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 sites 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-sites: 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 sites.

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 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 sites 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 for 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 sites which are highly relevant to the topic of the query, significantly more than most other sites. 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 ([24],[5]):

- 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 con-
siderably).

In addition to the classical issues in Information Retrieval, there is a Web-
specific obstacle which search engines must overcome, called search engine
persuasion ([22]). 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 busi-
nesses 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 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 ([5]). 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 hyper-
linked 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 recom-
mands, 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 \( q \) ([18]). 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 ([5]), such as intra-domain (inner) links (whose purpose is to provide navigational aid in a complex Web site of some organization), commercial/sponsor links, and links which result from link-exchange agreements. 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 WWW age, link structures were studied in the area of bibliometrics, which studies the citation structure of written documents ([27],[17]). Many works in this area were aimed at finding high-impact papers published in scientific journals ([12]), and at clustering related documents ([1]).

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 ([3]). Other works used link structures for categorizing pages and clustering them ([29],[25]).

Marchiori, in [22], 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 ([18]). In an attempt to impose some structure on the chaotic WWW, Kleinberg distinguished between two types of Web-sites which pertain to a certain topic: The first are authoritative pages in the sense described previously. The second type of sites are hub pages. Hubs are resource lists - They do not directly contain information pertaining to the topic, but rather point to many authoritative sites. 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 [13], and offers some enlightening practical remarks. The ARC
system, described in [8], 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 ([4]). Researchers from outside IBM, such as Henzinger and Brahat, have also studied Kleinberg’s approach and have proposed improvements to it ([14]).

Anchor text has also been used by Brin and Page in [2]. Another major feature of their work on the Google search engine ([16]) is a link-structure based site ranking approach called PageRank, which can be interpreted as a stochastic analysis of some random-walk behavior through the entire WWW.

In [21], 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 sites pertaining to a given topic should be split to hubs and authorities, we replace Kleinberg’s Mutual Reinforcement approach ([18]) 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 sites in $G$ should be highly visible (reachable) from many sites in $G$. Thus, we will attempt to identify these sites by examining certain random walks in $G$, under the proviso that such random walks will tend to visit these highly visible sites more frequently than other, less connected sites. 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 sites in the Stochastic approach may be calculated by
examining the weighted in/out degrees of the sites in $G$. This result yields
that SALSA is computationally lighter than the Mutual Reinforcement ap-
proach. We also discuss the reason for our success with analyzing weighted
in/out degrees of sites, which previous work has claimed to be unsatisfactory
for identifying authoritative sites.

The rest of the paper is organized as follows: Section 2 recounts Klein-
berg’s Mutual Reinforcement Approach. In section 3 we view Kleinberg’s
approach from a higher level, and define a meta-algorithm for link struc-
ture 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 sites. 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 is pro-
vided in appendix A. Appendix B contains a detailed proof of one of the
propositions which are brought in the paper.

The main contribution of the paper can be grasped without following
the full mathematical analysis.

2 Kleinberg’s Mutual Reinforcement Approach

The Mutual Reinforcement approach ([18]) starts by assembling a collection
$C$ of Web-sites, 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 sites on topic $t$.

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

- A root set $S$ of sites is obtained by applying a term based search engine,
such as AltaVista [10], to the query $q$. This is the only step in which
the lexical content of the Web sites is examined.
From $S$ we derive a base set $\mathcal{C}$ which consists of (a) sites in the root set $S$, (b) sites which point to a site in $S$ and (c) sites which are pointed to by a site 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 sites point to [a given url]".

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

Each site $s \in \mathcal{C}$ 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 site's hub weight should be proportional to the sum of the authority weights of the sites it points at.

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

The top ranking sites, 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 sites $s \in \mathcal{C}$.

2. Repeat the following three operations until convergence:

- Update the authority weight of each site $s$ (the $I$ operation): 
  $$a(s) \leftarrow \sum_{[x|x \text{ points to } s]} h(x)$$

- Update the hub weight of each site $s$ (the $O$ operation): 
  $$h(s) \leftarrow \sum_{[x|x \text{ points to } x]} a(x)$$

- Normalize the authority weights and the hub weights.

Note that applying the $I$ 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 $O$ operation is equivalent to assigning hub weights according to the result of multiplying the vector of all authority weights by the matrix $W$. 

6
Kleinberg showed that this algorithm converges, and that the resulting authority weights [hub weights] are the coordinates of the normalized principal eigenvector\(^1\) of \(W^T W\) [of \(WW^T\)]. \(W^T W\) and \(WW^T\) are well known matrices in the field of bibliometrics:

1. \(A \triangle \equiv W^T W\) is the \textit{co-citation matrix} ([27]) of the collection. \([A]_{i,j}\) is the number of sites 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 \(A\).

2. \(H \triangle \equiv WW^T\) is the \textit{bibliographic coupling matrix} ([17]) of the collection. \([H]_{i,j}\) is the number of sites 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 [7]:

- Given a topic \(t\), construct a site collection \(C\) which should contain many \(t\)-hubs and \(t\)-authorities, but should not contain many hubs or authorities for any other topic \(t'\). Let \(n = |C|\).

- Derive, from \(C\) and the link structure induced by it, two \(n \times n\) association matrices - a \textit{hub matrix} \(H\) and an \textit{authority matrix} \(A\). Association matrices are widely used in classification algorithms ([28]), and will be used here in order to classify the Web sites 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):

\[ M \text{ will have a unique real positive eigenvalue } \lambda(M) \text{ of multiplicity } 1, \text{ such that for any other eigenvalue } \lambda' \text{ of } M, \lambda(M) > \lambda'(M). \text{ Denote by } v_{\lambda(M)} \text{ the (unique) unit eigenvector which corresponds to } \lambda(M) \]

\(^1\text{The eigenvector which corresponds to the eigenvalue of highest magnitude of the matrix.}\)
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 $k$ sites that correspond to the largest coordinates of $v_{\lambda(A)}$ will form the principal algebraic community of authorities in $C$. The principal algebraic community of hubs in $C$ 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 $C$.

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 the Stochastic Approach for Link Structure Analysis - SALSA. The approach is based upon the theory of Markov chains, and relies on the stochastic properties of random walks performed on our collection of sites. 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 sites $C$ which is built around a topic $t$ in the manner described in section 2. Intuition suggests that authoritative sites on topic $t$ should be visible from many sites in the subgraph induced by $C$. 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 sites, 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 two 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 site two distinct scores, a hub score and an authority score.

The idea of ranking Web sites using random walks is not new. The search engine Google ([2],[16]) incorporates stochastic information into its ranking
of pages. The PageRank component of the search engine examines a single random walk on the entire WWW. Hence, the ranking of Web sites in Google is independent of the search query (a global ranking), and no distinction is made between hubs and authorities.

Let us build a bipartite undirected graph \( \tilde{G} = (V_h, V_a, E) \) from our site collection \( C \) and its link structure:

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

Each non-isolated site \( s \in C \) is represented by two nodes of \( \tilde{G}, s_h \) and \( s_a \). Each WWW-link \( s \rightarrow r \) is represented by an undirected edge connecting \( s_h \) and \( r_a \).

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 site.

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:

\[
\tilde{h}_{i,j} = \sum_{\{k|(i_h, k_a) \in \tilde{G}\}} \frac{1}{\text{deg}(i_h)} \cdot \frac{1}{\text{deg}(k_a)}
\]
2. The authority-matrix $A$, defined as follows:

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

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

Alternatively, the matrices $H$ and $A$ can be defined as follows: Let $W$ be the adjacency matrix of the directed graph defined by $C$ 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 $H$ consists of the non-zero rows and columns of $W_r W^T$, and $A$ consists of the non-zero rows and columns of $W_c W_r$. We ignore the rows and columns of $A, H$ which consist entirely of zeros, since (by definition) all the nodes of $\tilde{G}$ have at least one incident edge. The matrices $A$ and $H$ 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 $A$ and $H$ 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 $A$ and $H$ have multiple irreducible components.

Some properties of $H$ and $A$:

- 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 (since all the nodes are non-isolated). Hence, every state (=site) in each of the chains has a self-loop, causing the chains to be aperiodic.

- The adjacency matrix of the support graph of $A$ 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 $H$ 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 $k$ sites having the highest entries in the principal eigenvector of $\tilde{A}$ ($\tilde{H}$), for some user defined $k$. By the Ergodic Theorem ([11]), 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 sites 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 emphasize 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 sites. Roughly speaking, the TKC effect occurs when such a community scores high in link-analyzing algorithms, even though the sites 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 sites 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 $C_l$ is the dominant topic of the collection, and is probably of wider interest on the WWW. Since there are many $C_l$-authoritative sites, the hubs do not link to all of them, whereas the smaller $C_s$ community is densely interconnected. The TKC effect occurs when the sites 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 $n \geq (k+1)^2$ 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:

\[
\frac{h_l}{h_s} \frac{n}{m}
\]

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

\[
\frac{\binom{n-1}{k-1}}{2} < \left( \frac{n-1}{k-1} \right) - n = h_s < \left( \frac{n-1}{k-1} \right)
\]

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

Now:

\[
\frac{h_l}{h_s} \frac{n}{m} = \frac{\binom{n}{k} \binom{n-1}{k-1} - n}{\binom{n}{k}}, \frac{m}{n}
\]

\[
= \frac{\binom{n}{k} \binom{n-1}{k-1} - n}{\binom{n}{k}}, \frac{m}{n}
\]

\[
= \frac{n}{r \cdot \binom{n-1}{k-1}} \cdot \frac{k + 1}{(k+1)^2} \quad \text{(for some } 0.5 < r < 1) \]

\[
= \frac{n}{r \cdot k \cdot \frac{1}{k+1}} \quad \text{for some } 0.5 < r < 1 \]

\[
= \frac{(k+1)^2}{r \cdot k (k+1)} = \frac{k+1}{k} \cdot \frac{1}{r}
\]
And since $k \geq 3$, $0.5 < r < 1$ we have:

$$1 < \frac{k+1}{k} < \frac{h_i}{n} \frac{h_j}{m} < \frac{2(k+1)}{k} \leq \frac{8}{3}$$

Hence, both communities have roughly the same ratio of hubs to authorities. In appendix B we show:

**Proposition 1** On the collection $C_k$, SALSA will rank the $C_l$-authorities above the $C_s$-authorities.

While:

**Proposition 2** On the collection $C_k$, the Mutual Reinforcement approach will rank the $C_l$-authorities above the $C_s$-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).

Let us now change the collection $C_k$ a bit. Let $A_b$ be any nonempty proper subset of size $b$ of the $C_s$-authorities ($1 \leq b < m$). We add to $C_k$ 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 $\tilde{C}_k$. The resulting principal communities of authorities derived by the two approaches will be:

**Proposition 3** On the collection $\tilde{C}_k$, 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 $\tilde{C}_k$, 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 appendix B as well), we see that SALSA blends the authorities from the two communities in $\tilde{C}_k$, 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.
5.2 The WWW

We tested the different approaches on broad-topic WWW queries (both single topic queries and multi-topic queries). We obtained a collection of sites 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 [18], and are brought here for the sake of comparison. All collections were assembled during February, 1999. The root sets were compiled using AltaVista ([10]), which also provided the linkage information needed for building the base sets.

When expanding the root set to the entire collection, we filtered the links pointing to and from Web sites. Following [18], 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). We also ignored links to cgi scripts, and tried to identify ad-links and ignore them as well. 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. ([19]) 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, (2) denotes a site pointing into the root set, and (3) denotes a site pointed at by a member of the root set.
4. The value of the coordinate of this url in the principal eigenvector of the authority matrix.

Single-Topic Query: +censorship +net

For this query, both approaches produced the same top six sites (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:
<table>
<thead>
<tr>
<th>url</th>
<th>header</th>
<th>cat</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.eff.org/">http://www.eff.org/</a></td>
<td>EFFweb-The Electronic Frontier Foundation</td>
<td>(3)</td>
<td>0.5355</td>
</tr>
<tr>
<td><a href="http://www.epic.org/">http://www.epic.org/</a></td>
<td>Electronic Privacy Information Center</td>
<td>(3)</td>
<td>0.3584</td>
</tr>
<tr>
<td><a href="http://www.ctd.org/">http://www.ctd.org/</a></td>
<td>The Center For Democracy and Technology</td>
<td>(3)</td>
<td>0.3525</td>
</tr>
<tr>
<td><a href="http://www.eff.org/blueribbon.html">http://www.eff.org/blueribbon.html</a></td>
<td>Blue Ribbon Campaign For Online Free Speech</td>
<td>(3)</td>
<td>0.2810</td>
</tr>
<tr>
<td><a href="http://www.aclu.org/">http://www.aclu.org/</a></td>
<td>ACLU: American Civil Liberties Union</td>
<td>(3)</td>
<td>0.2800</td>
</tr>
<tr>
<td><a href="http://www.vtw.org/">http://www.vtw.org/</a></td>
<td>The Voters Telecommunications Watch</td>
<td>(3)</td>
<td>0.2539</td>
</tr>
</tbody>
</table>

Principal Community, SALSA:

<table>
<thead>
<tr>
<th>url</th>
<th>header</th>
<th>cat</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.eff.org/">http://www.eff.org/</a></td>
<td>EFFweb-The Electronic Frontier Foundation</td>
<td>(3)</td>
<td>0.3848</td>
</tr>
<tr>
<td><a href="http://www.eff.org/blueribbon.html">http://www.eff.org/blueribbon.html</a></td>
<td>Blue Ribbon Campaign For Online Free Speech</td>
<td>(3)</td>
<td>0.3207</td>
</tr>
<tr>
<td><a href="http://www.epic.org/">http://www.epic.org/</a></td>
<td>Electronic Privacy Information Center</td>
<td>(3)</td>
<td>0.2566</td>
</tr>
<tr>
<td><a href="http://www.ctd.org/">http://www.ctd.org/</a></td>
<td>The Center For Democracy and Technology</td>
<td>(3)</td>
<td>0.2566</td>
</tr>
<tr>
<td><a href="http://www.vtw.org/">http://www.vtw.org/</a></td>
<td>The Voters Telecommunications Watch</td>
<td>(3)</td>
<td>0.2405</td>
</tr>
<tr>
<td><a href="http://www.aclu.org/">http://www.aclu.org/</a></td>
<td>ACLU: American Civil Liberties Union</td>
<td>(3)</td>
<td>0.2405</td>
</tr>
</tbody>
</table>

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 sites 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:

<table>
<thead>
<tr>
<th>url</th>
<th>title</th>
<th>cat</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.jars.com/">http://www.jars.com/</a></td>
<td>EarthWeb’s JARS.COM Java Review Service</td>
<td>(3)</td>
<td>0.3341</td>
</tr>
<tr>
<td><a href="http://www.gamelan.com/">http://www.gamelan.com/</a></td>
<td>Gamelan - The Official Java Directory</td>
<td>(3)</td>
<td>0.3036</td>
</tr>
<tr>
<td><a href="http://www.javascripts.com/">http://www.javascripts.com/</a></td>
<td>Javascripts.com - Welcome</td>
<td>(3)</td>
<td>0.2553</td>
</tr>
<tr>
<td><a href="http://www.datamation.com/">http://www.datamation.com/</a></td>
<td>EarthWeb’s Datamation</td>
<td>(3)</td>
<td>0.2514</td>
</tr>
<tr>
<td><a href="http://www.roadcoders.com/">http://www.roadcoders.com/</a></td>
<td>Handheld Software Development® RoadCoders</td>
<td>(3)</td>
<td>0.2508</td>
</tr>
<tr>
<td><a href="http://www.earthweb.com/">http://www.earthweb.com/</a></td>
<td>EarthWeb</td>
<td>(3)</td>
<td>0.2494</td>
</tr>
<tr>
<td><a href="http://www.earthwebdirect.com/">http://www.earthwebdirect.com/</a></td>
<td>Welcome to Earthweb Direct</td>
<td>(3)</td>
<td>0.2475</td>
</tr>
<tr>
<td><a href="http://www.itknowledge.com/">http://www.itknowledge.com/</a></td>
<td>ITKnowledge</td>
<td>(3)</td>
<td>0.2469</td>
</tr>
<tr>
<td><a href="http://www.intranetjournal.com/">http://www.intranetjournal.com/</a></td>
<td>intranetjournal.com</td>
<td>(3)</td>
<td>0.2452</td>
</tr>
<tr>
<td><a href="http://www.javagoodies.com/">http://www.javagoodies.com/</a></td>
<td>Java Goodies JavaScript Repository</td>
<td>(3)</td>
<td>0.2388</td>
</tr>
</tbody>
</table>

Principal Community, SALSA:

<table>
<thead>
<tr>
<th>url</th>
<th>title</th>
<th>cat</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://java.sun.com/">http://java.sun.com/</a></td>
<td>Java™ Technology Home Page</td>
<td>(3)</td>
<td>0.3653</td>
</tr>
<tr>
<td><a href="http://www.gamelan.com/">http://www.gamelan.com/</a></td>
<td>Gamelan - The Official Java Directory</td>
<td>(3)</td>
<td>0.3637</td>
</tr>
<tr>
<td><a href="http://www.jars.com/">http://www.jars.com/</a></td>
<td>EarthWeb’s JARS.COM Java Review Service</td>
<td>(3)</td>
<td>0.3637</td>
</tr>
<tr>
<td><a href="http://www.javaworld.com/">http://www.javaworld.com/</a></td>
<td>IDG’s magazine for the Java community</td>
<td>(3)</td>
<td>0.2173</td>
</tr>
<tr>
<td><a href="http://www.yahoo.com/">http://www.yahoo.com/</a></td>
<td>Yahoo!</td>
<td>(3)</td>
<td>0.2144</td>
</tr>
<tr>
<td><a href="http://www.javaworld.com/">http://www.javaworld.com/</a></td>
<td>Java™ Technology Home Page</td>
<td>(3)</td>
<td>0.2031</td>
</tr>
<tr>
<td><a href="http://www.sun.com/">http://www.sun.com/</a></td>
<td>Sun Microsystems</td>
<td>(3)</td>
<td>0.1874</td>
</tr>
<tr>
<td><a href="http://www.javascripts.com/">http://www.javascripts.com/</a></td>
<td>Javascripts.com - Welcome</td>
<td>(3)</td>
<td>0.1385</td>
</tr>
<tr>
<td><a href="http://www.htmlgoodies.com/">http://www.htmlgoodies.com/</a></td>
<td>htmlgoodies.com - Home</td>
<td>(3)</td>
<td>0.1307</td>
</tr>
<tr>
<td><a href="http://javaboutique.internet.com/">http://javaboutique.internet.com/</a></td>
<td>The Ultimate Java Applet Resource</td>
<td>(1)</td>
<td>0.1181</td>
</tr>
</tbody>
</table>

Table 2: Authorities for WWW query ”Java”

Single-Topic Query: movies

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

Size of root size = 175, Size of collection = 4539
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 site. To understand how these sites scored so well, we turn to the principal community of hubs, shown in table 4:

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 sites 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 sites, and are much more relevant to the query:
<table>
<thead>
<tr>
<th>url</th>
<th>title</th>
<th>cat</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.imdb.com/">http://www.imdb.com/</a></td>
<td>The Internet Movie Database</td>
<td>(3)</td>
<td>0.2533</td>
</tr>
<tr>
<td><a href="http://www.mrshowbiz.com/">http://www.mrshowbiz.com/</a></td>
<td>Mr Showbiz</td>
<td>(3)</td>
<td>0.2233</td>
</tr>
<tr>
<td><a href="http://www.disney.com/">http://www.disney.com/</a></td>
<td>Disney.com—The Web Site for Families</td>
<td>(3)</td>
<td>0.2200</td>
</tr>
<tr>
<td><a href="http://www.hollywood.com/">http://www.hollywood.com/</a></td>
<td>Hollywood Online:...all about movies</td>
<td>(3)</td>
<td>0.2134</td>
</tr>
<tr>
<td><a href="http://www.imdb.com/">http://www.imdb.com/</a></td>
<td>The Internet Movie Database</td>
<td>(3)</td>
<td>0.2000</td>
</tr>
<tr>
<td><a href="http://www.paramount.com/">http://www.paramount.com/</a></td>
<td>Welcome to Paramount Pictures</td>
<td>(3)</td>
<td>0.1907</td>
</tr>
<tr>
<td><a href="http://www.mca.com/">http://www.mca.com/</a></td>
<td>Universal Studios</td>
<td>(3)</td>
<td>0.1800</td>
</tr>
<tr>
<td><a href="http://www.discovery.com/">http://www.discovery.com/</a></td>
<td>Discovery Online</td>
<td>(3)</td>
<td>0.1550</td>
</tr>
<tr>
<td><a href="http://www.film.com/">http://www.film.com/</a></td>
<td>Welcome to Film.com</td>
<td>(3)</td>
<td>0.1533</td>
</tr>
<tr>
<td><a href="http://www.mgmua.com/">http://www.mgmua.com/</a></td>
<td>mgm online</td>
<td>(3)</td>
<td>0.1300</td>
</tr>
</tbody>
</table>

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:

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

Principal Community, SALSA:

<table>
<thead>
<tr>
<th>url</th>
<th>title</th>
<th>cat</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.nrlc.org/">http://www.nrlc.org/</a></td>
<td>National Right To Life</td>
<td>(3)</td>
<td>0.3443</td>
</tr>
<tr>
<td><a href="http://www.prolife.org/ultimate/">http://www.prolife.org/ultimate/</a></td>
<td>The Ultimate Pro-Life Resource List</td>
<td>(3)</td>
<td>0.2847</td>
</tr>
<tr>
<td><a href="http://www.naral.org/">http://www.naral.org/</a></td>
<td>NARAL Choice for America</td>
<td>(3)</td>
<td>0.2402</td>
</tr>
<tr>
<td><a href="http://www.feminist.org/">http://www.feminist.org/</a></td>
<td>Feminist Majority Foundation</td>
<td>(3)</td>
<td>0.1808</td>
</tr>
<tr>
<td><a href="http://www.now.org/">http://www.now.org/</a></td>
<td>National Organization for Women</td>
<td>(3)</td>
<td>0.1739</td>
</tr>
<tr>
<td><a href="http://www.cais.com/agm/main/index.html">http://www.cais.com/agm/main/index.html</a></td>
<td>The Abortion Rights Activist</td>
<td>(1)</td>
<td>0.1661</td>
</tr>
<tr>
<td><a href="http://www.gynpages.com/">http://www.gynpages.com/</a></td>
<td>Abortion Clinics Online</td>
<td>(3)</td>
<td>0.1631</td>
</tr>
<tr>
<td><a href="http://www.plannedparenthood.org/">http://www.plannedparenthood.org/</a></td>
<td>Planned Parenthood Federation</td>
<td>(3)</td>
<td>0.1572</td>
</tr>
<tr>
<td><a href="http://www.all.org/">http://www.all.org/</a></td>
<td>What's new at American Life League</td>
<td>(3)</td>
<td>0.1424</td>
</tr>
<tr>
<td><a href="http://www.hli.org/">http://www.hli.org/</a></td>
<td>Human Life International</td>
<td>(3)</td>
<td>0.1424</td>
</tr>
</tbody>
</table>

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 sites and 4 pro-life sites (which are the same top 4 pro-life sites 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:**

<table>
<thead>
<tr>
<th>url</th>
<th>title</th>
<th>cat</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.aic.nrl.navy.mil/galist/">http://www.aic.nrl.navy.mil/galist/</a></td>
<td>The Genetic Algorithms Archive</td>
<td>(3)</td>
<td>0.2785</td>
</tr>
<tr>
<td><a href="http://alife.santafe.edu/">http://alife.santafe.edu/</a></td>
<td>Artificial Life Online</td>
<td>(3)</td>
<td>0.2762</td>
</tr>
<tr>
<td><a href="http://www.yahoo.com/">http://www.yahoo.com/</a></td>
<td>Yahoo!</td>
<td>(3)</td>
<td>0.2736</td>
</tr>
<tr>
<td><a href="http://www.geneticprogramming.com/">http://www.geneticprogramming.com/</a></td>
<td>The Genetic Programming Notebook</td>
<td>(1)</td>
<td>0.2559</td>
</tr>
<tr>
<td><a href="http://galt.ge.uiuc.edu/illigal.home.html">http://galt.ge.uiuc.edu/illigal.home.html</a></td>
<td>illigAL Home Page</td>
<td>(3)</td>
<td>0.2357</td>
</tr>
<tr>
<td><a href="http://www.cs.gmu.edu/research/gag/">http://www.cs.gmu.edu/research/gag/</a></td>
<td>The Genetic Algorithms Group...</td>
<td>(3)</td>
<td>0.2012</td>
</tr>
<tr>
<td><a href="http://www.scs.carleton.ca/csgs/resources/gaal.html">http://www.scs.carleton.ca/csgs/resources/gaal.html</a></td>
<td>Genetic Algorithms and Artificial Life Resources</td>
<td>(1)</td>
<td>0.1813</td>
</tr>
<tr>
<td><a href="http://lancet.mit.edu/ga/">http://lancet.mit.edu/ga/</a></td>
<td>GAlib: Matthew’s Genetic Algorithms Library</td>
<td>(3)</td>
<td>0.1812</td>
</tr>
</tbody>
</table>

**Principal Community, SALSA:**

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

Table 7: Authorities for WWW query "genetic"
6 SALSA and the In/Out Degrees of Sites

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 site confers to the pointed site. 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 site discusses the topic at hand ([8]).

- One of the link's endpoints being designated by the user as highly relevant to the search topic. When a site is known to be a good authority, it seems reasonable to raise the weights of the links which enter that site. Similarly, when a site 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 [6]. We coin it the anchor sites approach, since it uses user-designated sites 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.

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 given in ap-
Appendix A.)

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

- The weighted in-degree of site $i \in A$:
  \[ d_{\text{in}}(i) \triangleq \sum_{k \in H \; k \rightarrow i} w(k \rightarrow i) \]

- The weighted out-degree of site $k \in H$:
  \[ d_{\text{out}}(k) \triangleq \sum_{i \in A \; k \rightarrow i} w(k \rightarrow i) \]

- The sum of edge weights:
  \[ W = \sum_{i \in A} d_{\text{in}}(i) = \sum_{k \in H} d_{\text{out}}(k) \]

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

\[ P_A(i, j) = \sum_{k \in H \; k \rightarrow i, \; k \rightarrow j} \frac{w(k \rightarrow i)}{d_{\text{in}}(i)} \cdot \frac{w(k \rightarrow j)}{d_{\text{out}}(k)} \]

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 $k, l \in H$:

\[ P_H(k, l) = \sum_{i \in A \; k \rightarrow i, \; l \rightarrow i} \frac{w(k \rightarrow i)}{d_{\text{out}}(k)} \cdot \frac{w(l \rightarrow i)}{d_{\text{in}}(i)} \]

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

\[ R_A = \{(i, j) \mid P_A(i, j) > 0\} \]

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 $P(i, j) > 0$ implies $P(j, i) > 0$. 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:

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

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

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

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

By the Ergodic Theorem ([11]), any irreducible, aperiodic Markov chain has a unique stationary distribution vector. It will therefore suffice to show that the vector $\pi$ with the properties claimed in the proposition is indeed a stationary distribution vector of $M_A$.

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

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

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

$$[\pi P_A]_i = \sum_{j \in A} \pi_j P_A(j, i)$$

$$= \sum_{j \in A} \frac{d_{in}(j)}{W} \sum_{k \in H \mid k \rightarrow i, k \rightarrow j} \frac{w(k \rightarrow j) \cdot w(k \rightarrow i)}{d_{out}(k)}$$

$$= \frac{1}{W} \sum_{j \in A} \sum_{k \in H \mid k \rightarrow i, k \rightarrow j} \frac{w(k \rightarrow j) \cdot w(k \rightarrow i)}{d_{out}(k)}$$

$$= \frac{1}{W} \sum_{k \in H \mid k \rightarrow i} \sum_{j \in A \mid k \rightarrow j} \frac{w(k \rightarrow j) \cdot w(k \rightarrow i)}{d_{out}(k)}$$

$$= \frac{1}{W} \sum_{k \in H \mid k \rightarrow i} \frac{w(k \rightarrow i)}{d_{out}(k)} \sum_{j \in A \mid k \rightarrow j} w(k \rightarrow j)$$
Thus, when the (undirected) support graph of \( G \) is connected, SALSA assigns each site an authority weight which is proportional to the sum of weights of its incoming edges. The hub weight of each site is proportional to the sum of weights of its outgoing edges. In unweighted collections (with all edges having unit weight), each site’s Stochastic authority (hub) weight is simply proportional to the in(out) degree of the site.

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 sites, 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 \( A_i \cap A_j = \emptyset \) for all \( i \neq j \). 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 \( \epsilon \) 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 \( \pi^c(j) \) the entry which corresponds to \( j \) in \( \pi^c \) (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:

$$\lim_{n \to \infty} \epsilon P_A^n = \tilde{\pi} \quad \text{where} \quad \tilde{\pi}_j = \frac{|A_i(j)|}{|A|} \cdot \pi^{c(j)}$$

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

$$p_i^n = \sum_{j \in A_i} \epsilon_j = \frac{|A_i|}{|A|}$$

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

This proposition points out a natural way to compare the authoritativeness of sites from different irreducible components: Simply multiply each site’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 sites 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 site’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 sites 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 ([18]). 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 site belongs to). To gain more perspective on this apparent contradiction, 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-sites 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 sites 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 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 steps involved in the first stage are crucial to the success of the entire algorithm.

Considerable research efforts have been invested in improving the quality of the assembled graphs. The current state of the art techniques for these steps is now such that in many cases, simple (and efficient) algorithms and heuristics produce quite satisfying results on the assembled graphs.

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 sites correctly. It is entirely irrelevant if the site in place 98 is really better than the site in place 216. The Stochastic ranking, which turns out to be equivalent to a weighted in-degree ranking, discovers the most authoritative sites 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 sites (which constitute the vast majority of the collection).
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 sites 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 sites (or from finding authorities on all meanings/aspects of the query):

- 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.

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

We note that SALSA is less vulnerable to the TKC effect, and produces good results in many cases where the Mutual Reinforcement approach fails to do so.

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


[26] Ronny Roth. private communication.


A Mathematical Background

A.1 Irreducible Matrices

Let $B = [b_{i,j}]$ denote a square $n \times n$ real matrix with nonnegative entries. Denote by $\lambda_1(B), \lambda_2(B), \ldots, \lambda_n(B)$ the $n$ eigenvalues of $B$, ordered by non-increasing absolute value. In particular, $|\lambda_1(B)|$ is the spectral radius of $B$ (\cite{20}), and will be denoted $\lambda(B)$ (hence $\lambda(B)$ is a nonnegative real number).

Denote by $G(B)$ the (directed) support graph of $B$ (\cite{23}): $G(B)$ has $n$ nodes (corresponding to the $n$ rows of $B$), with a directed edge $i \to j$ if and only if $[B]_{i,j} = b_{i,j} > 0$.

**Definition 1** (\cite{23}) A nonnegative real square $n \times n$ matrix $B$ is irreducible if for every $1 \leq i, j \leq n$ there exists a nonnegative integer $l \geq 0$ such that $[B^l]_{i,j} > 0$.

**Definition 2** A directed graph $G = (V, E)$ is called irreducible if for every $i, j \in V$ there is a path in $G$ originating in $i$ and ending in $j$.

**Lemma 1** $B$ is irreducible if and only if $G(B)$ is irreducible (\cite{23}).

We now bring a version of the Perron-Frobenius Theorem, tailored for our needs.

**Theorem 1** (Perron-Frobenius Theorem for irreducible matrices,\cite{15}) Let $B$ be an irreducible matrix. Then

1. $\lambda(B) > 0$

2. $\lambda(B)$ is a simple eigenvalue of $B$ ($\lambda(B)$ is a simple root of the characteristic polynomial of $B$).

3. $B$ has positive (i.e. all components are positive) left and right eigenvectors corresponding to $\lambda(B)$.

**Lemma 2** Let $B = [b_{i,j}]$ be an irreducible $n \times n$ matrix. A sufficient condition which guarantees that $|\lambda_1(B)| > |\lambda_2(B)|$ is that for some $1 \leq i \leq n$, $b_{i,i} > 0$ (\cite{15}).

**Corollary 1** Let $B$ be an irreducible matrix for which $|\lambda_1(B)| > |\lambda_2(B)|$, and let $w$ be a real eigenvector of $B$ which does not correspond to $\lambda_1(B)$. Then $w$ has both positive and negative entries (\cite{11},\cite{20}).
Let $M$ be an irreducible $n \times n$ matrix with some non-zero main diagonal entry. We conclude that:

1. $\lambda(M) = \lambda_1(M) > |\lambda_2(M)|$

2. There is a unique positive unit eigenvector of $M$ corresponding to $\lambda(M)$, which we will denote by $v_{\lambda(M)}$. That is, every component of $v_{\lambda(M)}$ is positive, and $\|v_{\lambda(M)}\| = 1$.

### A.2 Irreducible Stochastic Matrices

A nonnegative real square $n \times n$ matrix $P = [p_{i,j}]$ is **stochastic** if for every row index $1 \leq i \leq n$,

$$\sum_{j=1}^{n} p_{i,j} = 1$$

**Definition 3** The period of a graph $G$ is the greatest common divisor of the lengths of all cycles in $G$. When $G$ has a period of 1, we say that $G$ is **aperiodic**.

**Definition 4** A matrix $B$ is called **primitive** if $G(B)$ is aperiodic.

**Theorem 2** (Ergodic Theorem, [11]) Let $P$ be an irreducible primitive stochastic matrix.

1. $\lambda(P) = \lambda_1(P) = 1$, and any other eigenvalue $\tilde{\lambda}$ of $P$ satisfies $|\tilde{\lambda}| < 1$.

2. There is a unique distribution row-vector $^2\pi_P$ which satisfies:

$$\pi_P \cdot P = \pi_P$$

The distribution $\pi_P$ is called the stationary distribution of the Markov chain defined by the (transition) matrix $P$.

3. For any distribution row-vector $q$:

$$\lim_{n \to \infty} q \cdot P^n = \pi_P$$

---

$^2$A nonnegative real vector whose sum of entries equals 1.
B Proof of propositions

Here we prove the propositions concerning the TKC effect. The reader should recall the constructions of the collections $\tilde{C}_k$ and $\tilde{C}_k$ from section 5.1.

We will first prove the propositions which apply to SALSA. We shall explicitly prove proposition 3; proposition 1 then follows as a special case where $A_b = \phi$.

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

**Proof:** The authority chain $M_A$ which results from $\tilde{C}_k$ is irreducible, since for every two authorities $i, j \in \tilde{C}_k$, $P_A(i, j) > 0$. This follows from the fact that for every two authorities, there exists at least one hub which links to both. Therefore, we can apply Proposition 5 and deduce that:

- for any $i \in C_i$:
  \[ \pi_i = \frac{d_{in}(i)}{W} = \frac{(n-1)}{k} + m \]

- for any $j \in C_s \setminus A_b$:
  \[ \pi_j = \frac{d_{in}(j)}{W} = \frac{h_s + n}{W} = \frac{(n-1)}{k} - n + n = \frac{(n-1)}{k} \]

- for any $t \in A_b$:
  \[ \pi_t = \frac{d_{in}(t)}{W} = \frac{h_s + n + (m + 1)}{W} = \frac{(n-1)}{k} + m + 1 \]

It follows that for any $i \in C_i, j \in C_s \setminus A_b, t \in A_b$ : $\pi_t > \pi_i > \pi_j$.

\[ \square \]

Before proving the claims about the rankings which are produced by the Mutual Reinforcement approach, we first consider a few properties of irreducible matrices.

**Definition 5** An $n \times n$ matrix $A = [a_{r,s}]$ is said to have the $(i, j)$-switch property ($1 \leq i, j \leq n, i \neq j$) if:

- $a_{i,i} + a_{i,j} = a_{j,i} + a_{j,j}$
• For all \( k \neq i, j \) : \( a_{i,k} = a_{j,k} \)

**Lemma 3** Let \( A = [a_{r,s}] \) be an \( n \times n \) matrix with the \((i,j)\)-switch property (for some \( i \neq j \)). Let \( \lambda \) be an eigenvalue of \( A \), and let \( w = (w_1, \ldots, w_n)^T \) be a corresponding eigenvector. Then: \( \lambda \neq a_{i,i} - a_{i,j} \implies w_i = w_j \).

**Proof:** Since \( w \) is an eigenvector which corresponds to \( \lambda \), we have:

\[
[Aw]_i = a_{i,i}w_i + \sum_{l \neq i,j} a_{i,l}w_l = \lambda \cdot w_i
\]

\[
[Aw]_j = a_{j,i}w_i + \sum_{l \neq i,j} a_{j,l}w_l = \lambda \cdot w_j
\]

Subtracting the second equation from the first, we get:

\[
(a_{i,i} - a_{j,i})w_i + (a_{i,j} - a_{j,j})w_j = \lambda(w_i - w_j)
\]

Since \( a_{i,i} - a_{j,i} = a_{j,i} - a_{i,j} \), we get:

\[
(a_{i,i} - a_{j,i})(w_i - w_j) = \lambda(w_i - w_j)
\]

Hence, \( \lambda \neq a_{i,i} - a_{j,i} \implies w_i - w_j = 0 \).

\[\square\]

**Lemma 4** Let \( A = [a_{r,s}] \) be a non-negative irreducible \( n \times n \) matrix, \( n > 1 \). For any \( 1 \leq i \leq n \), \( \lambda(A) > a_{i,i} \).

**Proof:** Let \( w = (w_1, \ldots, w_n)^T \) denote the positive eigenvector which corresponds to \( \lambda(A) \) (Perron-Frobenius Theorem). For any \( 1 \leq i \leq n \), we have:

\[
\lambda(A) \cdot w_i = a_{i,i} \cdot w_i + \sum_{j \neq i} a_{i,j} \cdot w_j
\]

Note that all the products \( a_{i,j} \cdot w_j \) are non-negative, since both \( a_{i,j} \geq 0 \) and \( w_j > 0 \) for all \( i, j \). Since \( A \) is irreducible, there is at least one index \( 1 \leq k \leq n \), \( k \neq i \) such that \( a_{i,k} > 0 \). Otherwise, in the support graph \( G(A) \), there will be no paths from node \( i \) to any node \( j \neq i \) and \( A \) would not be irreducible. Hence:

\[
\lambda(A) \cdot w_i = a_{i,i} \cdot w_i + a_{i,k} \cdot w_k + \sum_{j \neq i,k} a_{i,j} \cdot w_j \geq a_{i,i} \cdot w_i + a_{i,k} \cdot w_k > a_{i,i} \cdot w_i
\]

And since \( w_i > 0 \) we get \( \lambda(A) > a_{i,i} \).
The above proof is due in part to Ronny Roth ([26]).

**Corollary 2** Let $A = [a_{i,j}]$ be a non-negative irreducible $n \times n$ matrix, $n > 1$, with the $(i,j)$-switch property. Let $w = (w_1, \ldots, w_n)^T$ denote the positive eigenvector which corresponds to $\lambda(A)$. Then $w_i = w_j$.

**Proof:** By Lemma 4, $\lambda(A) > a_{i,i} \geq a_{i,i} - a_{j,i}$. Therefore, by Lemma 3, $w_i = w_j$.

**Definition 6** An $n \times n$ matrix $A = [a_{i,j}]$ is said to have the $(i,j)$-dominance property $(1 \leq i, j \leq n, i \neq j)$ if:

- $a_{i,i} + a_{i,j} > a_{j,i} + a_{j,j}$
- For all $k \neq i, j$ : $a_{i,k} \geq a_{j,k}$

**Lemma 5** Let $A = [a_{i,j}]$ be a non-negative, irreducible $n \times n$ matrix with the $(i,j)$-dominance property for some $i \neq j$. Let $w = (w_1, \ldots, w_n)^T$ denote the positive eigenvector which corresponds to $\lambda(A)$. Then $w_i > w_j$.

**Proof:** Let $\lambda \triangleq \lambda(A)$. By the definition of $w$ we have:

$$[Aw]_i = a_{i,i}w_i + a_{i,j}w_j + \sum_{l \neq i,j} a_{i,l}w_l = \lambda \cdot w_i$$

$$[Aw]_j = a_{j,i}w_i + a_{j,j}w_j + \sum_{l \neq i,j} a_{j,l}w_l = \lambda \cdot w_j$$

Subtracting the second equation from the first, we get:

$$(a_{i,i} - a_{j,i})w_i + (a_{i,j} - a_{j,j})w_j + \sum_{l \neq i,j} (a_{i,l} - a_{j,l})w_l = \lambda w_i - \lambda w_j$$

By the dominance of row $i$ over row $j$, we have:

$$(a_{i,i} - a_{j,i})w_i + (a_{i,j} - a_{j,j})w_j \leq \lambda w_i - \lambda w_j ,$$

which implies that

$$(\lambda - (a_{i,i} - a_{j,i})w_i \geq (\lambda - (a_{j,j} - a_{i,j}))w_j .$$
Since \( \lambda > a_k,k \) for all \( k \), both \((\lambda-(a_{i,j} - a_{i,i}))\) and \((\lambda-(a_{j,i} - a_{i,i}))\) are positive. By the \((i,j)\) dominance property, we have that \( a_{i,i} - a_{j,i} > a_{j,i} - a_{i,i} \). Therefore,

\[
\frac{w_i}{w_j} > \frac{\lambda - (a_{j,i} - a_{i,i})}{\lambda - (a_{i,i} - a_{j,i})} > 1,
\]

which completes the proof.

We now prove proposition 4. Proposition 2 will follow as a special case, where \( A_b = \phi \) (\(|A_b| = b = 0\)).

**Proposition 4** On the collection \( \tilde{C}_k \), 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.

**Proof:** Let \( W \) denote the adjacency matrix of \( \tilde{C}_k \), and consider the co-citation matrix \( W^T W \): The rows and columns which correspond to the hubs of \( \tilde{C}_k \) will contain only zeros, and we can analyze the ranking produced by the Mutual Reinforcement approach by considering only the sub-matrix of \( W^T W \) which contains the rows and columns which correspond to the authorities of \( \tilde{C}_k \). Denote this sub-matrix, which is positive (and thus clearly irreducible) by \( A = [a_{r,s}] \). \( A \) (like the co-citation matrix \( W^T W \)) is symmetric and has the following structure:

- For all \( t, a_{t,t} \) is the in-degree of \( t \). Therefore,
  - \( t \in C_l \Rightarrow a_{t,t} = \binom{n-1}{k-1} + m \)
  - \( t \in C_s \setminus A_b \Rightarrow a_{t,t} = \binom{n-1}{k-1} \)
  - \( t \in A_b \Rightarrow a_{t,t} = \binom{n-1}{k-1} + m + 1 \)

- For all \( t \in C_l, s \in C_s : a_{t,s} = a_{s,t} = 1 \).

- For all \( t_1, t_2 \in C_l \) \( (t_1 \neq t_2) : a_{t_1,t_2} = \binom{n-2}{k-2} \)

- For all \( t_1, t_2 \in A_b \) \( (t_1 \neq t_2) : a_{t_1,t_2} = h_s + (m + 1) = \binom{n-1}{k-1} - n + m + 1 \)

- For all \( t \in C_s \setminus A_b, s \in C_s \) \( (t \neq s) : a_{t,s} = a_{s,t} = h_s = \binom{n-1}{k-1} - n \)

Let \( \lambda \triangleq \lambda(A) \), and denote by \( w \) the unique positive unit eigenvector which corresponds to \( \lambda \). By the above, we have:
• For all $i, j \in A_b$, $A$ has the $(i, j)$-switch property.
• For all $i, j \in C_s \setminus A_b$, $A$ has the $(i, j)$-switch property.
• For all $i, j \in C_l$, $A$ has the $(i, j)$-switch property.
• For all $i \in A_b, j \in C_s \setminus A_b$, $A$ has the $(i, j)$-dominance property.

Thus there exist three positive values $\alpha, \beta, \gamma$ so that for all $i \in A_b$, $w_i = \alpha$, for all $j \in C_s \setminus A_b$, $w_j = \beta$ and for all $t \in C_l$, $w_t = \gamma$. In addition, $\alpha > \beta$. It remains to show that $\beta > \gamma$, which is what we do next. Choose arbitrary indices $i \in C_s \setminus A_b$ and $j \in C_l$. Then:

$$[Aw]_i = \lambda w_i = \lambda \beta = \alpha \sum_{t \in A_b} a_{i,t} + \beta \sum_{t \in C_s \setminus A_b} a_{i,t} + \gamma \sum_{t \in C_l} a_{i,t}$$

$$[Aw]_j = \lambda w_j = \lambda \gamma = \alpha \sum_{t \in A_b} a_{j,t} + \beta \sum_{t \in C_s \setminus A_b} a_{j,t} + \gamma \sum_{t \in C_l} a_{j,t}$$

Subtracting the second equation from the first, we get:

$$\lambda (\beta - \gamma) = \alpha \left( \sum_{t \in A_b} a_{i,t} - b \right) + \beta \left( \sum_{t \in C_s \setminus A_b} a_{i,t} - (m - b) \right) + \gamma \left( n - \sum_{t \in C_l} a_{j,t} \right)$$

$$\geq \beta \left( \sum_{t \in A_b} a_{i,t} - b \right) + \beta \left( \sum_{t \in C_s \setminus A_b} a_{i,t} - (m - b) \right) + \gamma \left( n - \sum_{t \in C_l} a_{j,t} \right)$$

Reorganizing the inequality yields:

$$\beta \left[ \lambda - \left( \sum_{t \in C_s} a_{i,t} - m \right) \right] \geq \gamma \left[ \lambda - \left( \sum_{t \in C_l} a_{j,t} - n \right) \right]$$

We now show a couple of short claims:
1. \( \lambda - (\sum_{t \in C} a_{i,t} - m) > 0: \)

\[
\lambda \beta = \alpha \sum_{t \in A_i} a_{i,t} + \beta \sum_{t \in C_i \setminus A_i} a_{i,t} + \gamma \sum_{t \in C_i} a_{i,t}
\geq \beta \sum_{t \in C_i} a_{i,t} + \gamma \sum_{t \in C_i} a_{i,t}
> \beta \sum_{t \in C_i} a_{i,t} > \beta (\sum_{t \in C_i} a_{i,t} - m)
\]

Dividing both sides by the positive constant \( \beta \) completes the claim.

2. \( \sum_{t \in C} a_{i,t} - m > \sum_{t \in C_i} a_{j,t} - n \). To prove this, we evaluate both expressions:

\[
\sum_{t \in C_i} a_{i,t} - m = (m - 1) h + \left( \frac{n - 1}{k - 1} \right) - m
= (m - 1) \left[ \frac{n - 1}{k - 1} \right] - n + \left( \frac{n - 1}{k - 1} \right) - m
= m \left( \frac{n - 1}{k - 1} \right) - n m + (n - m)
\]

\[
\sum_{t \in C_i} a_{j,t} - n = (n - 1) \left( \frac{n - 2}{k - 2} \right) + \left( \frac{n - 1}{k - 1} \right) + m - n
= (k - 1) \left[ \frac{n - 1}{k - 1} \right] - (n - m)
\]

Using the equality \( m = k + 1 = \sqrt{n} \), we now subtract the second expression from the first:

\[
(\sum_{t \in C_i} a_{i,t} - m) - (\sum_{t \in C_i} a_{j,t} - n) = \left( \frac{n - 1}{k - 1} \right) - n m + 2(n - m)
> \left( \frac{n - 1}{k - 1} \right) - n^2 > 0 \ \forall k \geq 3
\]
Using the first claim we can transform
\[
\beta \left[ \lambda - \left( \sum_{i \in C} a_{i,t} - m \right) \right] \geq \gamma \left[ \lambda - \left( \sum_{i \in C} a_{i,t} - n \right) \right]
\]
into
\[
\frac{\beta}{\gamma} \geq \frac{\lambda - \left( \sum_{i \in C} a_{i,t} - n \right)}{\lambda - \left( \sum_{i \in C} a_{i,t} - m \right)},
\]
and by the second claim we deduce that
\[
\frac{\beta}{\gamma} > 1 \implies \beta > \gamma
\]
which completes the proof. 
\[\square\]