \setminus A_b$*

Proposition 4 *On the collection $\tilde{\mathcal{C}}_k$, the Mutual Reinforcement approach will rank the A_b-authorities first, then the authorities of $C_s \setminus A_b$, and finally the C_l authorities.*

By these propositions (whose proofs appears in [Lempel and Moran 2000] as well), we see that SALSA blends the authorities from the two communities in $\tilde{\mathcal{C}}_k$, while the Mutual Reinforcement approach still ranks all of the C_s authorities higher than the C_l authorities.

Our constructions above are, of course, artificial. However, they do demonstrate that the Mutual Reinforcement approach is biased towards tightly knit communities, while our intuition suggests that communities of broad topics should be large, but not necessarily tightly knit. Experimental results which seem to support this intuition, and which demonstrate the bias of the Mutual Reinforcement approach towards tightly knit communities on the WWW, are shown in the next section.

We note here that a special case of the TKC effect has been identified by Bharat and Henzinger in [Bharat and Henzinger 1998]. They have dealt with the phenomena of *Mutually Reinforcing Relationships Between Hosts*, in which a single page from host a may contain links to many pages of host b (or, similarly, the page from host a may have many incoming links from pages of host b). Restricting our attention to the first case, the result of such a scenario is mass endorsement of (pages in) host b by the (single) page of host a. Now, if the same linkage pattern occurs in other pages of a, and they all massively endorse host b, the TKC effect can easily occur. The solution presented in [Bharat and Henzinger 1998] was to lower the weights of the links between the page of a to the pages of b. Specifically, if k such links existed, each link was assigned a weight of $1/k$, thus keeping the aggregate sum of weights on those links at 1. Hence, while a page p can link to any number of pages on any number of hosts, p will always endorse each of those hosts with a weight of 1.

This solution can deal with TKCs which involve multiple pages from a small set of hosts. However, in general, TKCs are not limited to cases of mass endorsement between specific pairs of hosts, and often occur when pages from many different hosts are all pointed at by other pages from different hosts.

5.2 The WWW

In the following, we present experimental results of the application of the different approaches on broad-topic WWW queries (both single topic queries and multi-topic queries). We obtained a collection of pages for each query, and then derived the principal community of authorities with both approaches. Three of these queries ("+censorship +net", "java", "abortion") were used by Kleinberg in [Kleinberg 1998], and are brought here for the sake of comparison. All collections were assembled during February, 1999. The root sets were compiled using AltaVista ([AltaVista]), which also provided the linkage information needed for building the base sets.

When expanding the root set to the entire collection, we attempted to filter non-informative links which exist between Web pages. This was done by studying the target URL of each link, in conjuction with the URL of the link's source.

- Following [Kleinberg 1998], we ignored intra-domain links (since these links tend to be navigational aids inside an intranet, and do not confer authority on the link's destination). Our heuristic did not rely solely on an exact match between the hosts of the link's source and target, and was also able to classify links between related hosts (such as "shopping.yahoo.com" and "www.yahoo.com") as being intra-domain.

- We ignored links to *cgi scripts* (as was done in [Brin and Page 1998]). These links are usually easily identified by the path of the target URL (e.g. *http://www.altavista.com/cgi-bin/query?q=car*).

- We tried to identify ad-links and ignore them as well. This was achieved by deleting links that contained certain characters in their

URL (such as '=','?' and others) which appear almost exclusively in advertisements and sponsorship links, and in links to dynamic content.

Overall, 38% of the links we examined were ignored. The collections themselves turn out to be relatively sparse graphs, with the number of edges never exceeding three times the number of nodes. We note that a recent work by Kleinberg et al. ([Kleinberg et al. 1999]) has examined some other connectivity characteristics of such collections.

For each query, we list the top authorities which were returned by the two approaches. The results are displayed in tables containing four columns:

1. The url.
2. The title of the url.
3. The category of the url: (1) denotes a member of the root set (as defined in page 6), (2) denotes a page pointing into the root set, and (3) denotes a page pointed at by a member of the root set.
4. The value of the coordinate of this url in the (normalized) principal eigenvector of the authority matrix.

Single-Topic Query: +censorship +net

For this query, both approaches produced the same top six pages (although in a different order). The results are shown in table 1.

```
Size of root size  = 150, Size of collection = 562
Principal Community, Mutual Reinforcement Approach:
```

url	title	cat	weight
http://www.eff.org/	EFFweb-The Electronic Frontier Foundation	(3)	0.5355
http://www.epic.org/	Electronic Privacy Information Center	(3)	0.3584
http://www.cdt.org/	The Center For Democracy and Technology	(3)	0.3525
http://www.eff.org/ blueribbon.html	Blue Ribbon Campaign For Online Free Speech	(3)	0.2810
http://www.aclu.org/	ACLU: American Civil Liberties Union	(3)	0.2800
http://www.vtw.org/	The Voters Telecommunications Watch	(3)	0.2539

Principal Community, SALSA:

url	title	cat	weight
http://www.eff.org/	EFFweb-The Electronic Frontier Foundation	(3)	0.3848
http://www.eff.org/blueribbon.html	Blue Ribbon Campaign For Online Free Speech	(3)	0.3207
http://www.epic.org/	Electronic Privacy Information Center	(3)	0.2566
http://www.cdt.org/	The Center For Democracy and Technology	(3)	0.2566
http://www.vtw.org/	The Voters Telecommunications Watch	(3)	0.2405
http://www.aclu.org/	ACLU: American Civil Liberties Union	(3)	0.2405

Table 1: Authorities for WWW query "+censorship +net"

Single-Topic Query: Java

The results for this query, with our first example of the TKC effect, are shown in table 2. All of the top ten Mutual Reinforcement authorities are part of the EARTHWEB Inc. network. They are interconnected, but since the domain names of the sites are different, the interconnecting links were not filtered out. Some of the pages are highly relevant to the query (and have many incoming links from sites outside the EarthWeb net), but most appear in the principal community only because of their EarthWeb affiliation. With SALSA, only the top three Mutual Reinforcement authorities are retained, and the other seven are replaced by other authorities, some of which are clearly more related to the query.

Size of root size = 160, Size of collection = 2810
Principal Community, Mutual Reinforcement Approach:

url	title	cat	weight
http://www.jars.com/	EarthWeb's JARS.COM Java Review Service	(3)	0.3341
http://www.gamelan.com/	Gamelan - The Official Java Directory	(3)	0.3036
http://www.javascripts.com/	Javascripts.com - Welcome	(3)	0.2553
http://www.datamation.com/	EarthWeb's Datamation.com	(3)	0.2514
http://www.roadcoders.com/	Handheld Software Development@ RoadCoders	(3)	0.2508
http://www.earthweb.com/	EarthWeb	(3)	0.2494
http://www.earthwebdirect.com/	Welcome to Earthweb Direct	(3)	0.2475
http://www.itknowledge.com/	ITKnowledge	(3)	0.2469
http://www.intranetjournal.com/	intranetjournal.com	(3)	0.2452
http://www.javagoodies.com/	Java Goodies JavaScript Repository	(3)	0.2388

`Principal Community, SALSA:`

url	title	cat	weight
http://java.sun.com/	Java(tm) Technology Home Page	(3)	0.3653
http://www.gamelan.com/	Gamelan - The Official Java Directory	(3)	0.3637
http://www.jars.com/	EarthWeb's JARS.COM Java Review Service	(3)	0.3039
http://www.javaworld.com/	IDG's magazine for the Java community	(3)	0.2173
http://www.yahoo.com/	Yahoo!	(3)	0.2141
http://www.javasoft.com/	Java(tm) Technology Home Page	(3)	0.2031
http://www.sun.com/	Sun Microsystems	(3)	0.1874
http://www.javascripts.com/	Javascripts.com - Welcome	(3)	0.1385
http://www.htmlgoodies.com/	htmlgoodies.com - Home	(3)	0.1307
http://javaboutique.internet.com/	The Ultimate Java Applet Resource	(1)	0.1181

Table 2: Authorities for WWW query "Java"

Single-Topic Query: movies

This query demonstrates the TKC effect on the WWW in a most striking fashion. First, consider the Mutual Reinforcement principal community of authorities, presented in table 3:

`Size of root size  = 175, Size of collection = 4539`

url	title	cat	weight
http://go.msn.com/npl/msnt.asp	MSN.COM	(3)	0.1673
http://go.msn.com/bql/whitepages.asp	White Pages - msn.com	(3)	0.1672
http://go.msn.com/bsl/webevents.asp	Web Events	(3)	0.1672
http://go.msn.com/bql/scoreboards.asp	MSN Sports scores	(3)	0.1672

Table 3: Mutual Reinforcement Authorities for WWW query "movies"

The top 30 authorities returned by the Mutual Reinforcement approach were all *go.msn.com* sites. All but the first received the exact same weight, 0.1672. Recall that we do not allow same-domain links in our collection, hence none of the top authorities was pointed at by a *go.msn.com* page. To understand how these pages scored so well, we turn to the principal community of hubs, shown in table 4:

url	title	cat	weight
http://denver.sidewalk.com/movies	movies: denver.sidewalk	(1)	0.1692
http://boston.sidewalk.com/movies	movies:boston.sidewalk	(1)	0.1691
http://twincities.sidewalk.com/movies	movies: twincities.sidewalk	(1)	0.1688
http://newyork.sidewalk.com/movies	movies: newyork.sidewalk	(1)	0.1686

Table 4: Mutual Reinforcement Hubs for WWW query "movies"

These innocent looking hubs are all part of the *Microsoft Network (msn)*, but when building the basic set we did not identify them as such. All these hubs point, almost without exception, to the entire set of authorities found by the MR approach (hence the equal weights which the authorities exhibit). However, the vast majority of the pages in the collection were not part of this "conspiracy", and almost never pointed to any of the *go.msn.com* sites. Therefore, the authorities returned by the Stochastic approach (table 5) contain none of those *go.msn.com* pages, and are much more relevant to the query:

url	title	cat	weight
http://us.imdb.com/	The Internet Movie Database	(3)	0.2533
http://www.mrshowbiz.com/	Mr Showbiz	(3)	0.2233
http://www.disney.com/	Disney.com–The Web Site for Families	(3)	0.2200
http://www.hollywood.com/	Hollywood Online:...all about movies	(3)	0.2134
http://www.imdb.com/	The Internet Movie Database	(3)	0.2000
http://www.paramount.com/	Welcome to Paramount Pictures	(3)	0.1967
http://www.mca.com/	Universal Studios	(3)	0.1800
http://www.discovery.com/	Discovery Online	(3)	0.1550
http://www.film.com/	Welcome to Film.com	(3)	0.1533
http://www.mgmua.com/	mgm online	(3)	0.1300

Table 5: Stochastic authorities for WWW query "movies"

A similar community is obtained by the Mutual Reinforcement approach, after deleting the rows and columns which correspond to the top 30 authorities from the matrix $W^T W$. This deletion dissolves the *msn.com* community, and allows a community similar to the one obtained by SALSA to manifest itself.

Multi-Topic Query: abortion

This topic is highly polarized, with different cyber communities supporting pro-life and pro-choice views. In table 6, we bring the top 10 authorities, as determined by the two approaches:

```
Size of root size  = 160, Size of collection = 1693
Principal Community, Mutual Reinforcement Approach:
```

url	title	cat	weight
http://www.nrlc.org/	National Right To Life	(3)	0.4208
http://www.prolife.org/ultimate/	The Ultimate Pro-Life Resource List	(3)	0.3166
http://www.all.org/	What's new at American Life League	(3)	0.2515
http://www.hli.org/	Human Life International	(3)	0.2129
http://www.prolife.org/cpcs-online/	Crisis Pregnancy Centers Online	(3)	0.1877
http://www.ohiolife.org/	Ohio Right to Life	(3)	0.1821
http://www.rtl.org/	Abortion, adoption and assisted-suicide Information at Right to Life...	(1)	0.1794
http://www.bethany.org/	Bethany Christian Services	(3)	0.1614
http://www.ldi.org/	abortion malpractice litigation	(1)	0.1401
http://www.serve.com/fem4life/	Feminists for Life of America	(3)	0.1221

```
Principal Community, SALSA:
```

url	title	cat	weight
http://www.nrlc.org/	National Right To Life	(3)	0.3440
http://www.prolife.org/ultimate/	The Ultimate Pro-Life Resource List	(3)	0.2847
http://www.naral.org/	NARAL Choice for America	(3)	0.2402
http://www.feminist.org/	Feminist Majority Foundation	(3)	0.1868
http://www.now.org/	National Organization for Women	(3)	0.1779
http://www.cais.com/agm/main/	The Abortion Rights Activist	(1)	0.1661
http://www.gynpages.com/	Abortion Clinics Online	(3)	0.1631
http://www.plannedparenthood.org/	Planned Parenthood Federation	(3)	0.1572
http://www.all.org/	What's new at American Life League	(3)	0.1424
http://www.hli.org/	Human Life International	(3)	0.1424

Table 6: Authorities for WWW query "Abortion"

All 10 top authorities found by the Mutual Reinforcement approach are pro-life resources, while the top 10 SALSA authorities are split, with 6 pro-choice pages and 4 pro-life pages (which are the same top 4 pro-life pages found by the Mutual Reinforcement approach). Again, we see the TKC effect: The Mutual Reinforcement approach ranks highly authorities on only one aspect of the query, while SALSA blends authorities from both aspects into its principal community.

Multi-Topic Query: genetics

This query is especially ambiguous in the WWW: It can be in the context of genetic engineering, genetic algorithms, or in the context of health issues and the human genome.

As in the *"abortion"* query, SALSA brings a diverse principal community, with authorities on the various contexts of the query, while the Mutual Reinforcement approach is focussed on one context (Genetic Algorithms, in this case). Both principal communities are shown in table 7:

```
Size of root size  = 120, Size of collection = 2952
Principal Community, Mutual Reinforcement Approach:
```

url	title	cat	weight
http://www.aic.nrl.navy.mil/galist/	The Genetic Algorithms Archive	(3)	0.2785
http://alife.santafe.edu/	Artificial Life Online	(3)	0.2762
http://www.yahoo.com/	Yahoo!	(3)	0.2736
http://www.geneticprogramming.com/	The Genetic Programming Notebook	(1)	0.2559
http://gal4.ge.uiuc.edu/illigal.home.html	illiGAL Home Page	(3)	0.2357
http://www.cs.gmu.edu/research/gag/	The Genetic Algorithms Group...	(3)	0.2012
http://www.scs.carleton.ca/∼csgs/resources/gaal.html	Genetic Algorithms and Artificial Life Resources	(1)	0.1813
http://lancet.mit.edu/ga/	GAlib: Matthew's Genetic Algorithms Library	(3)	0.1812

```
Principal Community, SALSA:
```

url	title	cat	weight
http://www.ncbi.nlm.nih.gov/	The National Center for Biotechnology Information	(3)	0.2500
http://www.yahoo.com/	Yahoo!	(3)	0.2278
http://www.aic.nrl.navy.mil/galist/	The Genetic Algorithms Archive	(3)	0.2232
http://www.nih.gov/	National Institute of Health (NIH)	(3)	0.1947
http://gdbwww.gdb.org/	The Genome Database	(3)	0.1770
http://alife.santafe.edu/	Artificial Life Online	(3)	0.1724
http://www.genengnews.com/	Genetic Engineering News (GEN)	(1)	0.1416
http://gal4.ge.uiuc.edu/illigal.home.html	illiGAL Home Page	(3)	0.1326

Table 7: Authorities for WWW query "genetic"

6 SALSA and the In/Out Degrees of Pages

In the previous sections we have presented the Stochastic approach as an alternative method for link-structure analysis, and have shown a few experimental results which compared its performance favorably with that of the Mutual Reinforcement approach. We have also presented the TKC effect, a topological phenomenon which sometimes derails the MR approach and prevents it from converging to a useful community of authoritative sites.

The sample results shown so far have all been produced on unweighted collections, in which all informative links have received unit weight. It is likely that both approaches will produce better rankings when applied on weighted collections, in which each informative link receives a weight which reflects the amount of authority that the pointing page confers to the pointed page. Possible factors which may contribute to a link's weight include the following:

- Anchor text which is relevant to the query. Such text around a link heightens our confidence that the pointed page discusses the topic at hand ([Chakrabarti et al. 1998b],[Fürnkranz 1998]).

- One of the link's endpoints being designated by the user as highly relevant to the search topic. When a page is known to be a good authority, it seems reasonable to raise the weights of the links which enter that page. Similarly, when a page is known to be a good hub, it seems reasonable to assign high weights to its outgoing links. This approach has been recently applied in [Chakrabarti et al. 1999b]. We coin it the *anchor pages* approach, since it uses user-designated pages as anchors in the collection, around which the communities of hubs and authorities are grown.

- The link's location in the pointing page. Many search engines consider the text at the top of a page as more reflective of its contents than text further down the page. The same line of thought can be applied to the links which appear in a page, with the links which are closer to the top of the page receiving more weight than links appearing at the bottom of the page.

While the above three heuristics amplify the weight of links which are deemed as especially informative, other heuristics also lower the weights of some links. We remind the reader that such an approach was taken by

Bharat and Henzinger in their attempt to counter the effects of mass endorsements between pairs of hosts ([Bharat and Henzinger 1998], see also section 5.1).

6.1 Analysis of the Stochastic Ranking

We now prove a general result about the ranking produced by SALSA in weighted collections. The required mathematical background is brought in [Lempel and Moran 2000]; however, the full mathematical analysis is not needed for following the presentation.

Let $G = (H; A; E)$ be a positively weighted, directed bipartite graph with no isolated nodes, and let all edges be directed from pages in H to pages in A. We will use the following notations:

- The weighted in-degree of page $i \in A$:

$$d_{in}(i) \triangleq \sum_{\{k \in H | k \to i\}} w(k \to i)$$

- The weighted out-degree of page $k \in H$:

$$d_{out}(k) \triangleq \sum_{\{i \in A | k \to i\}} w(k \to i)$$

- The sum of edge weights:

$$\mathcal{W} = \sum_{i \in A} d_{in}(i) = \sum_{k \in H} d_{out}(k)$$

Let M_A be a Markov chain whose states are the set A of vertices, with the following transition probabilities between every two states $i, j \in A$:

$$P_A(i,j) = \sum_{\{k \in H | k \to i,\ k \to j)} \frac{w(k \to i)}{d_{in}(i)} \cdot \frac{w(k \to j)}{d_{out}(k)}$$

Similarly, let M_H be a Markov chain whose states are the set H of vertices, with the following transition probabilities between every two states $k, l \in H$:

$$P_H(k,l) = \sum_{\{i \in A | k \to i,\ l \to i)} \frac{w(k \to i)}{d_{out}(k)} \cdot \frac{w(l \to i)}{d_{in}(i)}$$

We will denote by $P_A[P_H]$ the $|A| \times |A|$ $[|H| \times |H|]$ stochastic matrix which is implied by the transition probabilities defined above. Accordingly, $P_A^n(i,j)$ will denote the i,j entry of the n'th power of the matrix P_A, which also equals the probability of a transition from state i to state j in n steps.

Consider the following binary relation on the vertices of A (states of M_A):

$$R_A = \{(i,j) \mid \exists n \text{ such that } P_A^n(i,j) > 0\}$$

Since we assumed that there are no isolated nodes in G, it follows that for every $i \in A$, $P_A(i,i) > 0$. Hence, R_A is reflexive and M_A is aperiodic (primitive). From the definition of the transition probability $P_A(i,j)$, it is clear that $P_A(i,j) > 0$ implies $P_A(j,i) > 0$. Hence, R_A is symmetric. It is easily shown that R_A is also transitive, and is thus an equivalence relation on A. The equivalence classes of R_A are the irreducible components of M_A. Similar arguments hold for M_H.

We first deal with the case where R_A consists of one equivalence class (i.e., M_A is irreducible).

Proposition 5 *Whenever M_A is an irreducible chain (has a single irreducible component), it has a unique stationary distribution $\pi = (\pi_1, \ldots, \pi_{|A|})$ satisfying:*

$$\pi_i = \frac{d_{in}(i)}{\mathcal{W}} \ \textit{for all } i \in A$$

Similarly, whenever M_H is an irreducible chain, its unique stationary distribution $\pi = (\pi_1, \ldots, \pi_{|H|})$ satisfies:

$$\pi_k = \frac{d_{out}(k)}{\mathcal{W}} \ \textit{for all } k \in H$$

Proof: We will prove the proposition for M_A. The proof for M_H is similar.

By the Ergodic Theorem ([Gallager 1996]), any irreducible, aperiodic Markov chain has a unique stationary distribution vector. It will therefore suffice to show that the vector π with the properties claimed in the proposition is indeed a stationary distribution vector of M_A.

1. π is a distribution vector: Its entries are non-negative, and their sum equals one.

$$\sum_{i \in A} \pi_i = \sum_{i \in A} \frac{d_{in}(i)}{\mathcal{W}} = \frac{1}{\mathcal{W}} \sum_{i \in A} d_{in}(i) = 1$$

2. π is a stationary distribution vector of M_A. Here we need to show the equality $\pi P_A = \pi$:

$$
\begin{aligned}
[\pi P_A]_i &= \sum_{j\in A} \pi_j P_A(j,i) \\
&= \sum_{j\in A} \frac{d_{in}(j)}{\mathcal{W}} \sum_{\{k\in H | k\to i, k\to j\}} \frac{w(k\to j)}{d_{in}(j)} \frac{w(k\to i)}{d_{out}(k)} \\
&= \frac{1}{\mathcal{W}} \sum_{j\in A} \sum_{\{k\in H | k\to i, k\to j\}} \frac{w(k\to j)\cdot w(k\to i)}{d_{out}(k)} \\
&= \frac{1}{\mathcal{W}} \sum_{\{k\in H | k\to i\}} \sum_{\{j\in A | k\to j\}} \frac{w(k\to j)\cdot w(k\to i)}{d_{out}(k)} \\
&= \frac{1}{\mathcal{W}} \sum_{\{k\in H | k\to i\}} \frac{w(k\to i)}{d_{out}(k)} \sum_{\{j\in A | k\to j\}} w(k\to j) \\
&= \frac{1}{\mathcal{W}} \sum_{\{k\in H | k\to i\}} w(k\to i) \\
&= \frac{d_{in}(i)}{\mathcal{W}} \\
&= \pi_i
\end{aligned}
$$

□

Thus, when the (undirected) support graph of G is connected, SALSA assigns each page an authority weight which is proportional to the sum of weights of its incoming edges. The hub weight of each page is proportional to the sum of weights of its outgoing edges. In unweighted collections (with all edges having unit weight), each page's Stochastic authority(hub) weight is simply proportional to the in(out) degree of the page.

This mathematical analysis, in addition to providing insight about the ranking that is produced by SALSA, also suggests a very simple algorithm for calculating the Stochastic ranking: Simply calculate, for all pages, the sum of weights on their incoming(outgoing) edges, and normalize these two vectors. There is no need to apply any resource-consuming iterative method to approximate the principal eigenvector of the transition matrix of the Markov chain.

Markov chains with multiple irreducible components Consider the case in which the authority chain M_A consists of multiple irreducible components. Denote these (pairwise disjoint) components by $A_1, A_2, \ldots, A_k$ where $A_i \subset A, 1 \leq i \leq k$. What will be the outcome of a random walk performed on the set of states A according to the transition matrix P_A? To answer this question, we will need some notations:

- Let e denote the $|A|$-dimensional distribution vector, all whose entries equal $\frac{1}{|A|}$.
- For all vertices $j \in A$, denote by $c(j)$ the irreducible component (equivalence class of R_A) to which j belongs: $c(j) = l \iff j \in A_l$.
- Let $\pi^1, \pi^2, \ldots, \pi^k$ be the unique stationary distributions of the (irreducible) Markov chains induced by $A_1, \ldots, A_k$.
- Denote by $\pi_j^{c(j)}$ the entry which corresponds to j in $\pi^{c(j)}$ (the stationary distribution of j's irreducible component, $A_{c(j)}$).

Proposition 6 *The random walk on A, governed by the transition matrix P_A and started from all states with equal probability, will converge to a stationary distribution as follows:*

$$\lim_{n\to\infty} eP_A^n = \tilde{\pi} \quad \textit{where} \quad \tilde{\pi}_j = \frac{|A_{c(j)}|}{|A|} \cdot \pi_j^{c(j)}$$

Proof: Denote by p_i^n, $1 \leq i \leq k$ the probability of being in a page belonging to A_i after the n'th step of the random walk. This probability is determined by the distribution vector eP_A^n. Clearly,

$$p_i^0 = \sum_{j\in A_i} e_j = \frac{|A_i|}{|A|}$$

Since the transition probability between any two pages (states) which belong to different irreducible components is zero (probability does not shift from one component to another), $p_i^n = p_i^0$ for all n. Inside each irreducible component the Ergodic Theorem holds, thus the probabilities which correspond to the pages of A_i in $\lim_{n\to\infty} eP_A^n$ will be proportional to π^i, and the proposition follows.

□

This proposition points out a natural way to compare the authoritativeness of pages from different irreducible components: Simply multiply each page's authority score by the normalized size of the irreducible component to which it belongs. The underlying principle is obvious: The size of the community should be considered when evaluating the quality of the top pages in that community. The budget which the Mayor of New York City controls is much larger than that of the Mayor of Osh Kosh, Wisconsin.

The combination of a page's intra-community authority score and its community's size is one of the factors that enable SALSA to blend authorities from different aspects of a multi-topic query, and which reduces its vulnerability to the TKC effect.

6.2 In-Degree as a Measure of Authority (Revisited)

Extensive research in link-structure analysis has been conducted in recent years under the premise that considering the in-degree of pages as a sole measure of their authority does not produce satisfying results. Kleinberg, as a motivation to the Mutual Reinforcement approach, showed some examples of the inadequacy of a simple in-degree ranking ([Kleinberg 1998]). Our results in section 5.2 seem to contradict this premise: The Stochastic rankings seem quite satisfactory there, and since those collections were unweighted, the Stochastic rankings are equivalent to simple in-degree counts (normalized by the size of the connected component which each page belongs to). To gain more perspective on these conflicting results, let us elaborate on the first stage of the meta-algorithm for link-structure analysis (from section 3), in which the graph to be analyzed is assembled:

1. Given a query, assemble a collection of Web-pages which should contain many hubs and authorities pertaining to the query, and few hubs and authorities for any particular unrelated topic.

2. Filter out non-informative links connecting pages in the collection.

3. Assign weights to all non-filtered links. These weights should reflect the information conveyed by the link.

It is only after these steps that the weighted, directed graph is analyzed and that the rankings of hubs and authorities are produced. The analysis of the graph, however important, is just the second stage in the meta-algorithm, and the three steps detailed above, which comprise the first stage of the meta-algorithm, are crucial to the success of the entire algorithm.

We claim that the success of SALSA, as opposed to the earlier reported inadequacies of in-degree based ranking schemes, is mainly due to considerable research efforts which have been invested recently in improving the quality of the assembled WWW subgraphs. The techniques utilized in topic-specific WWW subgraph assembly are now such that in many cases, simple (and efficient) ranking algorithms produce quite satisfying results on the assembled subgraphs. The techniques for the three-step subgraph assembly process were described throughout the paper:

- Kleinberg's scheme of building a base set of pages around a small root set which pertains to a query (see both [Kleinberg 1998] and section 2) ensures, in most cases, that the collection will indeed be centered around the relevant topic.

- We have detailed the link-filtering schemes that we have applied in section 5.2. An additional filtering scheme is described in [Chakrabarti et al. 1998a], where links are considered to be navigational if they connected two servers which reside on the same IP-subnet.

- Several simple heuristics for assigning weights to links have been described in the beginning of section 6.

It is important to keep in mind the main goal of broad-topic WWW searches, which is to enhance the precision at 10 of the results, not to rank the entire collection of pages correctly. It is entirely irrelevant if the page in place 98 is really better than the page in place 216. The Stochastic ranking, which turns out to be equivalent to a weighted in-degree ranking, discovers the most authoritative pages quite effectively (and very efficiently) in many (carefully assembled) collections. No claim is made on the quality of its ranking on the rest of the pages (which constitute the vast majority of the collection).

Ranking hubs by out-degree One of the strengths of link-structure analysis in finding authoritative pages is that it is less vulnerable to spamming techniques than content analysis is. It is much easier for a Web master to manipulate the contents of his page than to manipulate the amount and origin of a page's incoming links. As a consequence, having many incoming links from prominent pages is viewed as a reliable measure of authority. However, when ranking pages as hubs, we are again susceptible to spamming, since the outgoing links of a page are in total control of the page's creator. SALSA, which ranks hubs according to their (weighted) out-degree, thus might seem especially vulnerable to spammers, which by simply adding many (irrelevant and noisy) links to their pages can increase the hub-scores of those pages.

We argue that while some susceptibility to spamming does exist, most spamming attempts should be thwarted by the nature of the graphs which are analyzed. The set of nodes of those graphs is determined by the neighbors of a small root set of pages (the size of the root set is typically less than ten percent of the size of the entire graph).

Consider an outlink-spammer page p, one that has many irrelevant outgoing links:

- When p is not a member of the root set, it is very likely that most of p's irrelevant links will not affect the analysis. This is because p does not take part in determining the set of nodes of the graph to be analyzed. Those nodes are determined by the root set, and only links of p which intersect with the root set or its immediate neighborhood take part in the analysis. Those links may very well be informative, and will credit p as a hub, while the spam links of p will not intersect with the neighborhood of the root set and thus will have no effect on the analysis.

- In the rare cases when p is part of the root set, its links will indeed fall inside the analyzed graph and affect the analysis. However, for p to be a member of the root set it must have ranked highly for the query in question in some search engine. Thus, its contents is probably somewhat relevant to the query, and therefore some of its outgoing links may also be relevant. In addition, many search engines devote a lot of effort to fighting spam pages. Thus, spam pages need to overcome many obstacles before infiltrating into the root set of link-analyzed collections.

As a final note, we recognize that hub scores present an opportunity to spammers. Denoting the set of outgoing links of a page p by $\mathcal{L}(p)$, we observe that both SALSA and the Mutual Reinforcement approach obey the following property:

$$\mathcal{L}(p) \subseteq \mathcal{L}(q) \Longrightarrow \text{hub-score}(p) \leq \text{hub-score}(q)$$

Thus, in both approaches, adding outgoing links to your page can only improve its hub score. In order to fight this sort of spam, link analysis must punish pages for having an excess of irrelevant links.

7 Conclusions

We have developed a new approach for finding hubs and authorities, which we call *SALSA*- The Stochastic Approach for Link Structure Analysis. SALSA examines random walks on two different Markov chains which are derived from the link structure of the WWW: The authority chain and the hub chain. The principal community of authorities (hubs) corresponds to the pages that are most frequently visited by the random walk defined by the authority (hub) Markov chain. SALSA and Kleinberg's Mutual Reinforcement approach are both in the framework of the same meta-algorithm.

We have shown that the ranking produced by SALSA is equivalent to a weighted in/out-degree ranking (with the sizes of irreducible components also playing a part). This makes SALSA computationally lighter than the Mutual Reinforcement approach.

Both approaches were tested on the WWW, where SALSA appears to compare well with the Mutual Reinforcement approach. These tests, as well as analytical consideration, have revealed a topological phenomenon on the Web called the TKC effect. This effect sometimes derails the Mutual Reinforcement approach, and prevents it from finding relevant authoritative pages (or from finding authorities on all meanings/aspects of the query):

- In single topic collections, the TKC effect sometimes results in the Mutual Reinforcement approach ranking many irrelevant pages as authorities.

- In multi-topic collections, the principal community of authorities found by the Mutual Reinforcement approach tends to pertain to only one of the topics in the collection. The Mutual Reinforcement approach can discover the other aspects of the collection by deriving *non-principal* communities of hubs and authorities [2]. These communities, as was demonstrated in [Kleinberg 1998], are often able to capture separately the top authorities[hubs] of the multiple aspects of multi-topic collections. While a thorough discussion of non-principal communities is out of the scope of this paper, we note that in the *abortion* collection (table 6), in which the principal community of the Mutual Reinforcement approach centered on pro-life pages, the pro-choice authorities are present in the first non-principal community. Likewise, in the

[2] communities which are derived from the non-principal eigenvectors of the co-citation and bibliographic coupling matrices

genetic collection (table 7), the two aspects not represented in the principal community of the Mutual Reinforcement approach (*genetic engineering* and the *human genome*) are found among the first few non-principal communities.

We note that SALSA is less vulnerable to the TKC effect, and produces good results in cases where the Mutual Reinforcement approach fails to do so. It is also frequently able to blend authorities from multiple aspects of multi-topic collections into its principal community of authorities.

The following issues are left for future research:

1. In collections with many connected components, we have studied one manner in which to combine the inner-component authority score with the size of the component. There may be better ways to combine these two factors into a single score.

2. We have found a simple property of the Stochastic ranking, which enables us to compute this ranking without the need to approximate the principal eigenvector of the stochastic matrix which defines the random walk. Is there some simple property which will allow us to calculate the Mutual Reinforcement ranking without approximating the principal eigenvector of $W^T W$? If not, can we alter the graph G in some simple manner (for instance, change some weights on the edges) so that the Stochastic ranking on the modified graph will be approximately equal to the Mutual Reinforcement ranking on the original graph?

Acknowledgments

The second author would like to thank Udi Manber for introducing him to the search problems studied in this paper, and Udi Manber and Toni Pitassi for delightful and interesting discussions at the early stages of this research.

References

[AltaVista] AltaVista Company. AltaVista. http://www.altavista.com/.

[Auguston and Minker 1970] J. Gary Auguston and Jack Minker. An analysis of some graph theoretical cluster techniques. *JACM*, 17(4):571–588, October 1970.

[Bharat and Henzinger 1998] Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *Proc. 21'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[Brin and Page 1998] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Proc. 7th International WWW Conference*, 1998.

[Botafogo et al. 1992] R.A. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180, April 1992.

[Chakrabarti et al. 1998a] S. Chakrabarti, B. Dom, D. Gibson, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Spectral filtering for resource discovery. *ACM SIGIR workshop on Hypertext Information Retrieval on the Web*, 1998.

[Chakrabarti et al. 1998b] Soumen Chakrabarti, Byron Dom, David Gibson, Jon M. Kleinberg, Prabhakar Raghavan, and Sridhar Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. *Proc. 7th International WWW Conference*, 1998.

[Chakrabarti et al. 1999a] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Hypersearching the web. *Scientific American*, June 1999.

[Chakrabarti et al. 1999b] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Mining the link structure of the www. *IEEE Computer*, August 1999.

[CLEVER] IBM Corporation Almaden Research Center. CLEVER. http://www.almaden.ibm.com/cs/k53.clever.html.

[Carriére and Kazman 1997] Jeromy Carrière and Rick Kazman. Webquery: Searching and visualizing the web through connectivity. *Proc. 6th International WWW Conference*, 1997.

[Canyontrace] LLC Canyontrace New Media Marketing. Link site with strategic link development services by canyontrace. http://www.canyontrace.com/strategic_link_development.htm.

[Frisse 1988] Mark E. Frisse. Searching for information in a hypertext medical handbook. *Communications of the ACM*, 31(27):880–886, July 1988.

[Fürnkranz 1998] Johannes Fürnkranz. Using links for classifying webpages. Technical Report TR-OEFAI-98-29, Austrian Research Institute for Artificial Intelligence, 1998.

[Gallager 1996] Robert G. Gallager. *Discrete Stochastic Processes.* Kluwer Academic Publishers, 1996.

[Garfield 1972] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.

[Gibson et al. 1998] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. *Proc. 9th ACM Conference on Hypertext and Hypermedia*, 1998.

[Google] Google Inc. Google search engine. http://www.google.com/.

[Grantastic] Grantastic Designs. Search engine optimization services from grantastic designs. http://www.grantasticdesigns.com/seo.html.

[Kessler 1963] M.M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[Kleinberg et al. 1999] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The web as a graph: Measurements, models and methods. *Proceedings of the Fifth International Computing and Combinatorics Conference*, 1999.

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

[Law et al. 1999] Ken Law, Thomas Tong, and Alan Wong. Automatic categorization based on link structure, 1999. http://www.stanford.edu/~tomtong/cs349/web.htm.

[Lempel and Moran 2000] Ronny Lempel and Shlomo Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. Technical Report CS-2000-06, Technion - Israel Institute of Technology, May 2000.

[LinkMe] Internet Marketing for Internet Business. Link management by linkme. http://www.linkme.com/.

[Marchiori 1997] Massimo Marchiori. The quest for correct information on the web: Hyper search engines. *Proc. 6th International WWW Conference*, 1997.

[Pirolli et al. 1996] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the web. *Proc. ACM SIGCHI Conference on Human Factors in Computing*, 1996.

[Papadimitriou et al. 1998] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Preliminary version appeared in PODS 98*, pages 159–168, 1998.

[Small 1973] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. American Soc. Info. Sci.*, 24:265–269, 1973.

[VanRijsbergen 1979] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, 1979.

[Weiss et al. 1996] R. Weiss, B. Vélez, M. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. Gifford. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. *Proc. 7th ACM Conference on Hypertext*, 1996.

[Yahoo!] Yahoo! Inc. Yahoo! http://www.yahoo.com/.