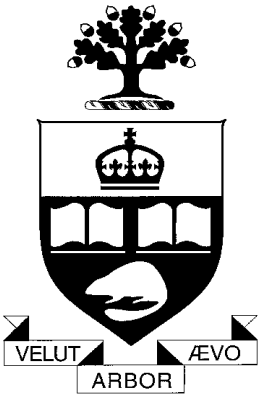

Downweighting Tightly Knit Communities in World Wide Web Rankings

by

Gareth O. Roberts
Department of Mathematics and Statistics
Lancaster University

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 0302 January 5, 2003

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Downweighting Tightly Knit Communities in World Wide Web Rankings

by

Gareth O. Roberts[1] and Jeffrey S. Rosenthal[2]

(January 2003.)

Abstract

We propose two new algorithms for using World Wide Web link structures to determine authority values of web pages from search queries. Both algorithms postulate an underlying latent cluster structure, in an effort to avoid the Tightly Knit Community (TKC) effect which can occur in the Kleinberg and SALSA algorithms. The first algorithm, Similarity Downweighting (SD), weights outlinks inversely with apparent cluster size. The second algorithm, Sequential Clustering (SC), first generates an underlying cluster structure consistent with the observed links, and then uses an empirical Bayes approach to compute authority values. We present experiments indicating that both algorithms do a fairly good job of selecting authoritative web pages for a given query, given only the link structure of the Base Set of pages, while largely avoiding the TKC effect. We also consider a fully Bayesian approach, but find that it is too sensitive to prior information to be useful at this point.

Keywords: tightly knit communities, link analysis, web searching, hubs, authorities, SALSA, Kleinberg's algorithm, clusters, Bayesian.

1 Introduction

In recent years, a number of papers [2, 5, 7, 6, 3, 1] have considered the use of hypertext links to determine the authority value of different web pages. In particular, these papers consider the extent to which hypertext links between World Wide Web documents can be used to determine the relative authority values of these documents for various web searches.

Such use of the Web's link structure poses a number of problems. One is the tightly knit community (TKC) effect [6], in which a number of web pages all link to each other and thus all receive high authority values, even though their true value may be much smaller.

In this paper, we present a new algorithm to deal with the TKC effect, and avoid giving too much weight to pages simply because they are members of TKCs. The key to our method is considering the underlying *cluster structure* of which web nodes tend to link to the same pages. Based on this, we propose several algorithms. The first is a simple model for computing the "similarity" of the link

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K. LA1 4YF. E-mail: g.o.roberts@lancaster.ac.uk. Web: http://www.maths.lancs.ac.uk/dept/people/robertsg.html

[2]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Supported in part by NSERC of Canada. E-mail: jeff@math.toronto.edu. Web: http://probability.ca/jeff/

structures of two pages, and using that to appropriately determine authority weights. The second is a sequential clustering algorithm which probabilistically generates an underlying cluster structure, and then uses this to compute appropriate authority weights. The third is a fully Bayesian model together with a Metropolis algorithm for computing cluster structures and authority values.

2 Base Sets and Cluster Structures

Based on the user's query term, a Base Set $\mathcal{N}$ of web page nodes is generated, as in [5]. Thus, web pages in $\mathcal{N}$ are somehow related to the query term, either because they contain the term, or they link to a page which contains the term, or they are linked to by a page which contains the term. However, the pages in $\mathcal{N}$ may or may not be good authorities for the query term. In this paper, we assume that $\mathcal{N}$ has already been generated; for more details, see e.g.

`http://www.cs.toronto.edu/~tsap/experiments/www10-experiments/help.html`

The data is in the form of adjacency values $\{A_{ij}\}$ for $i, j \in \mathcal{N}$, where $A_{ij} = 1$ if there is a link from i to j and $A_{ij} = 0$ otherwise. The question is, how can this data best be used to estimate authority values a_j for $j \in \mathcal{N}$. One simple method is to let $a_j = \sum_{i \in \mathcal{N}} A_{ij}$ count the number of pages in $\mathcal{N}$ which link to j; this is the pSALSA variant [1] of the SALSA algorithm [6]. However, this algorithm is still somewhat vulnerable to the TKC effect, since if many pages all link to each other, then they will all receive high authority values under pSALSA (and SALSA). The algorithm of Kleinberg [5] instead provides a more complicated "mutually reinforcing" structure, but this makes it even more vulnerable to the TKC effect. See [1] for further discussion of this.

To proceed, we postulate an underlying (latent) cluster structure $\mathcal{C}$ on the set $\mathcal{N}$ of nodes, where nodes in the same cluster tend to have similar sets of outlinks. The authority of a node is then proportional to the number of *clusters* which tend to link to it, rather than to the number of *nodes* which link to it. Thus, if there are lots of near-identical nodes which all link to the same page, then this page only receives a slight authority boost from the many near-identical nodes. This helps to avoid the TKC effect, as we shall discuss.

3 The Similarity-Downweighting Algorithm

For a node i, let $\ell(i)$ be the set of nodes linked to from i. In terms of this, we define the *similarity matrix* by

$$S_{ij} \;=\; \frac{|\ell(i) \cap \ell(j)|}{|\ell(i) \cup \ell(j)|}\,, \qquad i, j \in \mathcal{N}\,. \tag{1}$$

That is, S_{ij} measures the fraction of the outlinks from *either* i or j, which are common to *both* i and j. (For definiteness, if $\ell(i)$ and $\ell(j)$ are both empty then we set $S_{ij} = 1$, though in fact this will not effect our algorithm at all.)

Now, if the cluster structure provided a *perfect* description of the link structure, then we would have $S_{ij} = 1$ if i and j are in the same cluster, and $S_{ij} = 0$ otherwise. Of course, in practice we

will often have $0 < S_{ij} < 1$. However, this heuristic suggests that, roughly, the size of the cluster containing node k can be estimated as

$$|C(k)| \;\approx\; \sum_i S_{ki}\,. \tag{2}$$

We now describe how to estimate the authority a_j of a node j. Recall that this authority should be proportional to the number of *clusters* which tend to link to j. Hence, intuitively, we should have

$$a_j \;=\; \sum_{k:\, j\in\ell(k)} |C(k)|^{-1}\,, \tag{3}$$

where $C(k)$ is the size of the cluster containing node k.

Combining the intuition of (2) and (3), and ignoring nodes which do *not* link to j, we obtain the "similarity-downweighting" (SD) estimate

$$\widehat{a}_j \;=\; \sum_{k:\, j\in\ell(k)} \frac{1}{\sum_{i:\, j\in\ell(i)} S_{ik}}\,,$$

with S_{ik} as in (1). It is this estimate that we shall use for the "SD" authority rankings.

We note that $\widehat{a}_j = 0$ if and only if there are no links to j. If there are any links to j, then since $S_{ik} \le 1$, we must have $\widehat{a}_j \ge 1$. However, it is possible to have $\widehat{a}_j \approx 1$ even if there are many links to j, if those links come from pages which all have very similar outlinks.

4 Sequential Clustering (SC): An Empirical Bayes Approach

The SD algorithm provides authority weights $\widehat{a_j}$, while only *implicitly* considering the actual cluster structure on the nodes.

As an alternative, we now present a method of generating an actual cluster structure on the nodes, and then using that to compute authority weights $\widehat{a_j}$.

4.1 The Set-Up

We model the situation as follows. We postulate an underlying (latent) cluster structure $\mathcal{C}$ on the set $\mathcal{N}$ of nodes. (Thus, $\mathcal{C}$ is a partition of $\mathcal{N}$ into some number of clusters.) We denote a cluster structure by $\mathcal{C}$, and a cluster within that structure by C.

We then postulate parameters d_{Cj} for $C \in \mathcal{C}$ and $j \in \mathcal{N}$ such that

$$P[A_{ij} = 1 \,|\, \mathcal{C}, \{d_{Cj}\}, i \in C_0] \;=\; d_{C_0 j}\,,$$

independently for each $i, j \in \mathcal{N}$. Intuitively, d_{Cj} represents the probability that any given node from cluster C will link to the node j.

The likelihood function is then given by

$$L(A_{ij} \,|\, \mathcal{C}, \{d_{Cj}\}) \;=\; \prod_{C\in\mathcal{C}} \prod_{i\in C} \prod_{j\in\mathcal{N}} d_{Cj}^{A_{ij}} (1-d_{Cj})^{1-A_{ij}} \;=\; \prod_{C\in\mathcal{C}} \prod_{j\in\mathcal{N}} d_{Cj}^{N(C,j)} (1-d_{Cj})^{|C|-N(C,j)}\,, \tag{4}$$

where $|C|$ is the number of nodes in C, and $N(C,j)$ is the number of nodes in C which link to node j.

4.2 Computing the Authority Values

We note that the *conditional* distribution for d_{Cj}, given the cluster structure $\mathcal{C}$, is equal to

$$\pi(d_{Cj} \,|\, \mathcal{C}) \;\propto\; \prod_{C \in \mathcal{C}} \prod_{j \in \mathcal{N}} d_{Cj}^{N(C,j)} (1 - d_{Cj})^{|C| - N(C,j)} \,.$$

We recognise this as a beta distribution. Hence, the conditional expected value of d_{Cj} given $\mathcal{C}$, is equal to

$$\mathbf{E}(d_{Cj} \,|\, \mathcal{C}) \;=\; \frac{N(C,j) + 1}{|C| + 2} \,.$$

In terms of this, we thus define the authority value of the node j to be equal to

$$a_j \;=\; |\mathcal{C}|^{-1} \sum_{C \in \mathcal{C}} \frac{N(C,j) + 1}{|C| + 2} \,. \tag{5}$$

This corresponds to the probability that, if we choose one of the $|\mathcal{C}|$ different clusters uniformly at random, that any given node from that cluster will link to j. That is, j is considered to be authoritative if it is linked to by a large number of *clusters*, rather than by a large number of individual nodes.

4.3 Generating the Cluster Structure

We see from (5) that we can compute the authority values a_j once we know the cluster structure $\mathcal{C}$. But how do we compute the cluster structure?

We propose the following sequential clustering algorithm.

1. Choose a node uniformly at random from $\mathcal{N}$, and assign it to a fresh cluster.

2. Given clusters $C_1, \ldots, C_r$, choose uniformly at random a node $k \in \mathcal{N}$ which has not yet been assigned to a cluster.

3. For $1 \leq j \leq r$, define $\mathrm{aff}[j] = \sum_{i \in C_j} S_{ik} \,/\, |C_j|$, the average similarity value of k to elements of C_j, to be the "affinity" of node k for the cluster C_j. [Here S_{ik} is defined in (1).]

4. Let $\mathrm{aff}[r+1] = 1$ be affinity of node k for a *new* cluster.

5. Assign node k to cluster C_j with probability $\mathrm{aff}[j] \,/\, \sum_{q=1}^{r+1} \mathrm{aff}[q]$, for $1 \leq j \leq r+1$, where C_{r+1} corresponds to a new cluster.

6. If there are still unassigned nodes, then go to step 2, otherwise stop.

This algorithm thus puts nodes into clusters in a random fashion, so that a node is more likely to join up with a cluster to which it has similar outlinks on average, but so that each node always has some probability of starting a new cluster. Note also that the nodes are considered in random order, so that the resulting assignment is "label-independent" as defined in [1]. On the other hand, since the clusters are assigned by a direct construction rather than by a conditional probability distribution, this algorithm is not fully Bayesian. Rather, it represents an "empirical Bayes" approach.

4.4 The Final SC Algorithm

Once the (random) cluster assignment $\mathcal{C}$ is made as above, then (random) authority weights a_j can be computed as in (5). The final authority weights, $\widehat{a_j}$, are then computed by repeating this procedure a number of times (say, 1000 times), and *averaging* the authority weightings so obtained, to largely eliminate the randomness of the result.

We shall see in the experiments below that the SC algorithm does a very good job of selecting authoritative nodes.

5 A Fully Bayesian Approach

In this section, we discuss the search for a fully Bayesian approach to analysing the latent cluster structure of World Wide Web pages. One problem with a fully Bayesian approach on a model with a variable number of parameters, is that it is usually necessary to use complex reversible jump MCMC procedures, and this is likely to be computationally infeasible on any large data set. Here we shall look at one particular approach in which the parameters can be analytically integrated out, so that the MCMC algorithm to be implemented is considerably simpler.

We again begin with a latent cluster structure $\mathcal{C}$ and parameters d_{Cj} as above, with the likelihood function $L(A_{ij} \,|\, \mathcal{C}, \{d_{Cj}\})$ again given by (4). But this time, we put a *prior* distribution on $\mathcal{C}$ and $\{d_{Cj}\}$, as follows.

5.1 Prior and Posterior Distributions

We let $\mathcal{R}$ be the set of all possible cluster structures on $\mathcal{N}$, so that $|\mathcal{R}|$ is the well-known *Ramanujan Number*. We take the prior on cluster space $\mathcal{R}$ to be proportional to $\kappa^{|\mathcal{C}|}$, where $|\mathcal{C}|$ is the total number of clusters. Here $\kappa > 0$ is a parameter which controls average number of clusters; if $\kappa = 1$ then this correspond to the uniform prior on $\mathcal{R}$. We take the i.i.d. uniform $[0, 1]$ prior for the $\{d_{Cj}\}$.

This model and prior lead to a posterior distribution $\pi(\cdot)$ on the state space

$$\mathcal{X} \equiv \{(\mathcal{C}, d) \;:\; \mathcal{C} \in \mathcal{R},\; d : \mathcal{C} \times \mathcal{N} \to \mathbf{R}\}\,,$$

having density with respect to counting measure on $\mathcal{R}$, crossed with $\mathbf{R}$ to the power of the number of clusters in $\mathcal{C}$, given by

$$\pi(\mathcal{C}, d) \;=\; \kappa^{|\mathcal{C}|} \prod_{C \in \mathcal{C}} \prod_{j \in \mathcal{N}} d_{Cj}^{N(C,j)} (1 - d_{Cj})^{|C| - N(C,j)}\,.$$

Now, recall that for integers $a, b \geq 0$,

$$\int_0^1 x^a (1-x)^b \, dx \;=\; \beta(a+1, b+1) \;\equiv\; \frac{\Gamma(a+1)\,\Gamma(b+1)}{\Gamma(a+b+2)} \;\equiv\; \frac{a!\,b!}{(a+b+1)!}\,.$$

Using this, we can further simplify our model by integrating out the d_{Cj} from 0 to 1, to obtain a final posterior distribution on cluster structures, given by

$$\pi(\mathcal{C}) \;=\; \prod_{C \in \mathcal{C}} \prod_{j \in \mathcal{N}} \frac{N(C,j)!\,\Big(|C| - N(C,j)\Big)!}{\Big(|C| + 1\Big)!}\,. \tag{6}$$

This shows that the distribution on cluster structures $\mathcal{C}$ can be considered separately from the parameters $\{d_{Cj}\}$.

5.2 Bayesian Authority Weights

As before, we have that the conditional expected value of d_{Cj} given $\mathcal{C}$, is equal to

$$\mathbf{E}(d_{Cj} \,|\, \mathcal{C}) \;=\; \frac{N(C,j)+1}{|C|+2}\,.$$

In terms of this, we again define the authority value of the node j to be equal to

$$a_j \;=\; |\mathcal{C}|^{-1} \sum_{C \in \mathcal{C}} \frac{N(C,j)+1}{|C|+2}\,.$$

This provides us with authority weights, *if* we can first generate cluster structures $\mathcal{C}$ having the right distribution as in (6). We consider that next.

5.3 Applying the Bayesian Approach: MCMC Algorithms

To experiment with this fully Bayesian algorithm, we need to be able to simulate cluster structures $\mathcal{C}$ according to the posterior (6). To do this, we consider a random-walk Metropolis MCMC algorithm on this model. Because of the above simplification, we need only run MCMC on the cluster structure $\mathcal{C}$ itself, to converge to the stationary distribution $\pi(\mathcal{C})$ as above.

We have experimented with a number of random-walk Metropolis algorithms for this model. We have also augmented the algorithms with split/merge moves (see e.g. [4]). Unfortunately, it appears that the resulting model is too sensitive to the prior distribution (for example the value of κ), and tends to converge either to just a single cluster (in which case our algorithm reduces to pSALSA) or to a huge number of unstable clusters which do not help to counteract the TKC effect. This problem is a well-known phenomenon in Bayesian statistics, linked to the so called *Lindley's paradox*, and is caused by the large difference in dimensionality between models with different numbers of clusters. Thus, we do not report any experiments with this algorithm here, but rather leave further development to subsequent work.

6 Experiments

We run experiments on the web data from [1], available (as the "expanded" root sets) at

`http://www.cs.toronto.edu/~tsap/experiments/datasets/`

In particular, we consider the problematical query topics "abortion" and "gun control". These query topics are interesting because they represent hotly debated public policy matters, and therefore each give rise to two different, isolated, highly inter-connected "communities", one "pro" and one "con" for each issue. They thus provide a good test of algorithms which deal with web community issues.

We present the top-ten results of our SD and SC rankings, and also (for comparison) the results of the Kleinberg algorithm [5] and the SALSA algorithm [6], as presented in [1].

6.1 "Abortion" query

This Base Set consists of 2293 nodes, with adjacency list available at

`http://www.cs.toronto.edu/~tsap/experiments/datasets/abortion/expanded/adj_list`

6.1.1 Results of Kleinberg algorithm on "Abortion" query

1.	http://www5.dimeclicks.com (web marketing page)
2.	http://www.amazon.com/exec/obidos/redirect-home/youdebatecom (amazon.com page)
3.	http://rd1.hitbox.com/rd?acct=WQ590703J6FB45EN5 (amazon.com page)
4.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/electronics/misc/top-sellers.html (amazon.com page)
5.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/electronics/software/home.html (amazon.com page)
6.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst /music/charts/hot-100-music.html (amazon.com page)
7.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst /home/gifts.html (amazon.com page)
8.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst /video/sellers/amazon-top-100-dvd.html (amazon.com page)
9.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst /video/sellers/amazon-top-100-video.html (amazon.com page)
10.	http://www.nrlc.org (National Right to Life Organization home page)

6.1.2 Results of SALSA algorithm on "Abortion" query

1.	http://www.nrlc.org (National Right to Life Organization home page)
2.	http://www.plannedparenthood.org (Planned Parenthood Federation of America)
3.	http://www.naral.org (The National Abortion and Reproductive Rights Action League)
4.	http://www5.dimeclicks.com (web marketing page)
5.	http://www.amazon.com/exec/obidos/redirect-home/youdebatecom (amazon.com page)
6.	http://rd1.hitbox.com/rd?acct=WQ590703J6FB45EN5 (amazon.com page)
7.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/electronics/misc/top-sellers.html (amazon.com page)
8.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/electronics/software/home.html (amazon.com page)
9.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/music/charts/hot-100-music.html (amazon.com page)
10.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/home/gifts.html (amazon.com page)

6.1.3 Results of Similarity-Downweighting (SD) algorithm on "Abortion" query

1.	http://www.nrlc.org (National Right to Life Organization home page)
2.	http://www.priestsforlife.org (Catholic pro-life page)
3.	http://www.serve.com/fem4life (Feminist pro-life page)
4.	http://www.youdebate.com (debates page, including abortion debates)
5.	http://www.hli.org (Human Life Interational; international pro-life page)
6.	http://home.about.com (about.com home page)
7.	http://www.prolife.org (pro-life resource page)
8.	http://members.aol.com/abtrbng (The Abortion Law Homepage)
9.	http://www.prolife.org/ultimate (pro-life resource page)
10.	http://www.pregnancycenters.org (anti-abortion pregnancy support site)

6.1.4 Results of SC algorithm on "Abortion" query

1.	http://www.nrlc.org (National Right to Life Organization home page)
2.	http://www.plannedparenthood.org (Planned Parenthood Federation of America)
3.	http://www.naral.org (The National Abortion and Reproductive Rights Action League)
4.	http://www.gynpages.com (Abortion Clinics OnLine)
5.	http://www.prolife.org/ultimate (pro-life resource page)
6.	http://www.priestsforlife.org (Catholic pro-life page)
7.	http://www.prochoice.org (NAF – The Voice of Abortion Providers)
8.	http://www.hli.org (Human Life Interational; international pro-life page)
9.	http://www.cais.com/agm/main (The Abortion Rights Activist Home Page)
10.	http://www.pregnancycenters.org (anti-abortion pregnancy support site)

6.1.5 Discussion of experimental results for "Abortion" query

For this query, we see that both Kleinberg and SALSA have major problems. Kleinberg gets trapped in the amazon.com community, reporting 8 of the top 10 "abortion" authorities as being amazon.com sites (with the top site being an irrelevant marketing site, and only the tenth site being on-topic). Similarly, SALSA reports 6 of the top 10 "abortion" authorities as being amazon.com sites, and it also reports the same irrelevant marketing site, though SALSA's first three sites are on-topic and do a good job of representing disparate communities.

By contrast, the Similarity-Downweighting (SD) algorithm reports only 1 page which is completely off-topic (the home.about.com page). The other 9 authorities are all related to abortion issues, showing a very high degree of page relevance.

Indeed, the irrelevant pages appear in the Kleinberg and SALSA results because they are linked to by a large number (106) of near-identical pages, which causes Kleinberg and SALSA to weight them highly. By contrast, the SD algorithm considers all such links to be essentially *one* link from a single large cluster. Hence, the SD weights of these irrelevant pages are just slightly over 1.0, in contrast to weights of between 7.7 and 9.5 for the top-ten authorities. That is, the SD algorithm correctly identifies these irrelevant pages as being very *low* in authority value.

On the other hand, the authorities chosen by the SD algorithm tend to favour just one side

of the debate (pro-life), suggesting poor mixing of different communities. In addition, the pages selected by SD are quite different from those selected by the other algorithms, and it is not clear (aside from the irrelevant pages) if this is a good thing or not.

The SC algorithm performs at least as well as the SD algorithm, and perhaps better since it includes important pro-choice pages (e.g."Planned Parenthood") which SD omits. In fact, the first three authorities under SC are identical to those under SALSA, suggesting that SD manages to extract the best features of SALSA while still avoiding the TKC effect.

Furthermore, since the SC algorithm provides complete cluster structure output (not shown here), one can use it to "see" how the TKL effect is avoided. Indeed, the 106 near-identical pages which all link to the same irrelevant pages, are usually all put into just two or three different clusters by SC, so that their importance is vastly diminished.

6.2 "Gun Control" Query

This Base Set consists of 2137 nodes, with adjacency list available at

`http://www.cs.toronto.edu/~tsap/experiments/datasets/gun_control/expanded/adj_list`

6.2.1 Results of Kleinberg algorithm on "Gun Control" query

1.	http://www.astrology.com (Astrology page)
2.	http://www.parentsoup.com (Parenting site)
3.	http://www.allhealth.com (Health Information page)
4.	http://www.ivillagemoneylife.com (Debt Management page)
5.	http://www.corporate-ir.net/ireye/ir_site.zhtml?ticker=IVIL&script=2100 (iVillage Corporate Profile)
6.	http://www.lamaze.com (Lamaze birth and parenting page)
7.	http://www.ivillage.com/help/privacy.html (iVillage: The Women's Network – Privacy Policy)
8.	http://www.ivillage.com/help/tos.html (iVillage Terms of Service)
9.	http://www.ivillage.com/sponsors/sites (iVillage Sponsor Directory)
10.	http://www.ivillage.com/beauty (iVillage beauty advice)

6.2.2 Results of SALSA algorithm on "Gun Control" query

1.	http://www.nra.org (National Rifle Association home page)
2.	http://www.gunowners.org (Gun Owners of America home page)
3.	http://www.saf.org (Second Amendment Foundation home page)
4.	http://www.jpfo.org (Jews for the Preservation of Firearms Ownership)
5.	http://www.handguncontrol.org (Handgun Control home page)
6.	http://www.amazon.com/exec/obidos/redirect-home/youdebatecom (amazon.com page)
7.	http://rd1.hitbox.com/rd?acct=WQ590703J6FB45EN5 (internet statistics page)
8.	http://www5.dimeclicks.com (marketing page)
9.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/electronics/misc/top-sellers.html (amazon.com page)
10.	http://www.amazon.com/exec/obidos/redirect?tag=youdebatecom&path= subst/music/charts/hot-100-music.html (amazon.com page)

6.2.3 Results of SD algorithm on "Gun Control" query

1.	http://www.youdebate.com (online debating page including Gun Control debate)
2.	http://www.vote-smart.org (American Voting Advice/Issues page)
3.	http://www.nra.org (National Rifle Association home page)
4.	http://www.2ndlawlib.org (Second Amendment Law Library)
5.	http://www.lpnn.com (Libertarian Party News Network)
6.	http://www.avidoutdoors.com (Camping Equipment and Outdoor Gear on-line store)
7.	http://www.sas-aim.org (Second Amendment Sisters, a female pro-gun group)
8.	http://www.guncite.com (GunCite: gun control and Second Amendment issues)
9.	http://www.gunowners.org (Gun Owners of America home page)
10.	http://www.jpfo.org (Jews for the Preservation of Firearms Ownership)

6.2.4 Results of SC algorithm on "Gun Control" query

1.	http://www.nra.org (National Rifle Association home page)
2.	http://www.gunowners.org (Gun Owners of America home page)
3.	http://www.saf.org (Second Amendment Foundation home page)
4.	http://www.jpfo.org (Jews for the Preservation of Firearms Ownership)
5.	http://www.handguncontrol.org (Handgun Control home page)
6.	http://www.2ndlawlib.org (Second Amendment Law Library)
7.	http://www.ruger-firearms.com (Ruger Firearms gun store)
8.	http://www.shooters.com (on-line gun store)
9.	http://www.vpc.org (The Violence Policy Center: Research, Analysis, and Advocacy for Effective Gun Policy)
10.	http://rkba.org (Arms Rights and Liberty Information on the Internet)

6.2.5 Discussion of experimental results for "Gun Control" query

For the "Gun Control" query, we see that the Kleinberg algorithm performs very poorly. Its first page is an astrology page, its next pages are variously about parenting and health or debt management, and five of its top-ten pages are irrelevant pages from the "iVillage" on-line women's

magazine (though, to be fair, that magazine has included some discussion of gun control in the past). Essentially, none of Kleinberg's top ten pages are on topic.

By contrast, the SALSA algorithm begins very well, with its first five pages being on topic and very relevant (with the first four being pro-gun and the fifth favouring gun control). However, its last five pages are variously from amazon.com or marketing/statistics pages which are completely off-topic. Hence, overall SALSA's performance on this query is also quite poor.

The SD algorithm does somewhat better. Aside from its sixth page, the other nine pages are all somewhat related to gun control issues. However, the first two are general issues/debating pages which have included some gun control issues, as opposed to being specifically about gun control, and the fifth is a general Libertarian page rather than being a specifically anti-gun-control page. Also, the other pages selected by SD are somewhat different from the five selected by SALSA, for example including the "Second Amendment Sisters" as opposed to the "Second Amendment Foundation", and it is not clear if this is better or worse.

Once again, the SC algorithm does even better. Indeed, aside from its seventh and eight pages (which are gun stores rather than about gun control per se), all of its pages appear deeply relevant to gun control. Its first five pages are identical to SALSA's, again suggesting that SC manages to extract the best features of SALSA while avoiding the TKC effect. And its ninth and tenth pages are highly relevant pages which were missed by the other three algorithms.

6.3 Other queries

We have also tested the SD algorithm on some of the other queries taken from [1].

On the "Computational Geometry" query, the SD algorithm performs very well, but so does the Kleinberg algorithm. (By contrast, the SALSA algorithm does not perform as well.)

On the "Computational Complexity" query, all three algorithms perform rather poorly, including a number of sites about computational *geometry* as opposed to complexity per se. SD does at least as well as Kleinberg and SALSA, but none do very well.

On the "Net Censorship" query, all three algorithms again perform rather poorly, including various irrelevant sites such as CNN.com etc. As discussed in [1], this may be because Net Censorship is not well represented on the World Wide Web, and/or because it is too intermixed with other related issues. SD again does at least as well as Kleinberg and SALSA, but none do very well.

7 Conclusion

In this paper, we have presented two new algorithms (SD and SC) which use World Wide Web link structures to determine authority values of web pages. Both algorithms postulate an underlying cluster structure, in an effort to avoid the Tightly Knit Community (TKC) effect prevelant in the Kleinberg and SALSA algorithms.

Our experiments suggest that the SD algorithm does a fairly good job of selecting authoritative web pages for a given query, given only the link structure of the Base Set of pages. Indeed, it appears to perform at least as well as Kleinberg and SALSA on every query, and to sometimes

perform considerably better. Most obviously, it does a better job of avoiding selecting sites which are entirely off-topic but have many in-links due to a large collection of similar or identical pages.

In this sense, we feel that the SD algorithm does a good job of using the latent cluster structure of the web pages, in order to appropriately downweight links from similar or identical web pages.

The SC algorithm also performs very well, and in fact appears to do even better at finding the relevant authority pages. Furthermore, the SC algorithm provides full underlying cluster structures, allowing for deeper understanding of how the authority values come about.

We have also considered a fully Bayesian approach to this problem, complete with MCMC algorithms to sample from the posterior distribution. However, we have found such approaches to be too sensitive to the details of the prior distribution used, and we leave refinements of this approach to future work.

In this paper, we have used the SD and SC algorithms solely to define better authority weights a_j. However, the underlying cluster structures provided by the SC algorithm could also be used in other ways. For example, the algorithm could present the top authority in each cluster, thus giving a "representative sample" of authoritative web pages with different perspectives. Or, the algorithm could examine the largest or most authoritative clusters, with a view towards breaking up the Base Set $\mathcal{N}$ into different nodes from different communities. We leave this to future work as well.

Acknowledgements. We are very grateful to Panayiotis Tsaparas for making the Base Set data from [1] available on the World Wide Web.

References

[1] A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *10th International World Wide Web Conference*, 2001. Full version available at `http://probability.ca/jeff/research.html`.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *7th International World Wide Web Conference*, Brisbane, Australia, 1998.

[3] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. Preprint, 2000.

[4] S. Jain and R. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, to appear, 2000.

[5] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM (JASM)*, 46, 1999.

[6] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In *9th International World Wide Web Conference*, May 2000. Full version published in Computer Networks **33** (2000), 387–401.

[7] D. Rafiei and A. Mendelzon. What is this page known for? Computing web page reputations. In *9th International World Wide Web Conference*, Amsterdam, Netherlands, 2000.