

- $f(x)$ is a strictly increasing function of x , for all $x \in \mathbf{R} \setminus \{274\}$.

Hence, for all $x < 1$: $f(x) < f(1) < f(b) \implies b > 1 = c$, which completes the proof.

□

The proof is due in part to Ronny Roth ([27]).

We bring an additional lemma, which can be used for an alternative proof of proposition 2.

Lemma 4 *Let $A = [a_{i,j}]$ be a non-negative, irreducible $n \times n$ matrix with an (i, j) -switch property. Let λ be an eigenvalue of A , and let $w = (w_1, \dots, w_n)^T$ 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 λ , we have:

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

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

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 $a_{i,i} - a_{j,i} = a_{j,j} - a_{i,j}$, we get:

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

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

□

By [24], for any square irreducible $n \times n$ matrix B :

$$\lambda(B) \geq \min_{1 \leq i \leq n} \sum_{j=1}^n b_{i,j}$$

In our case, this lower bound yields $\lambda(A_k) \geq 1662$.

Exploiting the structure of A_k and its principal eigenvector, we arrive at the following 3×3 equation set:

$$\begin{pmatrix} 286 + 274 + 2k & 4 \cdot 274 & 12 \\ 2 \cdot 274 & 3 \cdot 274 + 286 & 12 \\ 2 & 4 & 11 \cdot 120 + 336 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \lambda \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

When scaling the eigenvector to obtain $c = 1$, we get:

$$\begin{aligned} (560 + 2k)a + 1096b + 12 &= \lambda a \\ 548a + 1108b + 12 &= \lambda b \\ 2a + 4b + 1656 &= \lambda \end{aligned}$$

Subtracting the second equation from the first yields:

$$(12 + 2k)a - 12b = \lambda(a - b)$$

or:

$$(\lambda - 12)(a - b) = 2k \cdot a > 0$$

Since $\lambda \geq 1662$, the term $(\lambda - 12)$ is positive, implying that $a - b > 0$, or $a > b$.

In order to prove $b > c$, we subtract 274 times the third equation from the second equation, to arrive at:

$$(1108 - 274 \cdot 4)b + 12 - 274 \cdot 1656 = \lambda(b - 274)$$

or:

$$12(b - 37811) = \lambda(b - 274)$$

Applying our lower bound for λ , we get:

$$138.5 = \frac{1662}{12} \leq \frac{\lambda}{12} = \frac{b - 37811}{b - 274}$$

Defining $f(x) \triangleq \frac{x-37811}{x-274}$, we get:

- $f(1) = 138.498 < 138.5 \leq f(b)$

1. For the first coordinate of $A\tilde{w}$:

$$\begin{aligned}
[A\tilde{w}]_1 &= a_{1,1}w_1 + a_{1,2}w_1 + \sum_{l=3}^n a_{1,l}w_l = \\
&= \sum_{l=1}^n a_{1,l}w_l + a_{1,2}(w_1 - w_2) = \\
&= \lambda \cdot w_1 + a_{1,2}(w_1 - w_2) > \lambda \cdot w_1 = \lambda \cdot [\tilde{w}]_1
\end{aligned}$$

2. For the second coordinate of $A\tilde{w}$:

$$\begin{aligned}
[A\tilde{w}]_2 &= a_{2,1}w_1 + a_{2,2}w_1 + \sum_{l=3}^n a_{2,l}w_l = \\
&= a_{1,1}w_1 + a_{1,2}w_1 + \sum_{l=3}^n a_{1,l}w_l = \\
&= a_{1,2}(w_1 - w_2) + \sum_{l=1}^n a_{1,l}w_l > \lambda \cdot w_1 = \lambda \cdot [\tilde{w}]_2
\end{aligned}$$

3. For all other coordinates ($3 \leq k \leq n$) of $A\tilde{w}$:

$$\begin{aligned}
[A\tilde{w}]_k &= a_{k,1}w_1 + a_{k,2}w_1 + \sum_{l=3}^n a_{k,l}w_l = \\
&= a_{k,2}(w_1 - w_2) + \sum_{l=1}^n a_{k,l}w_l = \\
&= a_{k,2}(w_1 - w_2) + \lambda \cdot w_k > \lambda \cdot w_k = \lambda \cdot [\tilde{w}]_k
\end{aligned}$$

Thus $A\tilde{w} > \lambda \cdot \tilde{w} \implies \lambda$ is not the principal eigenvalue of A .

□

Proof of proposition 2: The matrix A_k is irreducible. Therefore, by Theorem 1 it has a positive principal eigenvector. Part 2 of the proposition follows from lemma 3, since A_k has the (i, j) -switch property for all

$$(i, j) \in \{(1, 2)\} \cup \{3, 4, 5, 6\} \times \{3, 4, 5, 6\} \cup \{7, \dots, 18\} \times \{7, \dots, 18\}$$

It remains to prove that $a > b > c$.

Proposition 2 For any $k > 0$, the matrix A_k :

$$\begin{pmatrix} 286+k & 274+k & 274 & 274 & 274 & 274 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 274+k & 286+k & 274 & 274 & 274 & 274 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 274 & 274 & 286 & 274 & 274 & 274 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 274 & 274 & 274 & 286 & 274 & 274 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 274 & 274 & 274 & 274 & 286 & 274 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 336 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 336 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 336 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 336 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 336 & 120 & 120 & 120 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 120 & 336 & 120 & 120 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 120 & 120 & 336 & 120 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 336 & 120 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 336 & 120 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 336 & 120 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 336 & 120 \\ 1 & 1 & 1 & 1 & 1 & 1 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 120 & 336 \end{pmatrix}$$

has the following properties:

1. A_k has a positive principal eigenvector.
2. The positive principal eigenvector of A_k is of the form:

$$(a, a, b, b, b, b, c, c, c, c, c, c, c, c, c, c, c, c)^T$$

3. The entries of A_k 's principal eigenvector satisfy $a > b > c > 0$.

Before proving the proposition, we develop a few tools.

Definition 5 A non-negative, irreducible $n \times n$ matrix $A = [a_{i,j}]$ is said to have the (i, j) -switch property ($i \neq j$) 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 $\lambda > 0$ denote the principal eigenvalue of a positive $n \times n$ matrix $A = [a_{i,j}]$ with an (i, j) -switch property, and let $w = (w_1, \dots, w_n)^T$ be a corresponding positive principal eigenvector. Then $w_i = w_j$.

Proof: Assume to the contrary that $w_i \neq w_j$. Without loss of generality, let $i = 1, j = 2$ and let $w_1 > w_2$.

We define $\tilde{w} \triangleq (w_1, w_1, w_3, \dots, w_n)^T$. We will show that $A\tilde{w} > \lambda \cdot \tilde{w}$, thus contradicting the choice of λ as the principal eigenvalue of A ([20]).

- Property 6 implies that every topic is incident to the same number of hubs.
- Property 3 implies that every pair of distinct hubs is incident with the same number of topics.

An incidence relation on two sets of objects which exhibits properties 2,3 and 6 is called a *balanced incomplete block design (BIBD)* or as a (v, k, λ) *block design* with $(v = n_h, k = \frac{n_h \cdot c}{|T|}$ and $\lambda = \tau)$ ([22]).

In this proof we consider a smaller family of designs, called *symmetric BIBD's*. Here we add the constraint $|T| = n_h$. This implies that not only does each hub cover c topics, but also each topic is covered by $k = c$ hubs.

Such symmetric designs are closely related to *projective geometries*, and infinitely many of them exist. See [22] for constructions of block designs.

We now show how to expand a symmetric BIBD incidence relation between hubs and topics into a collection of $|T| \cdot A$ authorities and n_h hubs satisfying all eight properties. Note that requirements 2,3 and 6 have already been met.

For every topic $t \in T$, there are $k = c$ hubs which cover it. Set $A \triangleq k (= c)$ and $a \triangleq A - 1 (= k - 1)$. Let each t -hub cover a distinct set of a t -authorities: There are $\binom{A}{a} = \binom{A}{A-1} = A = k$ such sets, which is exactly the number of t -hubs. This satisfies properties 4 and 7 (as any authority will be pointed at by exactly $k - 1$ hubs). Also, every pair of t -authorities will be co-pointed at by $k - 2$ hubs, satisfying property 8.

Every two t -hubs cover different $(A - 1)$ -sized sets of the A t -authorities, hence they co-cover $\alpha = A - 2$ of the t -authorities. This satisfies property 5.

□

The proof is due in part to Tuvi Etzion ([9]).

Appendix A

Proofs of propositions

Proposition 1 *There are infinitely many symmetric, orthogonal T – topic collections, consisting of $|T| \cdot A$ authorities (A distinct perfect authorities per topic) and n_h hubs, which exhibit the following (not necessarily independent) properties:*

1. *The set of authorities and the set of hubs are disjoint.*
2. *Every hub covers c topics.*
3. *Every two hubs co-cover τ topics.*
4. *For each $t \in T$, every t -hub points to exactly a of the A t -authorities.*
5. *For each $t \in T$, every two t -hubs co-point to α of the A t -authorities.*
6. *Every topic is covered by the same number of hubs.*
7. *Every authority is pointed at the same number of times.*
8. *For each $t \in T$, every two t -authorities are co-pointed at by the same number of hubs.*

Proof: Define an incidence relation between the n_h hubs and the $|T|$ topics as follows: A hub h and topic t are incident if h covers t . For the time being, we ignore the individual authorities.

Consider properties 2, 3 and 6:

- Property 2 implies that every hub is incident to the same number of topics.

- [25] 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.
- [26] 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.
- [27] Ronny Roth. private communication.
- [28] 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.
- [29] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.
- [30] R. Weiss, B. Vález, M. Sheldon, C. Namprempre, P. Szilagy, A. Duda, and D. Gifford. Hypersuit: A hierarchical network search engine that exploits content-link hypertext clustering. *Proc. 7th ACM Conference on Hypertext*, 1996.

- [11] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.
- [12] D. Gibson, J. Kleinberg, and P. Raghavan. Structural analysis of the world wide web. *Web Characterisation Workshop*, November 1998. <http://www.w3.org/1998/11/05/WC-workshop/Papers/kleinber1.html>.
- [13] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. *Proc. 9th ACM Conference on Hypertext and Hypermedia*, 1998.
- [14] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- [15] Google Inc. Google search engine. <http://www.google.com/>.
- [16] Lycos Inc. Lycos internet guide. <http://www.lycos.com/>.
- [17] Alan Jennings. *Matrix Computation for engineers and scientists*. John Wiley & Sons, Ltd., 1977.
- [18] M.M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.
- [19] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [20] Peter Lancaster and Miron Tismenetsky. *The Theory of Matrices*. Academic Press, 1985.
- [21] Ken Law, Thomas Tong, and Alan Wong. Automatic categorization based on link structure, 1999. <http://www.stanford.edu/~tomtong/cs349/web.htm>.
- [22] Rudolf Lidl and Harald Niederreiter. *Finite Fields*. Addison-Wesley Publishing Company Inc., 1983.
- [23] Massimo Marchiori. The quest for correct information on the web: Hyper search engines. *Proc. 6th International WWW Conference*, 1997.
- [24] Brian H. Marcus, Ron M. Roth, and Paul H. Siegel. Constrained systems and coding for recording channels. Technical Report 0929, Technion - Israel Institute of Technology, March 1985.

Bibliography

- [1] J. Gary Auguston and Jack Minker. An analysis of some graph theoretical cluster techniques. *JACM*, 17(4):571–588, October 1970.
- [2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Proc. 7th International WWW Conference*, 1998.
- [3] Jeromy Carrière and Rick Kazman. Webquery: Searching and visualizing the web through connectivity. *Proc. 6th International WWW Conference*, 1997.
- [4] IBM Corporation Almaden Research Center. Clever. <http://www.almaden.ibm.com/cs/k53.clever.html>.
- [5] 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.
- [6] 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.
- [7] ThunderLink Communications. How to rank high in the search engines. <http://promotiontips.com/searchengine.shtml>.
- [8] Compaq Computer Corporation. Altavista net guide. <http://www.altavista.com/>.
- [9] TuvI Etzion. private communication.
- [10] Robert G. Gallager. *Discrete Stochastic Processes*. Kluwer Academic Publishers, 1996.

- Studying the performance of the algorithms on artificial multi-topic topologies. These topologies include collections of varying diversity (performance as a function of θ_c), and non-symmetric collections with topics of different weight structures.
- Benchmark artificial topologies could be devised, for testing and evaluating link-structure analyzing algorithms.
- Using directed-graph clustering techniques in conjunction with the Stochastic approach, in order to extract non-principal communities from the stochastic authority matrix.
- In [19], Kleinberg recounts that the Mutual Reinforcing approach *diffuses* when applied to a query which does not represent a topic of widespread interest on the WWW: The principal community found by the algorithm does not remain focussed on the topic of the query, but rather converges to a generalization of that topic which is of wide presence on the WWW. Following the results in section 4.1.1, it is possible that the δ -disparity coefficient can be used to avoid some cases of *diffusion*. Since the disparity coefficient further enhances the TKC effect, its use might enable the Mutual Reinforcing approach to find *small* tightly knit communities of authorities which otherwise do not score well.
- Currently, when applying the deletion method, the number of sites that are deleted between findings of successive communities is decided upon manually. Research might lead to determining this number in software, by examining the decline in the weights of a specific community or by some simple examination of the link structure within that community.

The main research goal is, however, to find efficient and effective search algorithms which utilize both the textual contents of hyperlinked documents and their link structure. Combining the information from both these sources should lead to improved page rankings than what is currently commercially available. Work in this direction has already begun ([2],[6],[13]), but there is still much to be done.

Chapter 5

Further Research and Conclusions

We have developed quantitative measures for assessing the informative link structures of hyperlinked collections, based on the paradigm of authorities and hubs. We have proposed a generalization to Kleinberg's Mutual Reinforcement approach for finding authorities and hubs ([19]), as well as a new Stochastic approach for the same task. Both approaches were tested on artificial topologies, on an electronic library of IEEE-IT documents, and on the WWW. The artificial topologies were built according to probabilistic models, which we have developed for hyperlinked collections, and by using combinatorial constructions. These tests, as well as analytical work, has shown that the rankings produced by both approaches do not always agree (specifically in the context of the TKC effect). Also demonstrated was the deletion method, an additional way to arrive at the non-principal communities of authorities in the Mutual Reinforcing approach, which compared well with the Gram-Schmidt method.

There is much room for future research on analyzing link structures. We point out some of the aspects which arise directly out of this work:

- Conducting sensitivity analysis of the approaches studied in this work (and new approaches) on artificial single-topic topologies: Finding how the ratio of authorities and hubs in the collection affects these algorithms, and how the performance of the algorithms varies as a function of the probabilities with which edges are drawn.

4.4.4 The Stochastic Approach

The Stochastic approach compares well with the Mutual Reinforcement approach (applied without the disparity coefficient). The main difference in the performance of the two approaches is that the Stochastic approach is less vulnerable to the TKC effect. In multi-topic collections this usually results in Stochastic principal communities which contain a mixture of authorities from the various topics of the collection. In single topic collections, this sometimes results in the Stochastic approach arriving at better principal communities than those of the Mutual Reinforcement approach (as in the case of the queries “java” and “movies” in section 4.3.1).

A major disadvantage of the Stochastic approach is that it converges slower than the Mutual Reinforcement approach. Its convergence can sometimes be expedited by applying (i, o) -trimming to the web graph G , but even then it still requires more computational resources. Also, currently we have not shown an ability to arrive at non-principal communities using the Stochastic approach. This issue is left for further research.

The acceleration caused by the use of trimming is not due just to the reduction in the number of required iterations in the power method. The time required for each iteration is also reduced: Recall that the WWW-graphs were relatively sparse. This means that (i, o) -trimming, even for small values of i and o , produces many isolated nodes in the graph. Isolated nodes result in all-zero-rows in the matrices A and H , and even naive implementations of the power method can be adapted to take advantage of zero-rows.

We conclude that trimming can accelerate both algorithmic approaches (especially the Stochastic approach) without jeopardizing the quality of the results.

4.4.2 The Effect of the δ Disparity Coefficient

The disparity coefficient has shown to help the Mutual Reinforcement approach handle the dense single topic artificial topologies of section 4.1.1.

Using positive values of δ further heightens the sensitivity of the Mutual Reinforcement approach to the TKC effect: It reduces the association between authorities that are pointed at by different hubs. This is exactly what enabled the Mutual Reinforcement approach to find the small tightly knit community of hubs and authorities in the dense artificial topologies (after the same approach failed to identify these communities with $\delta = 0$).

We conclude that this coefficient is useful in situations when users are interested in finding small, dense communities inside a much larger, noisy collection.

4.4.3 The Deletion Method

The deletion method has proven to be an effective tool for deriving non-principal communities in the Mutual Reinforcement approach. In all of the multi-topic queries we've run, it found the same communities as the Gram-Schmidt method, while being more efficient and not having the problem of reappearing communities. The disadvantage of this method, however, is the fact that it is not automated: The number of sites to delete between finding successive communities is currently set manually. This number must be set with care, as deleting too few sites will result in the same topic dominating more than one community, while deleting too many sites may result both in mixed-topic communities and in unrepresented topics.

4.4 Evaluating the Results

4.4.1 The Impact of Iterative (i, o) -Trimming

So far, all the results we have shown have been derived without the use of iterative (i, o) -trimming. We applied $(3, 3)$ -trimming to a few of the WWW collections which were discussed previously, and the output of running both algorithmic approaches on the resulting trimmed graphs indicates the following:

- The principal communities (resulting from both approaches) remained practically the same as those resulting from the untrimmed graphs. The ranking of the top sites in each collection changed very little, causing only minor changes in the principal communities.
- Applying trimming can significantly accelerate the convergence of the power method to the principal eigenvector, especially for the Stochastic approach.

Throughout all test cases, our convergence criterion was to run iterations of the power method until the maximal change of any coordinate between successive iterations did not exceed 0.00005. In the following table, we show the number of required iterations until convergence, with and without trimming, for a few WWW queries (and for both approaches):

Query	Mutual Reinforcement Approach		Stochastic Web Approach	
	no trimming	$(3, 3)$ -trimming	no trimming	$(3, 3)$ -trimming
java	18	17	342	41
movies	14	13	199	146
+censorship	20	13	241	19
+net				
genetic	38	35	109	65
abortion	23	22	163	35

Table 4.26: Number of Required Iterations Until Convergence

From the table we see that when using trimming, the number of iterations required in the Stochastic approach is about twice as much as what is required in the Mutual Reinforcement approach. The results for the query “movies” seem to contradict this, but recall that the Mutual Reinforcement approach failed to find a reasonable principal community there (see table 4.15).

Gram-Schmidt Method:

First Non-Principal Eigenvector, positive end.

url	header	cat	weight
http://www.ncbi.nlm.nih.gov/	The National Center for Biotechnology Information	(3)	0.473372
http://gdbwww.gdb.org/	The Genome Database	(3)	0.364426
http://www.nih.gov/	National Institute of Health (NIH)	(3)	0.255864
http://www-genome.wi.mit.edu/	Welcome To the Whitehead Institute Center for Genome Research	(3)	0.231098
http://www.ornl.gov/TechResources/ Human-Genome/home.html	Human Genome Project Information	(3)	0.157152
http://www.ebi.ac.uk/	New EBI Home Page	(3)	0.154447
http://cgsc.biology.yale.edu/top.html	E. coli Genetic Stock Center	(1)	0.151391

Second Non-Principal Eigenvector, negative end.

url	header	cat	weight
http://www.yahoo.com/	Yahoo!	(3)	0.484356
http://www.excite.com/	Excite	(3)	0.383734
http://www.lycos.com/	Lycos: Your Personal Internet Guide	(3)	0.326583
http://www.hotbot.com/	HotBot	(3)	0.283708
http://altavista.digital.com/	AltaVista: Main Page	(3)	0.185984
http://webcrawler.com/	WebCrawler	(3)	0.16824
http://www.infoseek.com/	GO Network-Start Here-	(3)	0.162115

Third Non-Principal Eigenvector, positive end.

url	header	cat	weight
http://www.netlink.de/gen/home.html	Genetic Engineering	(1)	0.482653
http://www.greenpeace.org/comms/cbio/ cbio/geneng.html		(3)	0.407808
http://www.bio-integrity.org/		(3)	0.393517
http://userwww.sfsu.edu/rone/ Genetic-Engineering.htm	Genetic Engineering and its Dangers	(1)	0.312838
http://www.geocities.com/Athens/1527/	Pure Food Campaign Homepage	(3)	0.237375
http://home1.swipnet.se/w-18472/ indexeng.htm	Genetically Engineered Foods - Safety Problems	(3)	0.224139
http://www.essential.org/crg/	Council For Responsible Genetics	(3)	0.216202

Table 4.25: WWW query “genetic”: Non-prin. authorities (Gram-Schmidt)

As with the previous multi-topic queries, we see that the Stochastic approach 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).

The following tables bring 2 deletion-induced communities and 3 non-principal Gram-Schmidt communities. The communities which focus on Genetic Engineering and the Genome come out exactly the same with both methods. The GS method produced an additional community of *search engines* (which didn't appear in the communities produced by deletion).

Deletion Method: second community, obtained by deleting the top 16 authorities of the Principal Community:

url	header	cat	weight
http://www.ncbi.nlm.nih.gov/	The National Center for Biotechnology Information	(3)	0.515507
http://gdbwww.gdb.org/	The Genome Database	(3)	0.395835
http://www.nih.gov/	National Institute of Health (NIH)	(3)	0.27598
http://www-genome.wi.mit.edu/	Welcome To the Whitehead Institute Center for Genome Research	(3)	0.254801
http://www.ornl.gov/TechResources/ Human-Genome/home.html	Human Genome Project Information	(3)	0.170042
http://www.ebi.ac.uk/	New EBI Home Page	(3)	0.168061
http://cgsc.biology.yale.edu/top.html	E. coli Genetic Stock Center	(1)	0.165061

Deletion Method: fourth community, obtained by deleting the top 16 authorities of the top three Communities:

url	header	cat	weight
http://www.netlink.de/gen/home.html	Genetic Engineering	(1)	0.483819
http://www.greenpeace.org/comms/cbio/ cbio/geneng.html		(3)	0.408652
http://www.bio-integrity.org/		(3)	0.394235
http://userwww.sfsu.edu/rone/ Genetic-Engineering.htm	Genetic Engineering and its Dangers	(1)	0.313719
http://www.geocities.com/Athens/1527/	Pure Food Campaign Homepage	(3)	0.237795
http://home1.swipnet.se/w-18472/ indexeng.htm	Genetically Engineered Foods - Safety Problems	(3)	0.224437
http://www.essential.org/crg/	Council For Responsible Genetics	(3)	0.216496

Table 4.24: WWW query “genetic”: Non-principal authorities (deletion)

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. The various communities which were derived by the different algorithmic approaches reflect all of these topics.

Size of root size = 120, Size of collection = 2952

Principal Community, Mutual Reinforcement Approach:

url	header	cat	weight
http://www.aic.nrl.navy.mil/galist/	The Genetic Algorithms Archive	(3)	0.27848
http://alife.santafe.edu/	Artificial Life Online	(3)	0.276159
http://www.yahoo.com/	Yahoo!	(3)	0.273599
http://www.geneticprogramming.com/	The Genetic Programming Notebook	(1)	0.25588
http://gal4.ge.uiuc.edu/illigal.home.html	illiGAL Home Page	(3)	0.235717
http://www.cs.gmu.edu/research/gag/	The Genetic Algorithms Group...	(3)	0.201237
http://www.scs.carleton.ca/csgs/ArtificialLifeResources/gaal.html	Genetic Algorithms and Resources	(1)	0.181315
http://lancet.mit.edu/ga/	GALib: Matthew's Genetic Algorithms Library	(3)	0.181157
http://www.santafe.edu/	Welcome to the Santa Fe Institute	(3)	0.180409
http://www.cis.ohio-state.edu/hypertext/faq/usenet/ai-faq/genetic/top.html	Genetic	(3)	0.175083

Principal Community, Stochastic Web Approach:

url	header	cat	weight
http://www.ncbi.nlm.nih.gov/	The National Center for Biotechnology Information	(3)	0.250012
http://www.yahoo.com/	Yahoo!	(3)	0.227782
http://www.aic.nrl.navy.mil/galist/	The Genetic Algorithms Archive	(3)	0.223191
http://www.nih.gov/	National Institute of Health (NIH)	(3)	0.194688
http://gdbwww.gdb.org/	The Genome Database	(3)	0.177001
http://alife.santafe.edu/	Artificial Life Online	(3)	0.172383
http://www.genengnews.com/	Genetic Engineering News (GEN)	(1)	0.141617
http://gal4.ge.uiuc.edu/illigal.home.html	illiGAL Home Page	(3)	0.13259
http://www.santafe.edu/	Welcome to the Santa Fe Institute	(3)	0.12818
http://lancet.mit.edu/ga/	GALib: Matthew's Genetic...	(3)	0.128173

Table 4.23: Authorities for WWW query "genetic"

When attempting to find non-principal communities with the Mutual Reinforcement approach, the pro-choice community is easily found both by the Gram-Schmidt method, and by the deletion method. We bring the top 8 pro-choice authorities, as found by the two methods.

Gram-Schmidt Method:

First Non-Principal Eigenvector, positive end.

url	header	cat	weight
http://www.naral.org/	NARAL Choice for America	(3)	0.375665
http://www.feminist.org/	Feminist Majority Foundation	(3)	0.316253
http://www.now.org/	National Organization for Women	(3)	0.241105
http://www.plannedparenthood.org/	Planned Parenthood Federation	(3)	0.240697
http://www.gynpages.com/	Abortion Clinics Online	(3)	0.22547
http://www.bodypolitic.org/	Body Politic Net News Home	(3)	0.20982
http://www.ms4c.org/	Medical Students for Choice	(3)	0.204592
http://www.cais.com/agm/main/index.html	The Abortion Rights Activist	(1)	0.196496

Deletion Method: second community, obtained by deleting the top 25 authorities of the Principal Community:

url	header	cat	weight
http://www.feminist.org/	Feminist Majority Foundation	(3)	0.354532
http://www.plannedparenthood.org/	Planned Parenthood Federation	(3)	0.284334
http://www.now.org/	National Organization for Women	(3)	0.271814
http://www.gynpages.com/	Abortion Clinics Online	(3)	0.265646
http://www.cais.com/agm/main/index.html	The Abortion Rights Activist	(1)	0.233572
http://www.bodypolitic.org/	Body Politic Net News Home	(3)	0.23076
http://www.ms4c.org/	Medical Students for Choice	(3)	0.23061
http://www.aclu.org/	ACLU: American Civil Liberties Union	(3)	0.209104

Table 4.22: Non Principal Authorities for WWW query “Abortion”

Query: abortion

First, 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	header	cat	weight
http://www.nrlc.org/	National Right To Life	(3)	0.420832
http://www.prolife.org/ultimate/	The Ultimate Pro-Life Resource List	(3)	0.316564
http://www.all.org/	What's new at American Life League	(3)	0.251506
http://www.hli.org/	Human Life International	(3)	0.212931
http://www.prolife.org/cpcs-online/	Crisis Pregnancy Centers Online	(3)	0.187707
http://www.ohiolife.org/	Ohio Right to Life	(3)	0.182076
http://www.rtl.org/	Abortion, adoption and assisted-suicide Information at Right to Life...	(1)	0.17943
http://www.bethany.org/	Bethany Christian Services	(3)	0.161359
http://www.ldi.org/	abortion malpractice litigation	(1)	0.140076
http://www.serve.com/fem4life/	Feminists for Life of America	(3)	0.122106

Principal Community, Stochastic Web Approach:

url	header	cat	weight
http://www.nrlc.org/	National Right To Life	(3)	0.344029
http://www.prolife.org/ultimate/	The Ultimate Pro-Life Resource List	(3)	0.284714
http://www.naral.org/	NARAL Choice for America	(3)	0.240227
http://www.feminist.org/	Feminist Majority Foundation	(3)	0.186843
http://www.now.org/	National Organization for Women	(3)	0.177946
http://www.cais.com/agm/main/ index.html	The Abortion Rights Activist	(1)	0.166083
http://www.gynpages.com/	Abortion Clinics Online	(3)	0.163117
http://www.plannedparenthood.org/	Planned Parenthood Federation	(3)	0.157186
http://www.all.org/	What's new at American Life League	(3)	0.142357
http://www.hli.org/	Human Life International	(3)	0.142357

Table 4.21: Authorities for WWW query "Abortion"

All 10 top authorities found by the Mutual Reinforcement approach are pro-life resources, while the top 10 authorities found by the Stochastic approach 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).

This is another example of the *TKC effect*: More sites in the collection are related to Michael Jordan, but the sites related to the Kingdom of Jordan are strongly interconnected.

Erasing the top 15 sites in the principal community of the MR approach, and reapplying the MR approach, yields the following community, which is focussed on Michael Jordan:

Deletion Method: second community, obtained by deleting the top 15 authorities of the Principal Community:

url	header	cat	weight
http://www.nba.com/	NBA.com	(3)	0.26521
http://www.unc.edu/lbrooks2/jordan.html	Michael Jordan	(1)	0.165756
http://www.wvu.edu/n9345228/tom1.html	The Michael Jordan VIRTUAL GALLERY	(3)	0.370176
http://www.fidelweb.com/alex/jordanp.html	Michael Jordan Page	(1)	0.29716
http://jordan.sportsline.com/	The Official Michael Jordan Web Site CBS SportsLine	(3)	0.291606
http://www.linkexchange.com/	LinkExchange	(3)	0.240629
http://www.vasia.com/2345/	The Jordan Dome	(2)	0.172743
http://members.tripod.com/mj23/mj.html	A Michael Jordan Fan's Heartbreak	(1)	0.1117

Table 4.20: WWW query “Jordan”: Non-principal authorities (deletion)

When applying Gram-Schmidt steps, two communities which correspond to the positive and negative ends of the first non-principal eigenvector come out very similar to the top two deletion-induced non-principal communities: The top 8 sites of the positive end are the exact 8 Michael Jordan sites shown above (in the exact same order), while the top 8 sites of the negative end are the same top 8 sites as in the principal MR community, with a somewhat different internal order.

Query: Jordan

This query was run a couple of days after the funeral of King Hussein of Jordan, and about a month after the announcement of Michael Jordan's retirement from the NBA.

Size of root size = 155

Size of collection = 1414

Principal Community, Mutual Reinforcement Approach:

url	header	cat	weight
http://www.jrtv.com/	Jordan Radio and Television	(3)	0.226732
http://www.interconti-jordan.com/	Hotel Inter.Continental Jordan	(1)	0.221885
http://www.nic.gov.jo/	National Information System	(3)	0.212707
http://www.noor.gov.jo/	Welcome to H.M. Queen Noor of Jordan	(1)	0.210997
http://www.cbj.gov.jo/	Welcome to Central Bank of Jordan	(3)	0.206245
http://www.movenpick-petra.com/	Movenpick Resort(Petra-Jordan)	(1)	0.204446
http://www.ajib.com/	Arab Jordan Investment Bank	(1)	0.188235
http://www.musicboxjo.com/	Music Box Jordan	(1)	0.183785
http://www.jiec.com/	Jordan Industrial Estates Corporation	(1)	0.182464

Principal Community, Stochastic Web Approach:

url	header	cat	weight
http://www.nba.com/	NBA.com	(3)	0.26521
http://www.linkexchange.com/	LinkExchange	(3)	0.265209
http://www2.cajun.net/ rs1864/halj.htm	Hal Jordan Memorial	(1)	0.190621
http://www.unc.edu/ lbrooks2/jordan.html	Michael Jordan	(1)	0.165756
http://www.jordancontrols.com/	Jordan Controls Electronic Actuators	(1)	0.15747
http://www.dakkak.com/	Dakkak Tours and Travel, Amman-Jordan	(1)	0.149178
http://www.nic.gov.jo/	National Information System	(3)	0.149176
http://www.interconti-jordan.com/	Hotel Inter.Continental Jordan	(1)	0.140889
http://www.access2arabia.com/	In Memory of His Majesty...	(3)	0.140888

Table 4.19: Authorities for WWW query "Jordan"

Clearly, the Mutual Reinforcement approach compiled a principal community on the topic of the Kingdom of Jordan, while two of the top four authorities in the Stochastic approach are on the topic of Michael Jordan, and only lower-rated authorities are concerned with the Kingdom.

4.3.2 Multi-Topic Queries

In the following subsections we set forth to examine two issues pertaining to multi-topic collections:

1. The difference in the principal communities of authorities which the two approaches find on such collections. For each query, we bring the principal communities found by the Mutual Reinforcement approach and by the Stochastic approach. The results show that the Mutual Reinforcement approach tends to find a principal community which is centered around one of the topics of the collection, while the Stochastic approach finds a principal community which contains authorities from various topics.
2. The different techniques to derive non-principal communities of authorities (in the Mutual Reinforcement approach only). We bring the non-principal algebraic communities which arise from the non-principal eigenvectors of $W^T W$ (calculated by using Gram-Schmidt steps in the Power method), as well as the deletion-induced non-principal communities of authorities. The results show that the deletion method induces the same non-principal communities as those found in the non-principal eigenvectors.

In section 3.7 we have stated that our experiments indicate that the non-principal eigenvectors of the stochastic matrices (used in the Stochastic approach) do not identify meaningful communities of authorities (or hubs). The deletion method fails here as well.

Therefore, the authorities returned by the Stochastic approach contain none of those *go.msn.com* sites, and are much more relevant to the query:

url	header	cat	weight
http://us.imdb.com/	The Internet Movie Database	(3)	0.253333
http://www.mrshowbiz.com/	Mr Showbiz	(3)	0.22335
http://www.disney.com/	Disney.com–The Web Site for Families	(3)	0.22003
http://www.hollywood.com/	Hollywood Online:...all about movies	(3)	0.213355
http://www.imdb.com/	The Internet Movie Database	(3)	0.199987
http://www.paramount.com/	Welcome to Paramount Pictures	(3)	0.196682
http://www.mca.com/	Universal Studios	(3)	0.180021
http://www.discovery.com/	Discovery Online	(3)	0.155024
http://www.film.com/	Welcome to Film.com	(3)	0.153347
http://www.mgmua.com/	mgm online	(3)	0.130012

Table 4.17: Stochastic authorities for WWW query “movies”

Similar results are obtained as a non-principal deletion-induced community by the Mutual Reinforcement approach, when deleting the top 30 sites of the principal community of that approach:

Deletion Method: second community, obtained by deleting the top 30 authorities of the Principal Community:

url	header	cat	weight
http://www.paramount.com/	Welcome to Paramount Pictures	(3)	0.339859
http://www.mca.com/	Universal Studios	(3)	0.314337
http://www.disney.com/	Disney.com–The Web Site for Families	(3)	0.275427
http://www.hollywood.com/	Hollywood Online:...all about movies	(3)	0.266036
http://www.mrshowbiz.com/	Mr Showbiz	(3)	0.2535
http://www.mgmua.com/	mgm online	(3)	0.24504
http://www.miramax.com/	Welcome to the Miramax Cafe	(3)	0.228061
http://us.imdb.com/	The Internet Movie Database	(3)	0.205971
http://www.film.com/	Welcome to Film.com	(3)	0.201761
http://www.warnerbros.com/	Warner Bros. Home Page	(3)	0.186598

Table 4.18: WWW query “movies”: Non-principal authorities (deletion)

Query: movies

This query demonstrates the *TKC effect* in a most striking fashion on the WWW.

Since we ignore, as a rule, links which are identified as intra-domain links, we were quite surprised to see the following principal community returned by the Mutual Reinforcement approach:

Size of root size = 175

Size of collection = 4539

url	header	cat	weight
http://go.msn.com/npl/msnt.asp	MSN.COM	(3)	0.167332
http://go.msn.com/bql/whitepages.asp	White Pages - msn.com	(3)	0.167202
http://go.msn.com/bsl/webevents.asp	Web Events	(3)	0.167202
http://go.msn.com/bql/maps.asp	Microsoft Expedia Maps-Home	(3)	0.167202
http://go.msn.com/bql/scoreboards.asp	MSN Sports scores	(3)	0.167202

Table 4.15: Mutual Reinforcement Authorities for WWW query “movies”

The top 30 authorities of the Mutual Reinforcement approach were all *go.msn.com* sites. All but the first received the exact same weight, 0.167202. Wondering how all these sites scored well even though we do not allow same-domain links in our collection, we looked at the list of hubs:

url	header	cat	weight
http://denver.sidewalk.com/movies	movies: denver.sidewalk	(1)	0.169197
http://boston.sidewalk.com/movies	movies:boston.sidewalk	(1)	0.169061
http://twincities.sidewalk.com/movies	movies: twincities.sidewalk	(1)	0.1688
http://newyork.sidewalk.com/movies	movies: newyork.sidewalk	(1)	0.168537

Table 4.16: 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.

Query: +censorship +net

For this query, both approaches produced the same top six sites (although in a different order).

Size of root size = 150

Size of collection = 562

Principal Community, Mutual Reinforcement Approach:

url	header	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, Stochastic Web Approach:

url	header	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 4.14: Authorities for WWW query “+censorship +net”

are different, the 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 the Stochastic approach, 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.

Query: Java

We bring the top ten authorities returned by both approaches.

Size of root size = 160

Size of collection = 2810

Principal Community, Mutual Reinforcement Approach:

url	header	cat	weight
http://www.jars.com/	EarthWeb's JARS.COM Java Review Service	(3)	0.334102
http://www.gamelan.com/	Gamelan - The Official Java Directory	(3)	0.303624
http://www.javascripts.com/	Javascripts.com - Welcome	(3)	0.255254
http://www.datamation.com/	EarthWeb's Datamation.com	(3)	0.251379
http://www.roadcoders.com/	Handheld Software Development@ RoadCoders	(3)	0.250816
http://www.earthweb.com/	EarthWeb	(3)	0.249373
http://www.earthwebdirect.com/	Welcome to Earthweb Direct	(3)	0.247467
http://www.itknowledge.com/	ITKnowledge	(3)	0.246874
http://www.intranetjournal.com/	intranetjournal.com	(3)	0.24518
http://www.javagoodies.com/	Java Goodies JavaScript Repository	(3)	0.238793

Principal Community, Stochastic Web Approach:

url	header	cat	weight
http://java.sun.com/	Java(tm) Technology Home Page	(3)	0.365264
http://www.gamelan.com/	Gamelan - The Official Java Directory	(3)	0.36369
http://www.jars.com/	EarthWeb's JARS.COM Java Review Service	(3)	0.303862
http://www.javaworld.com/	IDG's magazine for the Java community	(3)	0.217269
http://www.yahoo.com/	Yahoo!	(3)	0.21412
http://www.javasoft.com/	Java(tm) Technology Home Page	(3)	0.203099
http://www.sun.com/	Sun Microsystems	(3)	0.187355
http://www.javascripts.com/	Javascripts.com - Welcome	(3)	0.138548
http://www.htmlgoodies.com/	htmlgoodies.com - Home	(3)	0.130676
http://javaboutique.internet.com/	The Ultimate Java Applet Resource	(1)	0.118081

Table 4.13: Authorities for WWW query "Java"

This is our first WWW example of the *TKC effect*. 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

4.3 The WWW

We tested the different approaches on wide-topic WWW queries. We obtained a collection of sites per each query, and then ran a few tests on each collection.

Since the WWW is extremely dynamic, with a rapidly changing topology, it seems relevant to state that all collections were assembled during February, 1999. The root sets were compiled using AltaVista ([8]).

When expanding the root set to the entire collection, we *filtered* the links pointing to and from web-sites. Following [19], 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 (and averaging just about twice the number of nodes).

The results are displayed in tables containing four columns:

1. The url.
2. The header of the url.
3. The *category* of the url: (1) for a member of the root set, (2) for a site pointing into the root set, and (3) for a site pointed at by a member of the root set.
4. The value of the coordinate of this url in the eigenvector.

4.3.1 Single Topic Queries

We ran the following three single-topic queries:

- Java
- +censorship +net
- movies

We list the top authorities in the principal communities returned by both the Mutual Reinforcement approach and the Stochastic approach.

With the deletion method, we obtain similar results. Recall that the principal community was authoritative on Error-Correcting Codes (six of the top eight authorities in the principal community pertained to that topic). Here is the first non-principal community:

Deletion Method: second community, obtained by deleting the top 8 authorities of the Principal Community:

Topic	Previous rank	Current weight	In-degree rank
Data Compression/Source Coding	1	0.396911	2
Data Compression/Source Coding	4	0.309352	4
Data Compression/Source Coding	7	0.252421	13
Data Compression/Source Coding	9	0.22343	17(tie)
Data Compression/Source Coding	8	0.212298	14(tie)
Data Compression/Source Coding	10	0.189835	20

Table 4.12: IEEE-IT multi-topic non-prin. communities (deletion)

The top five authorities in the *fifth* community obtained by the deletion method (after deleting 8 authorities in each of the previous 4 communities) are exactly the same as those found in the positive side of the fifth non-principal eigenvector.

Gram-Schmidt Method:

First Non-Principal Eigenvector, positive end.

Topic	Previous rank	Current weight	In-degree rank
Data Compression/Source Coding	1	0.302627	2
Data Compression/Source Coding	4	0.234548	4
Data Compression/Source Coding	7	0.192748	13
Data Compression/Source Coding	2	0.179578	3
Data Compression/Source Coding	8	0.173155	14(tie)
Data Compression/Source Coding	9	0.171575	17(tie)

Gram-Schmidt Method:

First Non-Principal Eigenvector, negative end.

Topic	Previous rank	Current weight	In-degree rank
Error-Correcting Codes	1	-0.268493	1
Error-Correcting Codes	2	-0.133	14(tie)
Error-Correcting Codes	3	-0.117569	5(tie)
Error-Correcting Codes	4	-0.116671	17(tie)
Error-Correcting Codes	5	-0.116246	41
Error-Correcting Codes	6	-0.102028	68

Gram-Schmidt Method:

Fifth Non-Principal Eigenvector, positive end.

Topic	Previous rank	Current weight	In-degree rank
Shift Register Sequences	1	0.446948	5(tie)
Did not appear previously	-	0.302131	77(tie)
Did not appear previously	-	0.288628	77(tie)
Shift Register Sequences	4	0.275198	107(tie)
Shift Register Sequences	5	0.274505	107(tie)

Table 4.11: IEEE-IT multi-topic non-prin. communities (Gram-Schmidt)

4.2.2 Multi-Topic Queries

The query we present in this section combined all the keywords from the three single-topic queries shown previously. The corresponding collection consisted of 2099 documents. By the size of the three separate collections, we deduce that the combined collection should consist of two large topics of approximately equal size, and a third, minor topic.

In the following tables, we do not repeat the details of the papers. Instead, we refer to the occurrence of the paper in the separate, single topic queries, and its ranking (by the Mutual Reinforcement approach). We also mention the rank of the paper when sorting all of the collection's papers by their *in-degree*.

Topic	Previous rank	Current weight	In-degree rank
Error-Correcting Codes	1	0.427339	1
Error-Correcting Codes	3	0.282731	5
Error-Correcting Codes	4	0.231693	17
Data Compression/Source Coding	3	0.183893	7
Error-Correcting Codes	2	0.181836	14
Data Compression/Source Coding	2	0.169305	3
Error-Correcting Codes	7	0.157656	24
Error-Correcting Codes	13	0.157616	44

Table 4.10: IEEE-IT multi-topic Mutual Reinforcement authorities

The top authorities in the combined collection consist mainly of the top Error Correcting Codes authorities (as found by the Mutual Reinforcement approach). Proceeding to the non-principal communities, by using Gram-Schmidt steps, we obtain the results shown in table 4.11. We see that the first non-principal eigenvector distinguishes between the two major topics in the collection. The third, minor topic appears only in the fifth non-principal eigenvector.

Article Name	Author(s)	weight	Award
Universal noiseless coding	Lee D. Davisson	0.26185	
Quantizing for minimum distortion	Joel Max	0.239548	
Compression of individual sequences via variable-rate coding	Abraham Lempel Jacob Ziv	0.23506	
Noiseless coding of correlated information sources	Jack K. Wolf David S. Slepian	0.193377	PA
Asymptotically optimal block quantization	Allen Gersho	0.17138	
Universal codeword sets and representations of the integers	Peter Elias	0.159341	
Least squares quantization in PCM	Stuart P. Lloyd	0.155488	
Asymptotically efficient quantizing	Herbert Gish John N. Pierce	0.148111	
Computation of channel capacity and rate-distortion functions	Richard E. Blahut	0.145611	PA
Universal coding, information, prediction, and estimation	Jorma J. Rissanen	0.142992	PA
Source coding with side information and a converse for degraded broadcast channels	Rudolf F. Ahlswede János Kórnér	0.139838	
Fixed rate universal block source coding with a fidelity criterion	Lee D. Davisson Robert M. Gray David L. Neuhoff	0.136478	
The performance of universal encoding	Raphail E. Krichevsky Victor K. Trofimov	0.135542	
Variations on a theme by Huffman	Robert G. Gallager	0.12942	
Source coding theorems without the ergodic assumption	Lee D. Davisson Robert M. Gray	0.128805	PA

Table 4.9: Stochastic Authorities for “Source coding”

Article Name	Author(s)	weight	Award
Coset codes-II: Binary lattices and related codes	G. David Forney Jr.	0.2850	
Channel coding with multilevel/phase signals	Gottfried Ungerboeck	0.2133	PA
Minimum-distance bounds for binary linear codes	Hermann Josef Helgert Russell D. Stinaff	0.1657	
Efficient maximum likelihood decoding of linear block codes using a trellis	Jack K. Wolf	0.1616	
Generalized Hamming weights for linear codes (Corresp.)	Victor K. W. Wei	0.1490	
A new multilevel coding method using error-correcting codes	Shuji Hirakawa Hideki Imai	0.1438	GJA
Shift-register synthesis and BCH decoding	James L. Massey	0.1327	GJA
Coset codes-I: Introduction and geometrical classification	G. David Forney Jr.	0.1323	
New binary codes	Chin-Long Chen Sudhakar M. Reddy Neil J. A. Sloane	0.1267	
Convolutional codes I: Algebraic structure	G. David Forney Jr.	0.1209	GJA
Generalized minimum distance decoding	G. David Forney Jr.	0.1207	
Error bounds for convolutional codes and an asymptotically optimum decoding algorithm	Andrew J. Viterbi	0.1137	PA

Table 4.8: Stochastic Authorities for “Error correcting codes”

Article Name	Author(s)	weight	Award
Coset codes-II: Binary lattices and related codes	G. David Forney Jr.	0.5144	
Efficient maximum likelihood decoding of linear block codes using a trellis	Jack K. Wolf	0.2612	
Channel coding with multilevel/phase signals	Gottfried Ungerboeck	0.2431	PA
Coset codes-I: Introduction and geometrical classification	G. David Forney Jr.	0.2262	
Dimension/length profiles and trellis complexity of linear block codes	G. David Forney Jr.	0.2191	
Minimal trellises for block codes (Corresp.)	Douglas J. Muder	0.1934	
A new multilevel coding method using error-correcting codes	Shuji Hirakawa Hideki Imai	0.1934	GJA
The dynamics of group codes: State spaces, trellis diagrams, and canonical encoders	G. David Forney Jr. Mitchell D. Trott	0.1927	
On the optimum bit orders with respect to the state complexity of trellis diagrams for binary linear codes (Corresp.)	Toru Fujiwara Tadao Kasami Shu Lin Toyoo Takata	0.1815	
Generalized Hamming weights for linear codes (Corresp.)	Victor K. W. Wei	0.1686	
Maximum-likelihood soft decision decoding of BCH codes	Yair Be'ery Alexander Vardy	0.1620	
On complexity of trellis structure of linear block codes (Corresp.)	Toru Fujiwara Tadao Kasami Shu Lin Toyoo Takata	0.1532	

Table 4.7: Mutual Reinforcement Authorities for “Error correcting codes”

Principal Community, Mutual Reinforcement Approach:

Article Name	Author(s)	weight	Award
Shift-register synthesis and BCH decoding	James L. Massey	0.8520	GJA
A generalization of the Berlekamp-Massey algorithm for multisequence shift-register synthesis with applications to decoding cyclic codes	Kenneth K. Tzeng Gui-Liang Feng	0.2281	
On the minimum distance of cyclic codes	Richard M. Wilson Jacobus H. van Lint	0.1733	
Fast decoding of codes from algebraic plane curves	Tom Høholdt Helge Elbrønd Jensen Knud J. Larsen Jørn Justesen	0.1619	

Principal Community, Stochastic Web Approach:

Article Name	Author(s)	weight	Award
Shift-register synthesis and BCH decoding	James L. Massey	0.7833	GJA
An analysis of the structure and complexity of nonlinear binary sequence generators	Edwin L. Key	0.2806	
A generalization of the Berlekamp-Massey algorithm for multisequence shift-register synthesis with applications to decoding cyclic codes	Kenneth K. Tzeng Gui-Liang Feng	0.1598	
Sieves for low autocorrelation binary sequences	Marcel J. E. Golay	0.1443	

Table 4.6: Authorities for “Shift register sequences”

Query: Error correcting/detecting codes

Here, the collection contained 1117 documents. In tables 4.7 and 4.8 we list the top 12 authorities found by the two approaches.

Query: Data compression/Source coding

Papers on these topics have received many Paper Awards throughout the years. In table 4.9 we list the top 15 authorities for the Stochastic approach, of which 4 have received the Paper Award (The collection consisted of 1050 papers).

The community produced by the Mutual Reinforcement approach is similar: Of the top 15 papers, the two approaches intersect in 12, with 3 award-winning papers appearing in the principal MR community.

4.2 IEEE-IT Web

In the summer of 1998, the IEEE Information Theory Society published the entire IEEE-IT transactions on a set of 11 compact disks. This collection contained 6603 publications, dating from February, 1953 until November of 1997. Each paper in the collection was hyperlinked to all of its references and citations. In addition, each paper named a list of *keywords* - the topics to which that paper was related. The overall number of keywords mentioned in the collection was 1558.

Using the keywords and the hyperlinking of papers to their references and citations, we built collections pertaining to certain queries in the following manner: By stating a few keywords in our queries, we gathered a root set containing all papers whose keyword-list intersected with the query. We then added all papers referenced from within the root set, and all papers which cite a member of the root set.

We bring the outcome of running three single-topic queries and one multiple-topic query on this collection (The multiple topic query uses the union of all keywords used in the three single-topic queries). Our criterion for success is a small list of award-winning papers - we will count the number of award-winners in our communities. There are two kinds of awards:

1. *Information Theory Paper Award (PA)* - 31 of the IT Transactions' papers have received this award up to 1996.
2. *Golden Jubilee Award (GJA)* - 14 papers were given this award in August, 1998.

4.2.1 Single Topic Queries

For all three single-topic queries, both approaches rank the papers very similarly to a simple in-degree ranking (ranking the papers according to the number of other papers which cite them). This shouldn't come as a surprise, as the *citation count* of a scientific paper is a well accepted indicator of impact in bibliometric studies of scientific journals.

Query: Shift-register sequences

When given this query, both approaches rank a Golden Jubilee recipient high above all other papers in the field. In table 4.6 We list the top 4 authorities ranked by both approaches (on a collection of 142 papers).

distribution by examining a two-state $(\{t_1, t_2\})$ chain with the following transition matrix:

$$\begin{pmatrix} \frac{280}{286} & \frac{6}{286} \\ \frac{3}{336} & \frac{333}{336} \end{pmatrix}$$

The steady-state of the 18-state Markov chain assigns each t_2 -authority a probability of $\frac{168}{2874}$, and each t_1 -authority a probability of $\frac{143}{2874}$. Therefore the Stochastic approach finds t_2 as the principal topic of this collection.

Let us carry this example further, by adding a few more hubs, all of which point only to a_1^1 and a_1^2 . Adding k such hubs will transform the top 2×2 non-zero portion of A to be:

$$\begin{pmatrix} 286 + k & 274 + k \\ 274 + k & 286 + k \end{pmatrix}$$

In appendix A we prove:

Proposition 2 *For any $k > 0$, the entries in the principal eigenvector of A which correspond to t_1 -authorities will remain larger than those which correspond to t_2 authorities. Within the t_1 authorities, the entries of a_1^1 and a_1^2 will be larger than those of the other t_1 authorities.*

In the Stochastic approach, we now need to examine the three-state chain $(\{ \{a_1^1, a_1^2\}, \{a_1^3, a_1^4, a_1^5, a_1^6\}, \{a_2^j, 1 \leq j \leq 12\} \})$ with the following transition matrix:

$$\begin{pmatrix} \frac{k+6}{286+k} + \frac{274}{286+k} \cdot \frac{1}{3} & \frac{274}{286+k} \cdot \frac{2}{3} & \frac{6}{286+k} \\ \frac{274}{286} \cdot \frac{1}{3} & \frac{274}{286} \cdot \frac{2}{3} + \frac{6}{286} & \frac{6}{286} \\ \frac{1}{336} & \frac{2}{336} & \frac{333}{336} \end{pmatrix}$$

For values of $k \geq 50$, The Stochastic approach will rank a_1^1 and a_1^2 above the t_2 -authorities (which will remain above the rest of the t_1 -authorities). Thus, while the mutual reinforcement approach maintains that t_1 remains the principal authority, the Stochastic approach *blends* into its principal community authorities from both topics. As stated previously, this example may seem a bit contrived at this point. However, the bias of the Mutual Reinforcement approach towards tightly-knit communities (as compared with the ability of the Stochastic approach to blend authorities from different topics into its principal community) will be demonstrated on WWW queries.

- $H_1 = \{h_1^1, h_1^2, \dots, h_1^{274}\}$ is the set of 274 t_1 -hubs.
- $A_2 = \{a_2^1, a_2^2, \dots, a_2^{12}\}$ is the set of 12 t_2 -authorities.
- $H_2 = \{h_2^1, h_2^2, \dots, h_2^{792}\}$ is the set of $\binom{12}{5} = 792$ t_2 -hubs.
- Let there be $6 \cdot 12 = 72$ *noisy sites* n_i^j , $1 \leq i \leq 6, 1 \leq j \leq 12$.

The links in this topology are as follows:

- Each $h \in H_1$ points to all authorities in A_1 .
- Each $h \in H_2$ points to a *distinct* set of 5 authorities in A_2 .
- Each site n_i^j points to a_1^i and to a_2^j .

Topic t_1 is termed a *tightly-knit community*. It is a small, well connected topic. Topic t_2 , on the other hand, is much larger (of wider interest in the web context), with hubs that each covers only a portion of the topic's authorities.

Here is the non-zero portion of the authority matrix $A = W^T \cdot W$:

$$\begin{pmatrix} 286 & 274 & \cdots & 274 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 274 & 286 & \cdots & 274 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 274 & 274 & \cdots & 286 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & 336 & 120 & 120 & 120 & \cdots & 120 & 120 \\ 1 & 1 & \cdots & 1 & 120 & 336 & 120 & 120 & \cdots & 120 & 120 \\ 1 & 1 & \cdots & 1 & 120 & 120 & 336 & 120 & \cdots & 120 & 120 \\ 1 & 1 & \cdots & 1 & 120 & 120 & 120 & 336 & \cdots & 120 & 120 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & 120 & 120 & 120 & 120 & \cdots & 336 & 120 \\ 1 & 1 & \cdots & 1 & 120 & 120 & 120 & 120 & \cdots & 120 & 336 \end{pmatrix}$$

The normalized principal eigenvector has 0.270389 in the coordinates which correspond to t_1 -authorities, and 0.216283 in the coordinates which correspond to t_2 -authorities. We conclude that the Mutual Reinforcement approach identifies t_1 as the principal topic of this collection.

Let us now turn our attention to the Stochastic approach. Because of the complete symmetry inside each topic, we can compute the stationary

have already studied the spectrum of H in such designs, we deduce that the multiset of non-zero eigenvalues of A is: $\{27, 6, 6, 6, 6, 6, 6\}$. The principal eigenvector remains non-informative: It is the all-ones vector. However, there are six orthogonal eigenvectors which correspond to the eigenvalue 6 which identify the seven communities of same-topic authorities:

$$\begin{array}{cccccccc}
(1, 1, 1, & -1, -1, -1, & 0, 0, 0, & 0, 0, 0, & 0, 0, 0, & 0, 0, 0, & 0, 0, 0) \\
(1, 1, 1, & 1, 1, 1, & -2, -2, -2, & 0, 0, 0, & 0, 0, 0, & 0, 0, 0, & 0, 0, 0) \\
& & & & & & \vdots \\
(1, 1, 1, & 1, 1, 1, & 1, 1, 1, & 1, 1, 1, & 1, 1, 1, & 1, 1, 1, & -6, -6, -6)
\end{array}$$

Although informative non-principal eigenvectors exist, we cannot ensure convergence of the Power method to this set of vectors: Different choices of the initial vector (on which the Power method is applied) will cause us to converge to different eigenvectors in this eigenspace. The eigenvectors to which we will converge could thus turn out to be non-informative. We are, however, certain to converge to an eigenvector of the following “triplets” form: For every topic t , the three t -authorities will have equal entries in the eigenvector.

The Stochastic approach (as was the case with the hubs) computes a uniform distribution over all $|T| \cdot A$ sites and is not helpful in identifying authorities. However, the graph based clustering techniques mentioned in section 3.3 are ideal for such a topology, identifying $|T|$ clusters of A sites each when using the threshold $\tau = 2$. This shows that the information contained in the link structure of such combinatorial designs can be successfully tapped, although the Mutual Reinforcement approach and the Stochastic approach fail in this respect.

The Tightly-Knit-Community (TKC) Effect

In this section we show a topology on which the Mutual Reinforcement approach and the Stochastic approach arrive at different results. This topology might seem somewhat artificial at first, but the effect it produces is found in many real-life scenarios on the WWW, as will be demonstrated later in our results.

Consider a two-topic topology, $T = \{t_1, t_2\}$, with the following $n = |\mathcal{C}| = 1156$ sites:

- $A_1 = \{a_1^1, a_1^2, \dots, a_1^6\}$ is the set of 6 t_1 -authorities.

- The entries on the main diagonal of the co-citation matrix ($i = j, k = l$) measure the *in-degree* of each authority.
- The entries $A_{t_i^k, t_i^l}, k \neq l$ correspond to distinct authorities of the same topic. The total number of hubs which cover topic i is $\frac{n_h \cdot c}{|T|}$ (properties 2 and 7). Each i -hub points to a i -authorities, thus co-pointing at $\binom{a}{2}$ pairs of i -authorities. Overall, there are $\binom{A}{2}$ pairs of i -authorities. Hence, by property 8, each such pair is co-pointed at by $\frac{n_h \cdot c}{|T|} \cdot \binom{a}{2}$ hubs.
- The entries $A_{t_i^k, t_i^l}, i \neq j$ correspond to authorities of different topics. The properties of the design are not strong enough to uniquely determine the value of such entries.

It is interesting to note that although no information can be deduced about the hubs by either approach, it is sometimes possible to deduce information pertaining to the authorities. This is demonstrated by the following example, in which $|T| = 7$, $A = 3$, $n_h = 7$, $c = 3$, $a = 3$, $\tau = 1$ and $\alpha = 3$. Here is the $n_h \times |T| \cdot A$ non-zero submatrix of W :

	t_1^1, t_1^2, t_1^3	t_2^1, t_2^2, t_2^3	t_3^1, t_3^2, t_3^3	t_4^1, t_4^2, t_4^3	t_5^1, t_5^2, t_5^3	t_6^1, t_6^2, t_6^3	t_7^1, t_7^2, t_7^3
h_1	1,1,1	1,1,1	1,1,1	0,0,0	0,0,0	0,0,0	0,0,0
h_2	0,0,0	0,0,0	1,1,1	1,1,1	1,1,1	0,0,0	0,0,0
h_3	0,0,0	1,1,1	0,0,0	1,1,1	0,0,0	1,1,1	0,0,0
h_4	1,1,1	0,0,0	0,0,0	0,0,0	1,1,1	1,1,1	0,0,0
h_5	0,0,0	0,0,0	1,1,1	0,0,0	0,0,0	1,1,1	1,1,1
h_6	1,1,1	0,0,0	0,0,0	1,1,1	0,0,0	0,0,0	1,1,1
h_7	0,0,0	1,1,1	0,0,0	0,0,0	1,1,1	0,0,0	1,1,1

Denote by M_3 the 3×3 matrix in which all entries equal 3 and by M_1 the 3×3 matrix in which all entries equal 1. The 21×21 non-zero portion of the matrix A has the following structure:

$$\begin{pmatrix} M_3 & M_1 & M_1 & \cdots & M_1 \\ M_1 & M_3 & M_1 & \cdots & M_1 \\ M_1 & M_1 & M_3 & \cdots & M_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_1 & M_1 & \cdots & M_1 & M_3 \end{pmatrix}$$

The matrices H and A have the same multiset of eigenvalues. In particular, the multiset of non-zero eigenvalues is the same for both matrices. As we

The additional $|T| \cdot A$ eigenvalues of H are zero, and the corresponding eigenvectors are those that span the solution space of $Hv = 0$. Again, this eigenspace is non-informative in terms of identifying hubs.

The use of *trimming* or of some δ -*disparity coefficient* cannot break the inherent symmetry between the hubs. Thus, it too cannot help us to identify meaningful communities of hubs. Neither can the Stochastic approach, where the $n_h \times n_h$ block of \tilde{H} which corresponds to the hubs has the following structure:

$$\tilde{H}_{h_i, h_j} = \begin{cases} d_a^{-1} & i = j \\ \frac{\tau \cdot \alpha}{c \cdot a} \cdot d_a^{-1} & i \neq j \end{cases}$$

Explanations:

- Upon leaving h_i towards *any* authority that h_i points at, the probability of immediately returning to h_i is d_a^{-1} (since all authorities have the same in-degree, d_a).
- Two different hubs h_i and h_j co-point at $\tau \cdot \alpha$ authorities (properties 3 and 5). The probability of leaving h_i to one of the authorities which h_j also points at is thus $\frac{\tau \cdot \alpha}{c \cdot a}$. Once we arrive at such an authority, the probability of returning to h_j is d_a^{-1} . Multiplying the two probabilities gives the result.

The stationary distribution of the random walk inside the hub-block which results from this structure is a uniform probability distribution over the hubs.

Finally, it can be shown that the graph theoretical clustering techniques mentioned in section 3.3 when given the matrix H , can do no better than either list all hubs as one community, or listing each of them as a separate community. This is outside the scope of this work.

We now turn our attention to the matrix $A = W^T \cdot W$. We first study the non-zero entries of the matrix: Denote by t_i^k the k 'th authority of topic i ($1 \leq k \leq A, 1 \leq i \leq |T|$).

$$A_{t_i^k, t_j^l} = \begin{cases} d_a & i = j, k = l \\ \frac{n_h \cdot c}{|T|} \cdot \frac{a(a-1)}{A(A-1)} & i = j, k \neq l \\ \text{some non-constant values} & i \neq j \end{cases}$$

Explanations:

Proposition 1 *There are infinitely many such topologies.*

The Proof, which appears in appendix A, relies on the properties of *block designs* ([22]).

With these criteria, the matrix W has non-zero entries only in the $n_h \times (|T| \cdot A)$ submatrix corresponding to the hubs and authorities. It follows that the matrix $H = WW^T$ has just an $n_h \times n_h$ non-zero block, which we will denote by \hat{H} , and which is of the following structure:

$$\hat{H}_{h_i, h_j} = \begin{cases} c \cdot a & i = j \\ \tau \cdot \alpha & i \neq j \end{cases}$$

From this complete symmetry, it is evident that \hat{H} is *non-informative* as far as our mission (identifying good hubs and authorities) is concerned. The following spectral analysis of \hat{H} explains why this is true:

- $\lambda_1 = c \cdot a + (n_h - 1) \cdot \tau \cdot \alpha$ is the principal eigenvalue. The corresponding principal eigenvector is the all-ones vector, which is non-informative as it groups together all hubs in a single community.
- $\lambda_2 = \dots = \lambda_{n_h} = c \cdot a - \tau \cdot \alpha$. The corresponding eigenspace is the set of all vectors which are orthogonal to the principal eigenvalue: These are all the vectors whose sum of coordinates equals zero. Following is an orthogonal basis for this eigenspace:

$$\begin{aligned} v_{\lambda_2} &= (1, -1, 0, 0, \dots, 0, 0) \\ v_{\lambda_3} &= (1, 1, -2, 0, \dots, 0, 0) \\ &\vdots \\ v_{\lambda_{n_h}} &= (1, 1, 1, 1, \dots, 1, 1 - n_h) \end{aligned}$$

For *any* subgroup of the hubs of size $1 \leq h < n_h$, we can build a $c \cdot a - \tau \cdot \alpha$ eigenvector which distinguishes those hubs from the rest: The entries corresponding to the subgroup of hubs will equal $n_h - h$, and the entries corresponding to all other hubs will equal $-h$. Since we can distinguish any arbitrary proper subset of the hubs from the rest with this eigenspace, it is non-informative.

The eigenvalues of \hat{H} are also eigenvalues of H . To transform an eigenvector of \hat{H} to an eigenvector of H , simply add zeros in the coordinates which correspond to zero-rows of H .

4.1.2 Multi-Topic Topologies

In this section we discuss two theoretical families of topologies, and examine the behavior of the algorithms when encountering such topologies. First, we construct a topology whose link structure is informative but in which both the Stochastic approach and the Mutual Reinforcement approach do not produce a meaningful principal community. Then, we show a topology on which the two approaches result in different principal communities.

Symmetric Combinatorial Designs

Our purpose here is to show the existence of an infinite number of artificial topologies, which exhibit an informative link structure, but on which both approaches fail to identify meaningful communities of authorities.

We begin by introducing the notion of *topic coverage*: A hub h is said to *cover* topic t whenever h points to a significant number of t -authorities. This rather intuitive definition will suffice for our needs.

Consider a *symmetric, orthogonal T – topic* collection (see section 2.3.1), consisting of $|T| \cdot A$ authorities (A distinct perfect authorities per topic) and n_h hubs, with the following (not necessarily independent) properties:

1. The set of authorities and the set of hubs are disjoint.
2. Every hub covers c topics.
3. Every two hubs co-cover τ topics.
4. For each $t \in T$, every t -hub points to exactly a of the A t -authorities.
5. For each $t \in T$, every two t -hubs co-point to α of the A t -authorities.
6. Every topic is covered by the same number of hubs.
7. Every authority is pointed at the same number of times.
8. For each $t \in T$, every two t -authorities are co-pointed at by the same number of hubs.

Observe that the number of outgoing links from hubs is $n_h \cdot c \cdot a$ (properties 2 and 4). The number of authorities is $|T| \cdot A$. Since all authorities are pointed at the same number of times (property 7), The constant in-degree of all authorities $d_a \triangleq \frac{n_h \cdot c \cdot a}{|T| \cdot A}$.

Mutual Reinforcement approach - Authorities

$w_a(s)$	$r_a(s)$	Eigenvector entry	Algorithm rank
0.998009	3	0.238103	1
0.951735	8	0.207831	2
0.965965	7	0.195137	3
0.806887	27	0.187788	4
0.934406	9	0.179845	5
0.999949	2	0.176058	6
\vdots	\vdots	\vdots	\vdots
0.690844	46	.0247338	82
0.0	83	0.000218625	83

Mutual Reinforcement approach - Hubs

$w_h(s) = (p_s, \epsilon_s)$	$r_h(s)$	Eigenvector entry	Algorithm rank
(1.148293, 0.235992)	1	0.229579	1
(1.048627, 0.238877)	2	0.201086	2
(0.979556, 0.243141)	3	0.175397	3
(0.990383, 0.258689)	4	0.173469	4
(0.907366, 0.250568)	8	0.167869	5

Table 4.5: Dense Truncated Exponential Model (δ -Enhanced Results)

We conclude that both approaches deal well with the sparse models. This should not surprise us, as even naive methods such as ranking sites by their *in-degree* would identify the authorities nicely in sparse topologies. As for the two dense models, the disparity coefficient made the difference between complete failure and success.

Real WWW collections, assembled as described in section 3.2, come out neither as sparse nor as dense as in our extreme artificial models: In [19], Kleinberg states that in the setting of the WWW, ranking sites by their in-degree alone is unsatisfactory. From this we can deduce that the link structure of the WWW is not as favorable as in our sparse models. On the other hand, as the Mutual Reinforcement approach showed good results in [19] and in [13], we deduce that neither is the link-structure as noisy as in our dense models.

Mutual Reinforcement approach - Hubs

$w_h(s) = (p_s, \epsilon_s)$	$r_h(s)$	Eigenvector entry	Algorithm rank
(1.150618,0.247648)	1	0.17418	1
(1.113207,0.272110)	2	0.165026	2
(1.084835,0.267780)	5	0.163551	3
(1.073933,0.250976)	4	0.159574	4
(0.991324,0.259206)	6	0.154326	5
(1.051839,0.213985)	3	0.152324	6
(0.985765,0.312470)	12	0.147724	7
(0.948298,0.236848)	8	0.145934	8

Stochastic approach - Hubs

$w_h(s) = (p_s, \epsilon_s)$	$r_h(s)$	Eigenvector entry	Algorithm rank
(0.839214,0.634177)	60	0.200503	1
(0.825913,0.607245)	58	0.183756	2
(0.673375,0.608624)	69	0.15037	3
(0.765664,0.514528)	54	0.13781	4
(0.914116,0.402289)	29	0.133628	5
(1.113207,0.272110)	2	0.133595	6
(1.150618,0.247648)	1	0.133594	7
(1.084835,0.267780)	5	0.12942	8

Table 4.4: Sparse Truncated Exponential Model - Hubs

Both approaches failed in the dense truncated exponential model, with no authorities in the top 20 sites of the Stochastic approach, and only 4 authorities in the top 20 sites of the Mutual Reinforcement approach. But when using a disparity coefficient of $\delta = 0.2$, the Mutual Reinforcement approach finds the following principal community (There were 82 authorities in the collection):

The results of both exponential models are of similar nature to that of the zero-one model. Here, the parameters used to build the collections were the following: $p = 0.05$ and $a = 0.5$ (the minimal authority weight).

In the sparse model, the drawings produced 73 authorities. Both approaches performed well, finding close communities of authorities, as is demonstrated in the following tables. Also shown is the steep drop in weights between authorities and non-authorities in the Mutual Reinforcement Approach (There was no such drop in the Stochastic approach).

Mutual Reinforcement approach - Authorities

$w_a(s)$	$r_a(s)$	Eigenvector entry	Algorithm rank
0.969074	4	0.189347	1
0.908423	10	0.170358	2
0.837320	17	0.162352	3
0.972789	3	0.160031	4
0.900097	12	0.160024	5
0.962265	5	0.157974	6
1.000000	1	0.156302	7
⋮	⋮	⋮	⋮
0.532335	68	0.0655381	72
0.527816	71	0.0574048	73
0.0	74	0.0103725	74

Stochastic approach - Authorities

$w_a(s)$	$r_a(s)$	Eigenvector entry	Algorithm rank
0.96907	4	0.162634	1
0.908423	10	0.137614	2
0.900097	12	0.137608	3
1.0	1	0.137604	4
0.782264	27	0.133448	5
0.866696	15	0.133444	6
0.972789	3	0.133444	7

Table 4.3: Sparse Truncated Exponential Model - Authorities

Here we bring the top ranking hubs, as found by the two approaches. The Mutual Reinforcement approach finds better hubs than the stochastic approach, although both top-11 hub communities intersect in 5 hubs.

authorities in its top 50 (the first non-authority was ranked 37). The hubs were also identified well in both approaches.

When looking at the authority weights given to the sites by the Mutual Reinforcement approach, one notices a steep drop between places 50 (the rank of the last authority) and 51 (the rank of the first non-authority):

$w_a(s)$	Eigenvector entry	Algorithm rank
1.0	0.166657	1
1.0	0.162755	2
\vdots	\vdots	\vdots
1.0	0.0902727	49
1.0	0.0847680	50
0.0	0.0358433	51

Table 4.1: Mutual Reinforcement approach - Sparse 0/1 Model

No such drop exists in the weights given by the Stochastic approach: The decline of the weights there is very smooth.

In the dense model, both approaches failed to identify the authorities: The Mutual Reinforcement approach didn't have any authority in its top 50 sites, while the Stochastic approach ranked 3 authorities in its top 50 (but none in the top 25 sites). But when applying the Mutual Reinforcement approach with a *disparity coefficient* $\delta = 0.2$, all 50 authorities and 50 hubs were identified. Again, there is a significant drop in the authority weights assigned by the algorithm, between the authority in place 50 and the non-authority in place 51:

$w_a(s)$	Eigenvector entry	Algorithm rank
1.0	0.262892	1
1.0	0.244987	2
\vdots	\vdots	\vdots
1.0	0.0355103	49
1.0	0.0306712	50
0.0	0.0014187	51

Table 4.2: δ -Enhanced Mutual Reinforcement approach - Dense 0/1 Model

$$\begin{aligned}
&= \frac{\mathcal{A}_h \cdot (1 - \epsilon_h)}{\mathcal{A}_h} = \\
&= 1 - \epsilon_h
\end{aligned}$$

These stages conclude the drawing of edges between hubs and other sites (edges of E_1 and E_2). The following two proportions are now calculated:

$$\begin{aligned}
p_1 &\triangleq \frac{1}{n_h \cdot n_a} \mid \{ (h \in H, s \in A) : \text{the edge } (h, s) \text{ was drawn} \} \mid \\
p_2 &\triangleq \frac{1}{n_h \cdot (n - n_a)} \mid \{ (h \in H, s \in \mathcal{C} \setminus A) : \text{the edge } (h, s) \text{ was drawn} \} \mid
\end{aligned}$$

In the sparse model, edges between non-hubs and other sites (edges in E_3 and E_4) are i.i.d. w.p. p_2 . Afterwards, all sites are evaluated as hubs, and ranked.

Dense Truncated Exponential Model

This model differs from the Sparse Truncated Exponential Model by the probabilities with which edges in E_3 and E_4 are drawn. These probabilities are defined in a similar manner to that of the Dense Zero-One Model:

1. Edges in E_3 are drawn w.p. $q_1 \triangleq p_2 + \frac{p_1 - p_2}{n - n_h} \cdot (n_a - n_h)$.
2. Edges in E_4 are drawn w.p. $q_2 \triangleq p_2 + \frac{(p_1 - p_2)}{n - n_h} \cdot n_a$.

Here again, the choice of the four probabilities p_1, p_2, q_1 and q_2 causes the expected in-degree and expected out-degree of all sites to be the same.

Algorithmic Results

We tested the algorithms on 1500-site collections, with $\beta(t) \triangleq 0.4$.

For both sparse and dense Zero-One models, we defined $n_a = n_h \triangleq 50$, $p_1 \triangleq 0.35$ and $p_2 \triangleq 0.01$. These parameters cause each hub to have an average of 17.5 outgoing links to authorities, and about 15 links to non-authorities.

In the sparse model, both approaches succeeded very well in identifying the authorities: The Mutual Reinforcement approach ranked the 50 authorities as the top 50 sites, while the Stochastic approach ranked 46 of the 50

recall but weak precision and very precise hubs which point at relatively few authorities.

Now that the hub scores have been set, let h be a site with $\tilde{w}_h(h) > 0$ and $w_h(h) = (p_h, \epsilon_h)$. For all authorities s , we draw the edge from h to s with probability $\mu \cdot \beta(t) \cdot p_h \cdot w_a(s)$ (we assume that $\beta(t)$ is sufficiently low so that $\mu \cdot \beta(t) \leq 1$). Note that although we do not require the resulting topology to be consistent, this choice of probabilities causes hubs to point at strong authorities with higher probabilities than at weak ones.

Let d denote the number of edges from h to authorities that were successfully drawn in this stage, and let \mathcal{A}_h be the recall of h (the total authority collected by h through those d links):

$$\mathcal{A}_h = \sum_{s: h \rightarrow s} w_a(s)$$

The expected value of \mathcal{A}_h is $p_h \cdot \beta(t) \cdot \text{Sum}(t)$, as required by the recall measure of h 's hub score:

$$\begin{aligned} E[\mathcal{A}_h] &= E\left[\sum_{s: h \rightarrow s} w_a(s)\right] \\ &= \sum_{s: w_a(s) > a} P(h \rightarrow s) \cdot w_a(s) \\ &= \sum_{s: w_a(s) > a} \mu \cdot \beta(t) \cdot p_h \cdot w_a^2(s) \\ &= \mu \cdot \beta(t) \cdot p_h \sum_{s: w_a(s) > a} w_a^2(s) = p_h \cdot \beta(t) \cdot \text{Sum}(t) \end{aligned}$$

We then check the hub's precision. If $\frac{\mathcal{A}_h}{d} > 1 - \epsilon_h$, we need to add links from h to non-authorities in order to weaken h 's precision. We thus add

$$m \triangleq \left\lfloor \frac{1}{2} + \frac{\mathcal{A}_h - d(1 - \epsilon_h)}{1 - \epsilon_h} \right\rfloor$$

edges from h to non-authorities, and the precision of h becomes approximately the required value of $1 - \epsilon_h$:

$$\begin{aligned} \frac{\mathcal{A}_h}{d + m} &\approx \frac{\mathcal{A}_h}{d + \frac{\mathcal{A}_h - d(1 - \epsilon_h)}{1 - \epsilon_h}} = \\ &= \frac{\mathcal{A}_h \cdot (1 - \epsilon_h)}{d \cdot (1 - \epsilon_h) + \mathcal{A}_h - d \cdot (1 - \epsilon_h)} = \end{aligned}$$

These calculations assume that self loops (a link from a site to itself) are allowed. In practice, self looping edges are not drawn, but this has a negligible effect on the above calculations.

Sparse Truncated Exponential Model

This model is defined by the following parameters:

- n - The number of sites.
- a - The minimal authority weight / hub grade (not less than 0.5).
- $p = p_a = p_h$ - The desired fraction of hubs and authorities in the collection.
- $\beta(t)$ for the single topic t at hand.

First, authority weights are drawn: Each site s is given an authority weight $w_a(s)$ which equals $e^{-\lambda \cdot t}$, where $\lambda \triangleq \frac{\ln a}{-p}$ and t is drawn uniformly in $[0, 1)$. Then, all weights smaller than a are changed to zero, and the remaining weights are scaled linearly so that the maximal weight equals 1. The expected proportion of authorities (sites which drew authority weights $\geq a$) is thus p :

$$P(e^{-\lambda \cdot t} \geq a) = P(-\lambda \cdot t \geq \ln a) = P(t \leq \frac{\ln a}{-\lambda}) = P(t \leq p) = p$$

With the authority weights drawn, we define the following quantity, which is later used for scaling edge probabilities:

$$\mu \triangleq \frac{\sum_{s:w_a(s)>a} w_a(s)}{\sum_{s:w_a(s)>a} w_a^2(s)} = \frac{\text{Sum}(t)}{\sum_{s:w_a(s)>a} w_a^2(s)} \geq 1$$

We now draw the hub grades. This is done in the same manner as the authority weights. Then, for all sites s satisfying $\tilde{w}_h(s) > a$, we set

$$w_h(s) = (p_s, \epsilon_s) = (\tilde{w}_h(s) + f \cdot r, 1 - (\tilde{w}_h(s) - f \cdot r))$$

where r is drawn uniformly in $(-(1 - \tilde{w}_h(s)), 1 - \tilde{w}_h(s))$ and $f \triangleq 0.4$ is a constant factor. The factor $f \cdot r$ has been added in order to separate the two quality measures of the hub score from each other, without changing the hub grade of s . Thus, our collection will contain both hubs with very good

These probabilities cause the expected in-degree (out-degree) of all sites to be equal, regardless of whether or not the site is an authority (a hub):

$$\begin{aligned}
E_{in-degree}(s \in A) &= E[\text{no. of edges from hubs to } s] + \\
&\quad E[\text{no. of edges from non-hubs to } s] = \\
&= n_h \cdot p_1 + (n - n_h) \cdot q_1 = \\
&= n_h \cdot p_1 + (n - n_h) \cdot p_2 + (p_1 - p_2) \cdot (n_a - n_h) = \\
&= n_a \cdot p_1 + (n - n_a) \cdot p_2
\end{aligned}$$

$$\begin{aligned}
E_{in-degree}(s \in \mathcal{C} \setminus A) &= E[\text{no. of edges from hubs to } s] + \\
&\quad E[\text{no. of edges from non-hubs to } s] = \\
&= n_h \cdot p_2 + (n - n_h) \cdot q_2 = \\
&= n_h \cdot p_2 + (n - n_h) \cdot p_2 + (p_1 - p_2) \cdot (n_a) = \\
&= n_a \cdot p_1 + (n - n_a) \cdot p_2
\end{aligned}$$

$$\begin{aligned}
E_{out-degree}(s \in H) &= E[\text{no. of edges from } s \text{ to authorities}] + \\
&\quad E[\text{no. of edges from } s \text{ to non-authorities}] = \\
&= n_a \cdot p_1 + (n - n_a) \cdot p_2 = \\
&= n_a \cdot (p_1 - p_2) + n \cdot p_2
\end{aligned}$$

$$\begin{aligned}
E_{out-degree}(s \in \mathcal{C} \setminus H) &= E[\text{no. of edges from } s \text{ to authorities}] + \\
&\quad E[\text{no. of edges from } s \text{ to non-authorities}] = \\
&= n_a \cdot q_1 + (n - n_a) \cdot q_2 = \\
&= n_a \cdot p_2 + \frac{p_1 - p_2}{n - n_h} (n_a - n_h) n_a + \\
&\quad (n - n_a) \cdot p_2 + \frac{(p_1 - p_2)}{n - n_h} \cdot n_a \cdot (n - n_a) = \\
&= n \cdot p_2 + \frac{p_1 - p_2}{n - n_h} \cdot n_a \cdot (n_a - n_h + n - n_a) = \\
&= n_a \cdot (p_1 - p_2) + n \cdot p_2
\end{aligned}$$

Sparse Zero-One Model

This model is defined by the following parameters:

- n - The number of sites.
- n_a - The number of authorities.
- n_h - The number of hubs.
- $\beta(t)$ ¹ for the single topic t at hand.
- p_1, p_2 - two probabilities satisfying $p_1 > p_2$.

In this model, each edge $(x, y) \in E_1$ is independently drawn w.p. (*with probability*) p_1 while edges of E_2, E_3 and E_4 are i.i.d.² w.p. p_2 . After drawing all edges, each site $s \in \mathcal{C}$ is evaluated as a hub according to its set of outgoing links and given a hub-score (p_s, ϵ_s) . Then, $r_h(s)$ is calculated for each site.

Typically, $n \gg n_a, n_h$. This results in most edges being drawn w.p. p_2 (the smaller of the two probabilities), and hence the obtained topology is sparse. Other properties of this model are that the in-degree of authorities tends to be higher than that of non-authorities, and that the out-degree of hubs tends to be higher than that of non-hubs.

Dense Zero-One Model

This model differs from the Sparse Zero-One Model only by the probabilities with which each of the four sets of edges is drawn:

1. Edges in E_1 are drawn w.p. p_1 .
2. Edges in E_2 are drawn w.p. p_2 .
3. Edges in E_3 are drawn w.p. $q_1 \triangleq p_2 + \frac{p_1 - p_2}{n - n_h} \cdot (n_a - n_h)$.
4. Edges in E_4 are drawn w.p. $q_2 \triangleq p_2 + \frac{(p_1 - p_2)}{n - n_h} \cdot n_a$.

¹Defined in section 2.2.1

²identically, independently drawn

- $n \triangleq |\mathcal{C}|$
- $A \subset \mathcal{C}$ - The set of all authorities (sites with non-zero authority weights).
- $n_a \triangleq |A|$
- $H \subset \mathcal{C}$ - The set of “good” hubs.
- $n_h \triangleq |H|$

For each site $s \in \mathcal{C}$, five quantities are defined:

1. $w_a(s)$ - The (single topic) authority weight of s .
2. $w_h(s)$ - The hub score of s - the ordered pair (p_s, ϵ_s) .
3. $\tilde{w}_h(s)$ - The quantity $\frac{p_s + (1 - \epsilon_s)}{2}$ (termed the *hub-grade* of s).
4. $r_a(s)$ - The authority ranking of s . The place of s among all sites, when sorted by decreasing authority weights.
5. $r_h(s)$ - The hub ranking of h . The place of s among all sites, when sorted by decreasing hub grades.

We distinguish between four sets of potential directed edges (links) between pairs of sites:

1. $E_1 \triangleq \{(x, y) | x \in H, y \in A\}$.
These are the potential informative links - links which point from hubs to authorities.
2. $E_2 \triangleq \{(x, y) | x \in H, y \in \mathcal{C} \setminus A\}$.
These potential edges correspond to links from hubs to non-authorities.
3. $E_3 \triangleq \{(x, y) | x \in \mathcal{C} \setminus H, y \in A\}$.
These potential edges correspond to links from non-hubs to authorities.
4. $E_4 \triangleq \{(x, y) | x \in \mathcal{C} \setminus H, y \in \mathcal{C} \setminus A\}$.
These potential edges correspond to links from non-hubs to non-authorities.

We first describe in detail each of the four models. Afterwards, we bring the results of applying the algorithms to them.

Chapter 4

Results

The various algorithms were tested on three domains:

1. Simulated/Artificial web topologies.
2. The collection of all publications which appeared in the IEEE proceedings on Information Theory between 1953 and 1997.
3. The WWW.

On each domain, both single topic and multi-topic searches were carried out.

4.1 Simulated/Artificial Web Topologies

In these topologies we have complete knowledge on the identity of the authorities, and we can measure the effectiveness of algorithms which attempt to find them.

4.1.1 Single Topic Topologies

We tested the algorithms on simulated topologies according to four different probabilistic models for hyperlinked collections. The need for such models was noted in [12]. The models differ from each other in the way authority weights are given, in the way hubs are defined, and in the probabilities in which the edges are drawn.

Following are some notations, used in all four models:

- \mathcal{C} - The set of all sites.

algebraic *principal* communities of the matrices which result from deletion process.

Consider the effect which the deletion method has on the principal community of authorities: Its top authorities, which had high association scores with each other and with the lesser authorities of that community, are erased. Thus, the essence of that community is discarded, and the bonding of the lesser authorities to each other yields a much weaker reinforcing relationship. This allows other communities, which were less dominant relative to the (decimated) principal community, to emerge as the strong communities in the resulting matrix.

Note that applying the deletion method to nonnegative symmetric matrices (as are the association matrices in the Mutual Reinforcement approach) results in matrices which remain nonnegative and symmetric. Assuming that the resulting matrices are also irreducible, they meet the convergence requirements of the power method.

The deletion method trivially prevents the same sites from reappearing in multiple communities. In addition, it allows the number of iterations required per community to remain low, as for each community we only need to identify the large coordinates of the positive principal eigenvector of that stage's matrix.

other in a somewhat weaker fashion than do the weight sets which correspond to more dominant eigenvalues. The Stochastic approach, however, offers no intuitive explanation to its non-principal algebraic communities. There is no meaningful interpretation (in our context) to eigenvectors of stochastic matrices which correspond to eigenvalues smaller than 1. Experimental results also did not reveal any correlation between non-principal eigenvectors and authoritative sites. We therefore limit our discussion of using the meta-algorithm to discover non-principal algebraic communities to the Mutual Reinforcement approach.

Computational aspects of finding non-principal communities Using the power method, we avoided the need to compute with high precision the principal eigenvector of a matrix M . We were able to identify the coordinates of high absolute value in $v_{\lambda_1(M)}$ by performing only a small number of iterations. Kleinberg [19] suggests applying the power method with Gram-Schmidt steps, in order to compute the non-principal eigenvectors. This brings about two problems:

- In order to converge to a non-principal eigenvector $v_{\lambda_k(M)}$, the power method must use an initial vector \tilde{u}_0 which is orthogonal to the set of more dominant eigenvectors, $\{v_{\lambda_1(M)}, \dots, v_{\lambda_{k-1}(M)}\}$. In order to arrive at such an initial vector, Gram-Schmidt steps must be applied with accurate estimations of those more dominant eigenvectors. Hence, these eigenvectors must be calculated fairly accurately. This required accuracy demands more computational resources (more iterations) in each eigenvector approximation, than are needed for simply identifying the coordinates of high magnitude in each eigenvector.
- Experimental results ([19]) show that the same community of hubs and authorities tends to recur in the eigenvectors associated with the first few large eigenvalues.

We propose the following process for finding non-principal communities, called the *deletion method*: After finding each community, *erase* the rows and columns which correspond to the authorities(hubs) of that community from the matrix, and re-apply the power method - thus finding the principal community of the matrix which results from the erasure of previous communities. This changes the semantics of the algebraic non-principal communities: They are no longer derived from non-principal eigenvectors. Instead, the algebraic *deletion-induced* non-principal communities are the

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

- Both matrices are primitive, since the Markov chains which they represent are aperiodic: When visiting any non-isolated authority(hub), there is a positive probability to revisit it on the next entry to the authority(hub) side of the bipartite graph. Hence, non-isolated sites will be represented by states with self-loops, causing the chain to be aperiodic.
- The adjacency matrix of the support graph $G(\tilde{A})$ is symmetric, since $\tilde{a}_{i,j} > 0$ implies $\tilde{a}_{j,i} > 0$. Furthermore, $\tilde{a}_{i,j} > 0 \iff [W^T \cdot W]_{i,j} > 0$ (and the same is also true of \tilde{H} and $W \cdot W^T$).

Assuming that \tilde{H} and \tilde{A} are irreducible, both matrices comply with the convergence requirements of the power method (stated in section 3.4.3), and can be used (as association matrices) in the meta-algorithm.

The strong coordinates in the stationary distribution (see the *Ergodic Theorem*, page 22) of the authority chain should be authoritative pages - they correspond to the authority-side states of \tilde{G} most visited by the scout, and hence most likely to be part of dense bipartite subgraphs of \tilde{G} . The same argument holds for the most probable states of the hub chain.

Some stochastic information of random walks in the Web has already been used by the *Google* ([15]) team in Stanford ([2]). The *PageRank* component of the search engine examines a *single* random walk on the *entire* *WWW* (not on a subgraph induced by some specific base set). 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.

3.7 Finding Non-Principal Communities

Both the Mutual Reinforcement approach and the Stochastic approach offer intuitive arguments to explain why the principal algebraic community of authorities should resemble the actual principal community. For the Mutual Reinforcement approach, similar arguments can also be made on behalf of the non-principal algebraic communities: Each pair of (normalized) eigenvectors of A and H which correspond to the same eigenvalue, are simply sets of authority weights and hub weights which mutually reinforce each other to the extent of the magnitude of that eigenvalue. Hence, non-principal algebraic communities produced by the Mutual Reinforcement approach are simply sets of authority and hub weights, which mutually reinforce each

cause the scout to *cover little ground* for a while - he will spend some time *trapped* in the dense subgraph.

- When walking through nodes which do not belong to a dense portion of \tilde{G} , the scout will cover ground quickly, revisiting sites with low probability. The scout will tend to move about the graph rapidly, and will fall back into a dense portion of the graph (with high probability) after a small number of steps.

Armed with this intuition, we can hope to find hubs and authorities by examining the Markov chain that the scout's random walk defines. In particular, we hope that the most probable states of the walk will somehow correspond to good hubs and authorities. We will examine two different Markov chains: The chain of the scout's visits to the authority side of \tilde{G} (the *authority chain*), and the chain of his visits to the hub side (the *hub chain*). These chains naturally distinguish between the two types of sites.

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 h points to both pages i and j . This means that page j is reachable from page i by two steps: Retracting along the link $h \rightarrow i$ and then following the link $h \rightarrow j$.

An alternative way to define these matrices: Denote by W_r the matrix which results by dividing each entry of W by the sum of the entries in its row, and by W_c the matrix which results when dividing each element of W by the sum of the entries in its column (zero rows and columns of W are left unchanged in W_r and W_c , respectively). In terms of W_r and W_c , we get $\tilde{H} =$ the non-zero rows and columns of $W_r \cdot W_c^T$, and $\tilde{A} =$ the non-zero rows and columns of $W_c^T \cdot W_r$.

- The bibliographic coupling matrix and the co-citation matrix are constructed by choosing $\delta = 0$. Thus, Kleinberg’s formalization is a special case of this more general scheme.
- The matrices A and H remain symmetric and nonnegative regardless of the choice of δ . Assuming they are irreducible, they meet the convergence requirements of the power method.

Choosing the value of δ is actually tuning the sensitivity of the algorithm to how *tightly knit* the communities of hubs and authorities should be. The higher the value of δ is, the more biased the algorithm will become towards finding densely interconnected communities of hubs and authorities which have very few connections to extra-community sites. See section 4.4 for further discussion.

3.6.2 The Stochastic Approach: Analyzing a Random Walk on the Web

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

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

Each site s is represented by two nodes of \tilde{G} , s_h and s_a , with a link from site s to site q represented by an edge connecting s_h and q_a .

Communities of hubs and authorities correspond to relatively *dense bipartite subgraphs* of \tilde{G} . Informally, these are small subsets $U_h \subset V_h$ and $U_a \subset V_a$ ($|U_h| \ll |V_h|, |U_a| \ll |V_a|$) which have many interconnecting edges, but for which the cut $(U_h \cup U_a; (V_h \setminus U_h) \cup (V_a \setminus U_a))$ in \tilde{G} contains a small number of edges.

How can we find these dense bipartite subgraphs? Consider a scout performing a random walk in \tilde{G} , in a sense that when leaving a node, each of the edges incident to that node is chosen with equal probability. Each step of the scout causes him to *cross sides* of \tilde{G} , but intuition suggests the following:

- When entering a dense bipartite subgraph of \tilde{G} , the scout will tend to stay for a while in that subgraph, with high probability. This will

are symmetric, finding their irreducible components is equivalent to finding the connected components of their (undirected) support graph.

A proposed generalization of the Mutual Reinforcement Approach

Consider the bibliographic coupling matrix H . The large entries in this matrix correspond to pairs of pages which have a large intersection in the destination set of their outgoing hyper-links. The intuition about hubs suggests that two pages with a large intersection in the destination set of their outgoing links are likely to be hubs for the same topic or topics, since outgoing links represent areas of interest of the authors of the pages.

But why should an entry, $h_{i,j}$, reflect only the number of *intersecting out-links* (links which pages i and j have to jointly cited sites), and not the number of *non-intersecting out-links* (outgoing links from page i to pages that page j does not point at, and vice versa)? If the two sites are hubs for the same topic (and for no other topic), they should both be focussed on that topic and have a small number of non-intersecting links. We will try to take the non-intersecting links also into account, by building the matrix H in a more general manner, using the δ *disparity coefficient*.

Formally - denote by R_i the i 'th row of the matrix W , and by \tilde{R}_i its complement (recall that the rows of W are binary vectors). Earlier, we defined $h_{i,j} \triangleq \langle R_i, R_j \rangle$. We propose to alter this definition using a non-negative constant δ as follows:

$$h_{i,j} \triangleq \max \{ 0, \langle R_i, R_j \rangle - \delta \cdot \min (\langle \tilde{R}_i, R_j \rangle, \langle R_i, \tilde{R}_j \rangle) \}$$

A similar change is defined for the matrix A , where the inner products are calculated between the columns of W .

The motivation behind the *min* qualifier is to allow *topic generalization*, where one of hubs points to a broader topic, which includes the more narrow topic of the other hub as a subtopic. It also allows a good association score between a t -hub and a multi-topic hub, with t being one of the topics (however, two multi-topic hubs whose only joint topic of interest is t , will receive a low association score). The *max* qualifier ensures that the matrices remain non-negative.

After building these matrices, we proceed to identify the communities of hubs and authorities with the strong components of the first few principal eigenvectors of the matrices H and A , respectively.

Note the following two observations:

by the matrix W^T (recall that W is the adjacency matrix of the directed graph induced by the collection \mathcal{C} and its link-structure). 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 (it is a variation of the Power method from section 3.4.3). Furthermore, he argued that the resulting authority weights will be the coordinates of the normalized principal eigenvector of $W^T \cdot W$, and that the resulting hub weights will be the coordinates of the normalized principal eigenvector of $W \cdot W^T$.

Therefore, we can define the association matrices which are used by the meta-algorithm in this approach:

1. The *bibliographic coupling matrix* $H = W \cdot W^T$ ([18]), whose (i, j) entry is the number of sites jointly referred to (pointed at) by pages i and j . Multiplying a weights vector by H is equivalent to applying first the \mathcal{I} operation, and then the \mathcal{O} operation. Thus, the entries of the normalized principal eigenvector of H are exactly the hub weights to which Kleinberg’s iterative algorithm above will converge to.
2. The *co-citation matrix* $A = W^T \cdot W$ ([28]), whose (i, j) entry is the number of sites which jointly point at (cite) pages i and j . Multiplying a weights vector by A is equivalent to applying first the \mathcal{O} operation and then the \mathcal{I} operation. Thus, the entries of the normalized principal eigenvector of A are exactly the authority weights to which Kleinberg’s iterative algorithm will converge to.

The matrices A and H are the association matrices that are used in the meta-algorithm (as the authority and hub matrices, respectively) when exploiting the mutually reinforcing relationship between authorities and hubs. Both are symmetric and nonnegative. Whenever they are also irreducible, these matrices meet the convergence requirements of the power method.

The irreducibility constraint is not a limiting factor in this approach. A reducible authority matrix is usually a sign that the collection at hand is an orthogonal multi-topic collection, with each irreducible component corresponding to a particular topic. A reducible hub matrix is usually a sign that the collection is narrow minded (see section 2.3.1). In both cases, the reducibility of the association matrices helps us to distinguish between groups of authorities (or hubs) which pertain to different topics. So technically, if an association matrix is reducible, we simply find its irreducible components and apply the power method to each component. Since the matrices

3.5.2 Computational Aspects of the Meta-Algorithm

We shall compute the eigenvectors of the association matrices using the power method. The association matrices (used by both approaches) will be shown to comply with the convergence requirements of the power method.

In deriving the principal communities of hubs and authorities, it seems at first glance that we must compute the principal eigenvectors of both the hub matrix and of the authority matrix. However, note that we are only interested in identifying the few coordinates of highest absolute value in those eigenvectors. Typically, we are interested in the identities of about 20 coordinates of high absolute value in vectors consisting of hundreds of coordinates. The power method enables us to identify those top coordinates in far less iterations than are necessary in order to compute the entire eigenvector with high precision.

3.6 Two Approaches for Defining Association Matrices

3.6.1 The Mutual Reinforcement Approach

Kleinberg's formalization

In order to assign each site $s \in \mathcal{C}$ an authority weight $a(s)$ and a hub weight $h(s)$ according to the principals described in section 3.3, Kleinberg uses the following iterative algorithm ([19]):

1. Initialize $a(s) \leftarrow 1, h(s) \leftarrow 1$ for all sites $s \in \mathcal{C}$.
2. Repeat the following three operations:
 - Update the authority weight of each site s :
 $a(s) \leftarrow \sum_{\{x|x \text{ points to } s\}} h(x)$. This operation is called the \mathcal{I} operation.
 - Update the hub weight of each site s :
 $h(s) \leftarrow \sum_{\{x|s \text{ points to } x\}} a(x)$. This operation is called the \mathcal{O} operation.
 - 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

3.5.1 The Algorithmic Concept

- Derive, from \mathcal{C} (with $n = |\mathcal{C}|$), two $n \times n$ *association matrices* - A *hub matrix* and an *authority matrix*. Association matrices are widely used in classification algorithms ([29]), and will be used here in order to classify the web-sites into communities of hubs/authorities.
- Let $\lambda_1(M), \lambda_2(M), \dots, \lambda_{k+1}(M)$ denote the $k+1$ eigenvalues of highest magnitude of an $n \times n$ association matrix M , ordered by non-increasing absolute value (usually $k \ll n$). We assume, for the sake of simplicity, that these eigenvalues are all distinct and that $|\lambda_k(M)| > |\lambda_{k+1}(M)|$. Further, for all $1 \leq i \leq k$, denote by $v_{\lambda_i(M)}$ the unit eigenvector associated with $\lambda_i(M)$ whose first non-zero component is positive. We will identify $2k - 1$ communities of hubs/authorities with the coordinates of high absolute value in $v_{\lambda_1(M)}, \dots, v_{\lambda_k(M)}$.

We will refer to $v_{\lambda_1(M)}$ as the *principal eigenvector* of M . The association matrices we will use are such that all of the coordinates of $v_{\lambda_1(M)}$ will be positive, and so the sites which correspond to the largest coordinates will form the *principal algebraic community* of M .

The *non-principal eigenvectors* of M , $\{v_{\lambda_i(M)}, 1 < i \leq k\}$, will each have both positive and negative entries (Corollary 1). From each such eigenvector we will deduce two communities: One that corresponds to the positive coordinates of highest absolute value, and one that corresponds to the negative coordinates of highest absolute value.

For this meta-algorithm to be successful, the principal algebraic communities of hubs and authorities should resemble the “real” principal communities in \mathcal{C} , and this resemblance should continue to the non-principal communities as well (although, perhaps, in a weaker fashion).

Clearly, for a given collection \mathcal{C} , the algebraic communities produced by this meta-algorithm are determined solely by the definition of the association matrices³, and this is where the two approaches differ. Both approaches, presented in the next section, offer intuitive arguments for explaining why the principal algebraic communities, as determined by their matrices, should be close to the “real” communities.

³Under the simplifying assumption of distinct dominant eigenvalues

We can now state the following *convergence requirements* of the power method:

- When given an irreducible symmetric matrix M with some positive entries on the main diagonal, the power method (with u_0 being the all-ones vector) will converge to $v_{\lambda(M)}$ (This follows from conclusions 1 – 3 on page 21).
- When given P^T (with P being an irreducible primitive stochastic matrix), the power method (with any initial distribution vector and without any scaling) will converge to π_P (Part 3 of the Ergodic Theorem).

Calculating non-principal eigenvalues and eigenvectors of symmetric matrices

From the above description of the power method, one can see that after finding $\lambda_1(B)$ through $\lambda_{m-1}(B)$ and the corresponding eigenvectors of a symmetric matrix B , $\lambda_m(B)$ (and the corresponding eigenvector) can be calculated as follows:

- Start from an arbitrary vector $u_0 = \sum_{i=0}^n c_i v_{\lambda_i(B)}$, and use the Gram-Schmidt process ([20]) to orthogonalize it with respect to $v_{\lambda_1(B)}$ through $v_{\lambda_{m-1}(B)}$, producing $\tilde{u}_0 = \sum_{i=m}^n \tilde{c}_i v_{\lambda_i(B)}$ (Recall that eigenvectors which correspond to different eigenvalues of a symmetric matrix are orthogonal).
- Apply the power method with the initial vector \tilde{u}_0 .

Whenever $|\lambda_m(B)| > |\lambda_{m+1}(B)|$, this process will converge to $v_{\lambda_m(B)}$, provided that $c_m \neq 0$ (or, equivalently, $\tilde{c}_m \neq 0$). When $|\lambda_m(B)| = |\lambda_{m+1}(B)|$, the process still converges, but the resulting vector depends on the particular choice of u_0 .

3.5 The Meta-Algorithm for Analyzing Link Structures

In this section we describe a *single* meta-algorithm, which is applied by *both* the Mutual Reinforcement approach and the Stochastic approach for finding communities of hubs and authorities in \mathcal{C} .

Theorem 2 (Ergodic Theorem, [10]) 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 vector ² π_P which satisfies:

$$\pi_P^T \cdot P = \pi_P^T$$

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

3. For any distribution vector q :

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

3.4.3 The Power Method

The general framework

The power method ([17]) is an iterative algorithm which computes (under certain conditions which will be stated below) the eigenvalue of largest modulus and the corresponding eigenvector of a general matrix.

Let B be a square, $n \times n$ real matrix. Denote by $\lambda_1(B), \lambda_2(B), \dots, \lambda_n(B)$ the n eigenvalues of B , ordered by non-increasing absolute value. Assume that B has n linearly independent (unit) eigenvectors, and denote them by $v_{\lambda_1(B)}, v_{\lambda_2(B)}, \dots, v_{\lambda_n(B)}$.

Let $u_0 = \sum_{i=1}^n c_i v_{\lambda_i(B)}$ be an arbitrary vector, and define:

$$u_k \triangleq B^k \cdot u_0 = \sum_{i=1}^n c_i \lambda_i^k(B) v_{\lambda_i(B)}$$

When $|\lambda_1(B)| > |\lambda_2(B)|$ and $c_1 > 0$, we get:

$$\lim_{k \rightarrow \infty} \frac{u_k}{\lambda_1^k(B)} = c_1 v_{\lambda_1(B)}$$

In other words, when starting from an initial vector which is not orthogonal to $v_{\lambda_1(B)}$, and when $|\lambda_1(B)|$ is larger than the modulus of any other eigenvalue of B , then u_k tends to become proportional to $v_{\lambda_1(B)}$ as k grows.

The actual algorithm starts from an arbitrary vector u_0 . In the i 'th iteration, a vector \tilde{u}_i is calculated by $\tilde{u}_i = B \cdot u_{i-1}$. Then, \tilde{u}_i is scaled to unit length to produce u_i .

²A nonnegative real vector whose sum of entries equals 1.

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 ([10],[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 $\|v_{\lambda(M)}\| = 1$.
3. Let e denote the n -dimensional vector with 1 in all entries. Since $v_{\lambda(M)}$ is positive,

$$\langle v_{\lambda(M)}, e \rangle^1 = \sum_{i=1}^n v_{\lambda(M)}(i) > 0$$

4. If w is a real eigenvector of M which does not correspond to $\lambda(M)$, then w has both positive and negative entries.

The following are well known facts in linear algebra ([20]): Let C be a $n \times n$ real square *symmetric* matrix. Then, all of C 's eigenvalues are real. In addition, C has n linearly independent eigenvectors which span \mathbf{R}^n , and eigenvectors which correspond to different eigenvalues are orthogonal in \mathbf{R}^n .

3.4.2 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.*

¹The dot product of the two vectors

them. The issue of combining the graph theoretical clustering techniques with some form of ranking sites inside the resulting clusters is left for future research.

3.4 Mathematical Background

3.4.1 Irreducible Symmetric Matrices

Let $B = [b_{i,j}]$ denote a square $n \times n$ real matrix with nonnegative entries. Denote by $\lambda_1(B), \lambda_2(B), \dots, \lambda_n(B)$ the n eigenvalues of B , ordered by non-increasing absolute value. In particular, $|\lambda_1(B)|$ is the spectral radius of B ([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 ([24]): $G(B)$ has n nodes (corresponding to the n rows of B), with a directed edge $i \rightarrow j$ if and only if $[B]_{i,j} = b_{i,j} > 0$.

Definition 1 ([24]) *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 ([24]).*

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

Theorem 1 (Perron-Frobenius Theorem for irreducible matrices, [14]) *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 $1 \leq i \leq n$, $b_{i,i} > 0$ ([14]).*

The Stochastic Approach is based upon the theory of Markov chains and random walks. When our base set \mathcal{C} is built around a topic t , authoritative sites on topic t should be visible (pointed at) from many sites in the subgraph induced by \mathcal{C} . This, in turn, suggests that a random walk in this subgraph will visit t -authorities with high probability.

The idea of using random walks to rank web-sites is not new. The *PageRank* algorithm used in the *Google* search engine ([2], [15]) incorporates such stochastic information into its ranking of pages. However, 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. State transitions in these chains do not arise from following a single hyperlink of the Web, but rather from following one link forward and one link “backwards” (or vice versa). Analyzing both chains allows our approach to give each web site two distinct rankings, a hub rank and an authority rank, a distinction which the “conventional” random walk cannot make.

Both approaches make use of $|\mathcal{C}| \times |\mathcal{C}|$ nonnegative *association matrices*, which define some measure of association between each ordered pair of sites $(i, j) \in \mathcal{C} \times \mathcal{C}$. When the association matrix is symmetrical (as is the case in the Mutual Reinforcement approach), a weighted undirected graph with $|\mathcal{C}|$ nodes can be derived from it, and classic graph theoretical clustering techniques ([1],[29]) can be applied. These techniques set a positive threshold τ , delete all the edges in the graph whose weight is less than τ , and then identify the clusters in one of the following manners:

- The clusters are the connected components of the resulting graph whose size exceeds some threshold.
- The clusters are the maximal complete subgraphs of the resulting graph.
- The clusters are some combination of the maximal complete subgraphs.

The same techniques can be applied to directed graphs, derived from non-symmetric association matrices.

In this work, we do not study the application of these clustering techniques on our collection \mathcal{C} (except for some comments in section 4.1.2). Our work is mostly concerned with *ranking* web sites, in addition to *classifying*

Formally, we have two thresholds - i for the in-degree and o for the out-degree, and we repeat the following iteration until G reaches a steady state:

- For each node $v \in V(G)$ such that $in - degree(v) < i$ - delete all incoming edges to v .
- For each node $v \in V(G)$ such that $out - degree(v) < o$ - delete all outgoing edges from v .

This iterative trimming of G can clear irrelevant links from the graph, and possibly help the algorithms converge more quickly and more accurately to the communities of hubs and authorities.

From the (trimmed) graph G , we derive the $|\mathcal{C}| \times |\mathcal{C}|$ adjacency matrix W . Thus, $w_{i,j} = 1$ if and only if page i ($i \in \mathcal{C}$) has an untrimmed outgoing hyper-link to page j ($j \in \mathcal{C}$).

3.3 Overview of the Algorithms

This section presents the basic ideas behind the algorithmic approaches which we apply for analyzing link structures. The exact details will be explained in later sections.

The Mutual Reinforcement Approach was presented by Kleinberg in [19]. It aims to assign each site $s \in \mathcal{C}$ a pair of weights: A hub-weight $h(s)$ and an authority weight $a(s)$, based on the following two principals:

- The quality of a hub is determined by the quality of the authorities it points at. Therefore, a site’s hub weight should reflect 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. Hence, a site’s authority weight should reflect 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.

We propose a generalization of this approach, which enables it to control how *tightly knit* the relationship between the hubs and authorities should be: It essentially “fines” hubs which point to sites outside the community of authorities, and fines authorities which are pointed at by sites outside the community of hubs.

There are a few computational and algebraic aspects which have to do with the application of the meta-algorithm to identify the non-principal communities of hubs and authorities in T . In section 3.7 we present these aspects, and suggest an alternative method for arriving at these non-principal communities.

3.2 Description of the Data

In this section, we describe the common starting point for all algorithms and heuristics - the method for obtaining the relevant data. The data will consist of a collection \mathcal{C} of pages (web-sites), which should contain communities of hubs and authorities pertaining to a topic of interest. Denote by q a term-based search query to which pages in our topic of interest are deemed to be relevant. The collection \mathcal{C} is assembled in the following manner (first described in [19]):

- A *root set* S of k pages is obtained by applying a term based search engine, such as AltaVista [8], to our query q . This is the only step in our work in which the lexical contents of the web pages are examined.
- From S we arrive at a *base set* \mathcal{C} which consists of the pages in S , pages which point to a page in S , and pages which are pointed to by a page in S .

We will search for hubs and authorities within the base set \mathcal{C} , by examining the hyper-links between the web sites in this set.

3.2.1 Refining the Link-Structure Induced by \mathcal{C}

Denote by G the directed subgraph of the WWW induced by the pages in \mathcal{C} and their link structure. We introduce a heuristic called *iterative (i,o)-trimming* which will attempt to alter G in a manner that will facilitate our algorithms in finding hubs and authorities.

This heuristic follows the notion that nodes (=web sites) which have a low in-degree in G are unlikely to be authoritative, and nodes which have a low out-degree are unlikely to be hubs. All links outgoing from the non-hubs and all links incoming to non-authorities are thus considered *noisy*, and are deleted from the graph G (G is *trimmed*). Trimming results in new nodes having low in/out degrees, and the process is carried on iteratively until a steady state is achieved.

Chapter 3

Proposed Algorithms and Heuristics

3.1 Introduction

In the previous chapters, we recounted Kleinberg’s definitions of hubs and authorities in the setting of the WWW, the Mutually Reinforcing relationship which exists between the two types of pages, and the communities that groups of hubs and authorities are expected to form. We now set forth in examining known algorithms, and proposing new approaches, for finding communities of hubs and authorities in hyper-linked media.

We will assume a collection \mathcal{C} of Web-pages, which contains communities of authorities and hubs pertaining to topics of interest with broad representation in the Web. The method for arriving at this set is described in section 3.2.

Using only the link-structure of the pages in \mathcal{C} , our main goal will be to identify the *principal community* of hubs and authorities which lie therein. In section 3.3, we provide an overview of the algorithms and heuristics which are discussed in detail later in this chapter. In section 3.4, we bring the essential mathematical background on which the algorithmic approaches rely.

We propose a generalization of Kleinberg’s *Mutual Reinforcement approach*, as well as a new *Stochastic approach* which relies on the stochastic properties of random walks performed on our set of pages, for finding the communities of hubs and authorities. Interestingly, both approaches apply the same meta-algorithm, which will be discussed in section 3.5. The approaches themselves will be presented with detail in section 3.6.

The weight structure of the various topics has implications on the communities of hubs and authorities (see page 5) in the collection: In collections which contain an obvious major topic, we expect the authorities of that topic to dominate the principal community of authorities, and the authorities of the minor topics to dominate distinct non-principal communities of authorities.

2.3.3 Noise in Multi-Topic Collections

In addition to noisy links stemming from sites which are not hubs for any of the topics, we must observe that the set of informative links for all topics other than topic t act as noise for any algorithm aiming to find (or distinguish) t -authoritative pages. This noise is more difficult to filter than noise in a single topic setting, since unlike the random noise of that scenario, the t -noisy links in a multi-topic setting are ordered in patterns similar to that of the t -informative links. Especially disruptive is noise stemming from an overlap between the set of t -hubs and hubs for other topics.

By the above definitions, two topics $t, u \in T$ will be called *orthogonal* if they have no common authoritative sites (hence $\langle \mathcal{V}_t \cdot \mathcal{V}_u \rangle = 0$), and a collection will be called an *orthogonal multi-topic collection* if every two different topics in the collection are orthogonal. In orthogonal collections, each site is an authority on at most one topic.

Another interesting special case of collections, related to orthogonal collections, is one in which each site is a good hub for at most one topic. We call such a collection *narrow minded*, since it seems that any site which is at all interested in some area, has just a single area of interest. However, hubs for different topics in a narrow minded collection may not be orthogonal: A good *t-hub* might still have a link to a *u-authority* without being a good *u-hub*. Such links are non-informative from the standpoint of topic t , but do not collect enough *u-weight* in order to qualify the hub as a *t-hub*.

2.3.2 Weight-Structures in Multi-Topic Collections

Multi-topic collections may consist of one major topic, and a few minor ones. Other collections may contain a few equally dominant topics. These cases might prove to pose different challenges for algorithms which aim to find the topics and distinguish them from each other.

What are the parameters that influence us in saying that some topics are *major* and others are *minor*? We claim that the *dominance* of a topic is a result of its *weight structure*, which is a combination of the number of authorities on that topic, and their quality. We therefore define the *weight structure of topic t* , \mathcal{W}_t , to be the *multi-set* of *t-weights*:

$$\mathcal{W}_t \triangleq \{a_t(s_1), a_t(s_2), \dots, a_t(s_n)\}$$

There is no natural meaningful ordering between every two structures \mathcal{W}_t and \mathcal{W}_u . For instance, which is more dominant, a topic with $k+1$ perfect authorities or a topic with 1 perfect authority and $2k$ authorities with weight 0.5?

We should, however, comment on the issue of *symmetric topics*: Topics $t, u \in T$ are called *symmetric* if $\mathcal{W}_t = \mathcal{W}_u$, and a T -topic collection \mathcal{C} is called symmetric if every $t, u \in T$ are symmetric. Note that symmetry between topics implies that both the number and the quality of authorities for the two topics are the same. It does not imply, however, that the multi-sets of hub weights for the two topics will be equal, or even similar. Also, symmetry is independent of the issue of diversity: $\theta_{\mathcal{C}}$ cannot insinuate symmetry (or lack thereof) of any two topics in \mathcal{C} .

is the amount of noisy links it can be subjected to while still finding the authoritative sites for the (single) topic at hand.

2.3 Multi-Topic Collections

Let us now assume that our collection of sites consists of authoritative pages for a set $T = \{t_1, t_2, \dots, t_m\}$ of topics. Such a collection \mathcal{C} will be called a *T-topic collection*. A site $s \in \mathcal{C}$ may be authoritative on any subset of the topics. We thus associate with each site an *m-dimensional vector* $\mathcal{A}(s) = (a_{t_1}(s), a_{t_2}(s), \dots, a_{t_m}(s))$, where $a_{t_i}(s)$ is the authority weight of site s on topic t_i . Accordingly, hubs are now defined by an *m-dimensional vector of ordered pairs* $\mathcal{H} = ((p_{t_1}, \epsilon_{t_1}), \dots, (p_{t_m}, \epsilon_{t_m}))$. We will call a multi-topic collection consistent when it is consistent for each of its topics.

2.3.1 Measuring Topic Diversity

We now observe that a hub can focus on two different topics $t, u \in T$ (have low ϵ values for both topics) only when there is some meaningful similarity between the set of t -authorities and the set of u -authorities. Specifically, those two sets must have a large intersection. This brings about the following question: Since we identify topics only by their respective sets of authorities, can we expect any algorithm to distinctly identify topics with similar authority sets?

To quantify this notion, we associate with each topic $t \in T$ the *n-dimensional weights vector* $\mathcal{V}_t \triangleq (a_t(s_1), a_t(s_2), \dots, a_t(s_n))$ - The ordered set of t -weights in our n -site collection \mathcal{C} . Following [25], for topics $t, u \in T$ we define

$$\theta_{t,u} \triangleq \arccos \frac{\langle \mathcal{V}_t \cdot \mathcal{V}_u \rangle}{\|\mathcal{V}_t\| \cdot \|\mathcal{V}_u\|}$$

$\theta_{t,u}$ is the angle between the weight-vectors of topics t and u , and the greater this angle is, the more distinct the two topics are. With every *T-topic* collection \mathcal{C} we associate a value $\theta_{\mathcal{C}}$ as follows :

$$\theta_{\mathcal{C}} = \min_{t,u \in T (t \neq u)} \theta_{t,u}$$

$\theta_{\mathcal{C}}$ is the minimal angle observed in \mathcal{C} between any two different topics. The smaller $\theta_{\mathcal{C}}$ is, the harder it will be for any algorithm to distinguish between all of the different topics in \mathcal{C} .

by a search engine and the person does not publicize it, chances are that no links will point to that site ([23]).

We thus see that posing requirements only on the outgoing links of hubs may result in a link structure which suggests an authority distribution that does not reflect the “true” picture, as given by the predetermined authority weights. We would like to have a tighter coupling between strong authorities and good hubs: Good hubs should point at strong authorities, and strong authorities should be pointed at by good hubs (this coupling is in the heart of Kleinberg’s *Mutual Reinforcing Relationship*, [19]). Thus, we conclude that for a link structure to be informative, it must demand some properties of the incoming links of authorities.

We therefore introduce the notion of a *consistent collection*, which poses restrictions on the incoming links to authorities from high-quality hubs:

For a site s , denote by $\mathcal{L}_s(p, \epsilon)$ the set of links outgoing from all (p, ϵ) t -hubs to s . A single-topic collection (with topic t) is called (p, ϵ) -*consistent* if the following property holds for all site pairs s_i, s_j :

$$a_t(s_i) > a_t(s_j) \implies |\mathcal{L}_{s_i}(p, \epsilon)| \geq |\mathcal{L}_{s_j}(p, \epsilon)|$$

This consistency constraint is quite *minimal*, in a sense that it does not require $|\mathcal{L}_{s_i}(p, \epsilon)|$ to be strictly greater than $|\mathcal{L}_{s_j}(p, \epsilon)|$. An even stronger requirement could have tied the two ratios $\frac{a_t(s_i)}{a_t(s_j)}$ and $\frac{|\mathcal{L}_{s_i}(p, \epsilon)|}{|\mathcal{L}_{s_j}(p, \epsilon)|}$ in some quantitative manner.

2.2.3 Noisy Links In Single-Topic Collections

In our Web model so far, we have assumed the existence of authorities and hubs. However, in reality, most of the sites might not be authorities ($a_t(s) = 0$ for most sites s), and certainly not all sites are hubs. For our model to be complete, we must address the question of the outgoing links from non-hubs, as well as that of the non-informative outgoing links of hubs (those links which connect a hub to non-authoritative sites). When a site has d non-informative outgoing links, we assume all selections of d non-authorities as destinations for those links to be equiprobable. Note that d is not necessarily the out-degree of the site.

The non-informative links are called *noisy* since they act as *chaff* - they hide and camouflage the informative links. Since, in realistic situations, most of the sites in the collection will neither be authorities nor hubs - most of the links will be noisy, and some measure of the quality of an algorithm

have $p = 1$ and $\epsilon = 0$. Clearly, such hubs may not always exist: A necessary condition for the existence of perfect t -hubs is that $|\mathcal{Best}(t)| \geq \beta(t) \cdot \mathcal{Sum}(t)$.

In addition to perfect hubs, there may be many unattainable values of the pair (p, ϵ) for t -hubs in a given collection. To see this, assume (without loss of generality) that $a_t(s_1) \geq a_t(s_2) \geq \dots a_t(s_n)$. Then:

- Given a recall constraint \tilde{p} , let $1 \leq j \leq n$ be minimal such that

$$\sum_{i=1}^j a_t(s_i) \geq \tilde{p} \cdot \beta(t) \cdot \mathcal{Sum}(t)$$

Every (\tilde{p}, ϵ) t -hub satisfies

$$\epsilon \geq 1 - \frac{\sum_{i=1}^j a_t(s_i)}{j}$$

- Given a precision constraint $\tilde{\epsilon}$, let $1 \leq j \leq n$ be maximal such that

$$1 - \frac{\sum_{i=1}^j a_t(s_i)}{j} \leq \tilde{\epsilon}$$

Every $(p, \tilde{\epsilon})$ t -hub satisfies

$$p \leq \frac{\sum_{i=1}^j a_t(s_i)}{\beta(t) \cdot \mathcal{Sum}(t)}$$

2.2.2 Consistent Collections

We have defined hubs for single-topic collections of sites with authority weights. The recall and precision criteria by which hubs are judged require good hubs to point to strong authorities. However, no requirements were posed so far on the incoming links of authorities. It is entirely possible for good and even perfect hubs to exist, without any of them pointing at some perfect authority s (hubs are not expected to point at all of the authorities, just at a $\beta(t) \cdot \mathcal{Sum}(t)$ portion of them). Such a situation will clearly prevent us from identifying s as an authority. Note that such isolated authorities are very realistic in the setting of the WWW - A person can set up a web-site which is extremely authoritative on a topic, but if that site is never crawled

although the situation may be such that the vast majority of sites in the collection are not t -authoritative.

2.2.1 Probabilistic Definitions for Hubs

As a part of our single-topic collection of sites, we expect the existence of *hubs*. In order to define the properties of hubs, we need to address the issue of *qualities of a good hub* in a single topic collection. The following are well-accepted criteria in the field of *Information Retrieval*, which can be applied to hubs:

- *recall* - A good hub should point at many authoritative pages.
- *precision* - A good hub should remain focussed on the topic at hand, and not point at many non-authoritative pages.

Still, these intuitions do not supply concrete enough grounds for a formal definition of a hub. In particular, two problems arise from our requirement of *recall*:

1. Should a hub point to a big portion of the authorities regardless of the breadth of the topic, or the number of authorities in the collection of sites? Clearly, in the setting of the WWW, one does not expect a hub for a broad topic to include hundreds of links to authorities on the topic. Therefore, for each topic t , we introduce a *breadth coefficient*, $0 \leq \beta(t) \leq 1$, which defines the portion of authorities for topic t which we would expect a hub to point at.
2. How do the authority weights $a(s)$ come into effect? Should a hub point only to very good authorities? Should links to sites with different authority weights be treated differently?

To cope with the above issues, we introduce the following definition. A site h is a (p, ϵ) t -hub if the following two conditions are satisfied:

1. $\sum_{\{s|h \text{ points to } s\}} a_t(s) \geq p \cdot \beta(t) \cdot \text{Sum}(t)$
2. $\frac{\sum_{\{s|h \text{ points to } s\}} a_t(s)}{\text{The out degree of } h} \geq 1 - \epsilon$

Hence, a good t -hub in a single-topic collection should have a high p value (exhibit high recall) and a low ϵ value (be precise). A *perfect* hub will

site s is either relevant to topic t (in this case $a_t(s) = 1$), or is irrelevant ($a_t(s) = 0$).

Now that we have authority weights, let us define the following:

- $Sum(t) \triangleq \sum_{i=1}^n a_t(s_i)$ - The sum of all t - *weights* in the collection.
- $Best(t) \triangleq \{s | a_t(s) = 1\}$ - The *perfect* authorities for topic t .

We assume that $Best(t) \neq \phi$ - That is, for every topic, there is at least one perfect authority. This implies that authority weights are normalized, in the following two senses :

1. Intra-topic aspect: The set of t -authoritative pages are measured in a relative manner, with the most authoritative site being called *perfect*.
2. Inter-topic aspect: The best authorities for *any* topic receive the *same* authority weight of 1.

2.1.2 Informative Link Structure

We are interested in finding authoritative web sites by studying the link structure of the WWW. For this to be at all possible, we must assume that the link structure of the web reflects, in some manner, the notion of authority we have defined above. For any link-based algorithm to succeed, the links present in the web must carry some information on the identities of the authoritative sites.

Therefore, when given a collection of sites with authority weights already assigned to them, the link structure must depend in some manner on those weights. Our quantitative measures attempt to assess how well the link structure reflects those weights. Following Kleinberg's interpretation of the significance of links in hyper-linked media [19], we will assume that a link from page p to page q conveys some measure of confidence which page p has in the contents of q .

2.2 Single Topic Collections

A collection of sites will be termed a *single topic collection* when $|T| = 1$. In such a collection, all the authoritative sites correspond to the same topic t . Since we are concentrating on methods of finding authoritative pages for *broad topic queries* (see page 2), we shall assume that $Sum(t)$ is quite large,

Chapter 2

Quantitative Measures for Assessing Informative Link Structures

2.1 Introduction

We shall define a few quantitative measures for assessing collections of hyperlinked media. The definitions will be incremental, and will rely on the notions of *hubs* and *authorities*, introduced in [19]. Throughout these definitions, we shall assume we have a collection of n hyper-linked sites $\mathcal{C} = \{s_1, s_2, \dots, s_n\}$ whose contents is relevant to a set of *topics* T .

2.1.1 Authority Weights

We shall assume that each site $s \in \mathcal{C}$ has an *authority weight*, $0 \leq a_t(s) \leq 1$ for every topic $t \in T$. Possible ways for assigning authority weights are:

- *Keyword Appearance Rule* - In cases where topic t is defined by a query containing m keywords, each site can be given a t -weight which equals the percentage of all t -keywords which appear in that site.
- *Human Judgment Rule* - The sites can be given weights according to human evaluation of their relevancy to each topic.

As noted above, the weights may be any real numbers in $[0, 1]$. A special important case is when all weights are binary (the *Zero-One Rule*): Each

and authorities, both Kleinberg's Mutual Reinforcement approach and our Stochastic approach employ the same *meta-algorithm*.

Chapter 4 recounts theoretical and experimental results of applying the proposed algorithms to many hyperlinked collections, including artificial topologies, digital libraries, and the WWW. In particular, these experiments point out the *TKC Effect* - A phenomenon which demonstrates an inherent difference in the way in which the Mutual Reinforcement approach and the Stochastic approach rank authorities in certain topologies.

The importance of experimenting on the WWW and on digital libraries is self evident. The artificial topologies are of importance as well: These topologies (which are built according to our probabilistic models and by combinatorial methods) allow us to control many topological parameters, over which we have little or no control (and sometimes even no knowledge) in the WWW. Controlling the topology facilitates the evaluation of link-analyzing algorithms. For example, in artificial topologies we have complete knowledge on the identities of the authorities, and we can measure the effectiveness of algorithms which attempt to find them.

Finally, chapter 5 brings our conclusions, and suggestions for further research.

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, and the smaller (or less densely connected) communities in the collection (if they exist) are called the *non-principal* communities. Kleinberg also suggested an algorithm to identify these communities. His algorithm is described in detail in chapter 3, along with our suggestions for ways to generalize and improve it.

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 [6], augments Kleinberg's link-structure analysis by considering also the textual surrounding of the hyperlinks. 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]).

The text found around hyperlinks has also been used by Brin and Page in [2]. They consider *anchor text*, the text which describes the hyperlink in the *pointing* page, as a source of information on the contents of the *pointed* page. Another major feature of their work on the *Google* search engine ([15]) 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 WWW.

In [21], the authors use the link-structure of an initial 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 contents of this collection is then analyzed in order to rank the relevancy of its individual pages.

This work We start by defining quantitative measures for assessing informative link structures (chapter 2). These measures provide the basis for defining theoretical, probabilistic models for collections of hyperlinked media which exhibit informative link structures.

In Chapter 3 we generalize Kleinberg's approach for finding authoritative pages, and suggest some alternative computational procedures for various stages in it. We also present a new *stochastic* approach for identifying authorities, which examines random walks on graphs derived from the link structure. We show that in finding the principal communities of hubs

sored site. There is no necessary mutual topic of interest between the advertiser/sponsor and the ad-hosting site.

- Links resulting from link-exchange agreements. In order to increase visibility, many sites negotiate link-exchanges with other sites. Again, no mutual topic of interest must be present, just a mutual will for improved visibility. These links increase the *link popularity* (number of incoming links) of both participating sites, thus improving their rankings with search engines that consider this measure (such as *Lycos* [16]).

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 ([28],[18]). Many works in this area were aimed at finding high-impact papers published in scientific journals ([11]), and at clustering related documents ([1]).

Some works have studied the Web's link structure, in addition to the textual contents 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 ([30],[26]).

Marchiori, in [23], 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.

The main previous work on which this work is based, is due to Jon Kleinberg ([19]). 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. Note that a page can be both a hub and an authority.

In light of the mutually reinforcing relationship, hubs and authorities should form *communities*, which can be pictured as dense bipartite portions

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],[7]). Here, the authors repeat certain keywords many times, and include other keywords which are more remotely connected to their site's context, in order to "lure" search engines into ranking them highly for many queries. Many adult sites, for instance, *spam* their sites with names of actresses and models.

Informative link structure - The answer? The WWW is a hyper-linked collection. In addition to the textual contents 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 (sometimes 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 page p thinks highly of q 's contents. Such a link, called an *informative link*, is p 's way to *confer authority* on q ([19]). 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 information derived from q 's textual contents, which is under complete control of q 's author). This makes the information extracted from such links less vulnerable to manipulative techniques (such as spamming).

Alas, not all links are informative. There are many kinds of links which confer little or no authority ([5]), such as the following link types:

- Intra-domain (*inner*) links - These links connect pages that belong to the same complex site of some organization. Many times, their purpose is to provide navigational aid in the complex site, helping surfers find different sections of the site comfortably. In addition, as these links are under the direct control of the owning organization, the authority which they confer is dubious ([23]).
- Commercial links and sponsor links - Many sites host hyperlinked commercials, and many sites are sponsored by other commercial entities, which in return get a promotional hyperlinked banner in the spon-

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 sea of resources - Most users will be quite satisfied with a few dozen *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 will focus on finding authoritative resources which pertain to broad-topic queries.

Term-based WWW 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 ([25],[5]):

- Synonymy - Retrieving documents containing the term “car” when using the query “automobile”.
- Polysemy/Ambiguity - When given the query “Jordan”, which kind of pages should be retrieved? Those pertaining to the Hashemite Kingdom of Jordan, or those pertaining to 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).

These issues make retrieving the set of Web-pages which pertain to a broad-topic query very difficult. Then, there is the task of *ranking* the relevance of the many retrieved pages, in which those same three issues reappear. This leads us to a specific problem of searching the WWW, called *search engine persuasion* ([23]). There are sometimes millions of sites pertaining in some manner to broad-topic queries, but most users will not browse through more than the first few dozen 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

Chapter 1

Introduction

Searching the WWW - The Challenge The WWW is a rapidly expanding hyperlinked collection of unstructured information. Its volume has increased exponentially in recent years, along with a huge rise in the number of page creators and Web surfers. The complete lack of structure (practically anyone can post information of any kind in whatever way, shape or form on the Web) and the enormous volume of 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.

Traditionally, Information Retrieval systems are evaluated by two parameters:

- *Recall*: The fraction of documents relevant to the query retrieved, out of all relevant documents in the collection.
- *Precision*: The fraction of documents relevant to the query retrieved, out of all retrieved documents.

When discussing the difficulties which WWW search engines face, we distinguish between *narrow-topic* queries and *broad-topic* queries. The 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.

ABSTRACT

Today, when searching for information on the WWW, one usually performs a query through a search engine. Many search engines are *term-based*, and return, as the query's result, a list of Web pages whose *contents* match the query. For *wide 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 hyperlinked media. Information such as which pages are linked to by others, how many such hyperlinks exist, etc. can be used to augment the search results. In this context, Jon M. Kleinberg (in his paper *Authoritative Sources in a Hyperlinked Environment* from 1998) introduced the following notions:

1. *Authoritative pages* - A small subset containing the most definitive pages, from amongst the large number of pages which match the query. These pages will usually have many incoming links, and will also be termed *authorities*.
2. *Hub pages* - Pages which have links to multiple authoritative pages.

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 web sites.

In our work, we define quantitative measures for assessing *informative* link structures. Based on these measures, we provide theoretical, probabilistic models for hyperlinked media with an informative link structure. In these artificial topologies we have complete knowledge on the identities of the authorities, and we can measure the effectiveness of algorithms which attempt to find them. We also generalize Kleinberg's approach for finding authoritative pages, and suggest some alternative computational procedures for various stages in it. We then present a new *stochastic* approach for identifying authorities, which examines random walks on graphs derived from the link structure. We show that both Kleinberg's Mutual Reinforcement approach and our Stochastic approach employ the same *meta-algorithm*. Finally, we compare the results of applying our algorithmic ideas to the results derived through Kleinberg's approach on the WWW and on other hyperlinked topologies. By analyzing cases where the Stochastic approach yields different results than Kleinberg's approach, we isolate a particular topological phenomenon which causes the two approaches to disagree.

List of Tables

4.1	Mutual Reinforcement approach - Sparse 0/1 Model	40
4.2	δ -Enhanced Mutual Reinforcement approach - Dense 0/1 Model	40
4.3	Sparse Truncated Exponential Model - Authorities	41
4.4	Sparse Truncated Exponential Model - Hubs	42
4.5	Dense Truncated Exponential Model (δ -Enhanced Results) .	43
4.6	Authorities for “Shift register sequences”	53
4.7	Mutual Reinforcement Authorities for “Error correcting codes”	54
4.8	Stochastic Authorities for “Error correcting codes”	55
4.9	Stochastic Authorities for “Source coding”	56
4.10	IEEE-IT multi-topic Mutual Reinforcement authorities	57
4.11	IEEE-IT multi-topic non-prin. communities (Gram-Schmidt)	58
4.12	IEEE-IT multi-topic non-prin. communities (deletion)	59
4.13	Authorities for WWW query “Java”	61
4.14	Authorities for WWW query “+censorship +net”	63
4.15	Mutual Reinforcement Authorities for WWW query “movies”	64
4.16	Mutual Reinforcement Hubs for WWW query “movies”	64
4.17	Stochastic authorities for WWW query “movies”	65
4.18	WWW query “movies”: Non-principal authorities (deletion) .	65
4.19	Authorities for WWW query “Jordan”	67
4.20	WWW query “Jordan”: Non-principal authorities (deletion) .	68
4.21	Authorities for WWW query “Abortion”	69
4.22	Non Principal Authorities for WWW query “Abortion”	70
4.23	Authorities for WWW query “genetic”	71
4.24	WWW query “genetic”: Non-principal authorities (deletion) .	72
4.25	WWW query “genetic”: Non-prin. authorities (Gram-Schmidt)	73
4.26	Number of Required Iterations Until Convergence	74

Contents (continued)

3.5.1	The Algorithmic Concept	24
3.5.2	Computational Aspects of the Meta-Algorithm	25
3.6	Two Approaches for Defining Association Matrices	25
3.6.1	The Mutual Reinforcement Approach	25
3.6.2	The Stochastic Approach: Analyzing a Random Walk on the Web	28
3.7	Finding Non-Principal Communities	30
4	Results	33
4.1	Simulated/Artificial Web Topologies	33
4.1.1	Single Topic Topologies	33
4.1.2	Multi-Topic Topologies	44
4.2	IEEE-IT Web	51
4.2.1	Single Topic Queries	51
4.2.2	Multi-Topic Queries	57
4.3	The WWW	60
4.3.1	Single Topic Queries	60
4.3.2	Multi-Topic Queries	66
4.4	Evaluating the Results	74
4.4.1	The Impact of Iterative (i, o) -Trimming	74
4.4.2	The Effect of the δ Disparity Coefficient	75
4.4.3	The Deletion Method	75
4.4.4	The Stochastic Approach	76
5	Further Research and Conclusions	77
A	Proofs of propositions	82

Contents

Abstract	1
1 Introduction	2
2 Quantitative Measures for Assessing Informative Link Structures	8
2.1 Introduction	8
2.1.1 Authority Weights	8
2.1.2 Informative Link Structure	9
2.2 Single Topic Collections	9
2.2.1 Probabilistic Definitions for Hubs	10
2.2.2 Consistent Collections	11
2.2.3 Noisy Links In Single-Topic Collections	12
2.3 Multi-Topic Collections	13
2.3.1 Measuring Topic Diversity	13
2.3.2 Weight-Structures in Multi-Topic Collections	14
2.3.3 Noise in Multi-Topic Collections	15
3 Proposed Algorithms and Heuristics	16
3.1 Introduction	16
3.2 Description of the Data	17
3.2.1 Refining the Link-Structure Induced by \mathcal{C}	17
3.3 Overview of the Algorithms	18
3.4 Mathematical Background	20
3.4.1 Irreducible Symmetric Matrices	20
3.4.2 Irreducible Stochastic Matrices	21
3.4.3 The Power Method	22
3.5 The Meta-Algorithm for Analyzing Link Structures	23

THE RESEARCH THESIS WAS DONE UNDER THE
SUPERVISION OF PROF. SHLOMO MORAN IN THE
FACULTY OF COMPUTER SCIENCE.

I wish to thank Shlomo Moran for his devoted and thorough guidance
throughout the course of this research.

THE GENEROUS FINANCIAL HELP OF THE TECHNION
IS GRATEFULLY ACKNOWLEDGED.

**FINDING AUTHORITATIVE SITES ON THE WWW
(AND OTHER HYPERLINKED MEDIA)
BY ANALYZING THE WEB'S LINK-STRUCTURE**

RESEARCH THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN COMPUTER SCIENCE

RONNY LEMPEL

SUBMITTED TO THE SENATE OF
THE TECHNION – ISRAEL INSTITUTE OF TECHNOLOGY

AV 5759

HAIFA

JULY 1999

**FINDING AUTHORITATIVE SITES ON THE WWW
(AND OTHER HYPERLINKED MEDIA)
BY ANALYZING THE WEB'S LINK-STRUCTURE**

RONNY LEMPEL