Web Communities Identification from Random Walks

Jiayuan Huang[1,2], Tingshao Zhu[2], and Dale Schuurmans[2]

[1] University of Waterloo, Waterloo, Canada
[2] University of Alberta, Edmonton, Canada

Abstract. We propose a technique for identifying latent Web communities based solely on the hyperlink structure of the WWW, via random walks. Although the topology of the Directed Web Graph encodes important information about the content of individual Web pages, it also reveals useful meta-level information about user communities. Random walk models are capable of propagating local link information throughout the Web Graph, which can be used to reveal hidden global relationships between different regions of the graph. Variations of these random walk models are shown to be effective at identifying latent Web communities and revealing link topology. To efficiently extract these communities from the stationary distribution defined by a random walk, we exploit a computationally efficient form of directed spectral clustering. The performance of our approach is evaluated in real Web applications, where the method is shown to effectively identify latent Web communities based on link topology only.

1 Introduction

Increasingly, the World Wide Web is playing an important role in peoples' lives as a main destination for information. However, the sheer heterogeneity of Web users and authors—given diverse backgrounds and interests—hampers traditional information retrieval approaches that rely on content analysis alone. The Web is comprised of multiple communities [5] created by different groups of people having common interests. The identification of Web communities can help users with their information retrieval goals, by allowing the construction of preclassified directories and the creation of more effective recommendation services. Random walk models have been successfully used for Web ranking in the past [12,11], and have also raised interest in identifying Web communities.

In this paper we investigate how the *directed* hyperlink information conveyed via random walks can help one efficiently identify latent Web communities from the hyperlink topology alone. Our work on identifying Web communities exploits recent progress on *directed* spectral clustering [16], and contributes further understanding to the nature of such clustering techniques. Here, we analyze directed spectral clustering from a random walk perspective.

Intuitively, a coherent Web community can be identified by a subset of Web pages that is strongly connected within the subset, while only being weakly connected with pages outside the subset. We assume that if two pages are directly

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 187–198, 2006.

linked, their interests are assumed to be somewhat related [7]. We also take *co-citation* [15] and *co-reference* [9] relations into account to accurately identify latent Web communities. Such high level connections provide useful relationship information, since sometimes connections between Web pages might not be as obvious as such direct links.

For Web community identification, we first examine a one-step random walk model that captures low level aspects of hyperlink connectivity. We then consider a two-step random walk model with different variations that captures higher level information by exploiting the existence of co-reference and co-citation relationships in the link topology. For the two-step model, we introduce a damping process that samples the entire Web uniformly, which allows the random walk to be properly applied to general Web graphs.The selection of the damping factor is essential to both computation and performance. Finally, we examine the performance of different random walk models and damping factors in identifying Web communities from pure graph topology. The empirical results demonstrate that our random walk models are sufficiently flexible to capture different levels of relationships in the link topology to achieve significant performance in Web community identification. This provides a practical understanding how various random models behave in Web community identification.

2 Background

Before analyzing our specific models, we briefly review related work on identifying Web communities in general.

The problem of identifying Web communities is clearly related to the more fundamental problem of graph partitioning. For general graph partitioning, one can often resort to straightforward principles such as unrestricted minimal cut, or the dual principle of maximum flow. However, the graphs used by most such techniques are undirected, and therefore they ignore the directionality information encoded in Web hyperlinks [4,8]. Another simple approach is to extract similarity measurements between neighboring vertices (Web pages) directly from the link structure to perform a generic clustering method [9]. However, the similarity should be measured from the *global* structure of the graph. A more global approach to Web graph clustering suggests, therefore, that some sort of aggregate similarity measure be used, such as those based on the spectrum of the connectivity matrix. For *undirected* graph clustering, a common suggestion is to partition by performing a singular value decomposition (SVD) on W [13]. However again, the connectivity matrix W is *not* symmetric.

By considering the directed links of Web pages, Kleinberg showed that the HITS ranking algorithm [10] converges to a spectral method that uses the principle eigenvectors of $W^T W$ and WW^T—the final weight scores for the authorities and hubs. Later, it was observed that this technique can in fact be used to identify web communities, where Web pages with highest authority and hub scores are used to define the core of a community [6]. However, one can see that this approach reduces to SVD on an *undirected* graph weight matrices WW^T and

$W^T W$. In fact, this approach suffers from two drawbacks: first, a straightforward graph partition method based on simply computing the principle eigenvectors is not very effective in general; and second, the directed hyperlink information is significantly diminished through the symmetric transformations. Regarding the first drawback, a more appropriate way to solve the graph partitioning problem is to consider it as a *balanced* minimum cut problem, which usually results in more accurate clusters being obtained. Although most versions of the balanced minimum cut are NP-complete, the eigenvectors of graph Laplacians [2] provide a good approximation to this NP-hard problem. The efficiency and effectiveness of such *balanced* spectral clustering methods has been demonstrated in many domains, e.g. [14]. Unfortunately, these methods have only been developed for undirected graphs, and do not consider directionality information.

To address these shortcomings, we require a balanced spectral clustering principle that can take into account the directionality of Web hyperlinks. Recently, a new approach to directed graph clustering has been proposed in [16], which offers a mathematically clean solution to this problem. It minimizes a balanced cut criterion for directed graphs that has a very natural interpretation in a random walk framework. Unfortunately, the work presented in [16] does not address the specific role of random walks in Web graph clustering. A Web graph differs from a general directed graph in that it posesses particular topological properties. It is therefore critical to formulate a proper random walk model that ensures similar pages are grouped into coherent Web communities. In this paper, we analyze two random walk models with their variants that are sufficiently flexible to capture important aspects of Web graph topology, and disclose how walk connectivity is related to page similarity in directed spectral clustering. Below we investigate the performance of these random walk models in comparison with standard models of spectral clustering on undirected graphs [6].

3 Directed Spectral Clustering

To identify Web communities in a Directed Web Graph we employ the efficient spectral clustering technique for directed graphs developed in [16]. The criterion for directed graph partitioning is given by a combinatorial partition criterion that generalizes the normalized cut criterion for undirected graphs [14]. It requires no transformation of the asymmetric adjacency matrix into a symmetric one.

A directed graph $G = (V, E)$ can be associated with a Markov chain defined by a random walk on the graph. The stationary distribution π of this random walk gives a probability of occupancy over a vertex v given infinite time. So given a subset S of vertices in G, we define the probability with which the random walk occupies vertices in S as $P(S) = \sum_{v \in S} \pi(v)$. Let S^c denote the complement of S. Obviously, $P(S) + P(S^c) = 1$. Define the probability with which the random walk jumps to S^c from S as $P(S \to S^c) = \sum_{u \in S, v \in S^c} \pi(u) p(u, v)$. We then consider partitioning the directed graph G into two nonempty subsets S and S^c by minimizing the following

$$\text{cut}(S) = \frac{P(S \to S^c)}{P(S)} + \frac{P(S^c \to S)}{P(S^c)} \tag{1}$$

Intuitively, a good partitioning of a directed graph under this criterion corresponds to a cut such that the probability of escaping from one community to another is small, whereas the probability of remaining in the current community is high. Note that these escape and retention probabilities are measured with respect to a long run of the random walk, so the optimal partition is determined by the global link topology of the graph. Minimizing this objective is NP-hard [2] but an approximation can be efficiently obtained by solving for the eigenvectors of a directed graph Laplacian Δ defined as follows [16]. Let Π denote the diagonal matrix with $\Pi(v,v) = \pi(v)$ for all $v \in V$. Let P denote the transition probability matrix and P^T the transpose of P. Then define

$$\Theta = \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2}}{2}. \tag{2}$$

Then, $\Delta = I - \Theta$, where I denotes the identity. The directed spectral clustering algorithm is then to compute the eigenvector Φ of Θ corresponding to the second largest eigenvalue, and then partition the vertex set V of G into two parts according to the sign. In practice, for multiple clusters, it is standard to detect and visualize clusters based on the sorted eigenvalues [1].

4 Random Walks on Digraphs

The random walk model is a free parameter in the directed spectral clustering framework outlined above. Technically, the only requirement is that the transition probabilities of the random walk satisfies the *balance equation* $\pi(v) = \sum_{u \to v} \pi(u) p(u,v)$ where $u \to v \in E$ denotes page v is pointed by page u. However, for the purposes of identifying latent Web communities in a Directed Web Graph, we need to specify an appropriate random walk to ensure that tightly coupled Web pages share a common topic or interest. This provides a practical understanding of different behavior of random walk models in Web clustering.

One-Step Random Walk

The one-step random walk model we examine initially is the *teleporting random walk* model of [12]. Given that the random surfer is currently at a vertex u: (a) with probability ϵ it chooses an outlink uniformly at random and follows the link to the next page; or (b) with probability $1 - \epsilon$ it jumps to a Web page uniformly at random over the entire Web (excluding itself). Here, a damping factor ϵ ($0 < \epsilon < 1$) is introduced in the case where the current page has no outlink. Such a random walk is guaranteed to converge to a unique stationary distribution which can be computed by numerically solving the balance equation. The transition probability $p_{tele}(u,v)$ between u and v under this model can be written as $p_{tele}(u,v) = \epsilon \frac{w(u,v)}{d^+(u)} + p_\epsilon(u,v)$, where $p_\epsilon(u,v) = w(u,v)/\operatorname{vol} G$ if $d^+(u) = 0$ and $p_\epsilon(u,v) = (1-\epsilon) w(u,v)/\operatorname{vol} G$ if $d^+(u) > 0$; $\operatorname{vol} G = \sum_u (d^+(u) +$

$d^-(u)$). Here $w(u,v)$ is the weight value along each edge; $d^-(v) = \sum_{u \to v} w(u,v)$ and $d^+(v) = \sum_{u \leftarrow v} w(v,u)$ are the *in-degree* and *out-degree* of v.

This random walk makes the simple assumption that similar pages are directly linked. The stationary probability of a Web page corresponds to the frequency that a surfer visits the page following forward links. This can be viewed as an authority effect in the Web page ranking. We refer to this random walk as the one-step authority model (***OneStepA***). Conversely, we can consider another random walk that traverses *backward* along the hyperlinks [3]. This is equivalent to the hub effect, since a good hub page should be able to visit many other related pages. Therefore, we refer to this random walk as the one-step hub model (***OneStepH***).

Two-Step Random Walk

Web pages are "connected" by more than their direct hyperlinks. Intuitively, commonality between two Web pages is revealed by the presence of common co-citation or co-reference pages. The random walk we employ should therefore also consider these implicit connections in Web community identification.

We now consider a *two-step random walk* model motivated by the Hubs and Authorities model in [10]. Assume temporarily that each Web page has inlinks and outlinks. Then, starting from a page u, the random surfer first jumps backward to an adjacent hub vertex h with probability $p^-(u,h) = w(h,u)/d^-(u)$, then it jumps forward to a page v adjacent from h with probability $p^+(h,v) = w(h,v)/d^+(h)$. Then the two-step transition probability $p^A(u,v)$ between two authorities u and v is given by

$$p^A(u,v) = \sum_h p^-(u,h)p^+(h,v) \tag{3}$$

The stationary distribution π^A of this random walk is $\pi^A(u) = d^-(u)/\operatorname{vol} G^-$ where $\operatorname{vol} G^- = \sum_{u \in V} d^-(u)$. This follows from the fact that

$$\begin{aligned}\sum_{u \in V} \pi^A(u)p^A(u,v) &= \sum_{u \in V} \frac{d^-(u)}{\operatorname{vol} G^-} \sum_{h \in V} \frac{w(h,u)w(h,v)}{d^-(u)d^+(h)} \\ &= \frac{1}{\operatorname{vol} G^-} \sum_{h \in V} \frac{w(h,v)}{d^+(h)} \sum_{u \in V} w(h,u) = \frac{d^-(v)}{\operatorname{vol} G^-} = \pi^A(v)\end{aligned}$$

This random walk is performed by treating pages as authorities.

Using the same argument, we can define a two-step random walk by treating pages as hubs. The random walk performs among hubs u and v by first taking a forward step and then a backward step along the edges $u \to a$ and $a \leftarrow v$, yielding the transition probability between hubs

$$p^H(u,v) = \sum_a p^+(u,a)p^-(a,v) \tag{4}$$

Similarly, this random walk between hubs has the stationary distribution $\pi^H(u) = d^+(u)/\operatorname{vol} G^+$. The two-step random walk exploits the co-citation and co-reference effects in the high level Web link topology. The assumption here is that two similar pages should share more common hubs or authorities.[1]

The above two-step random walks require that each Web page has inlinks and outlinks, but this is not always true for real Web graphs. To be able to handle the general case, we propose to combine the two-step random walk with a teleporting step, so that each forward and backward step through an outlink and a inlink has a damping factor. Therefore, to obtain the mixed two-step random walk, simply plug the modified transition probabilities p^- and p^+ into formulas (3) and (4) to modify p^A and p^H among authorities and hubs. In our experiments below we only use the mixed version of the two-step random walks, but for simplicity we just refer to them as ***TwoStepA*** and ***TwoStepH*** respectively. Finally, we consider a convex combination of the two types of two-step random walks that address the hyperlink structure in a more flexible manner $P = \beta P^A + (1-\beta)P^H$, where β is a tuning parameter that controls the different weights of co-citation and co-reference effects. The advantage of this combination is that it can help us determine which effect is dominant in the link structure, based on the results. Or conversely, given some prior knowledge about the levels of link structure, we can set a proper value for β that consistently matches the hyperlink topology.

Spectral Clustering with Random Walks

To partition a Directed Web Graph, we can simply use the adjacency matrix A with unit weights (i.e., $a(u, v) = 1$ when $u \to v$). It is interesting to compare the results of the different random walk models and the symmetrized transformation models in this case. To demonstrate the differences in a simple toy example, we computed the second eigenvectors of Θ formulated as in (2) for both the one-step and two-step random walks on the graph in Figure 1. We set $\epsilon = 0.95$. We also obtain the principal eigenvectors of $A^T A$ and AA^T, corresponding to the symmetrized authority and hub scores mentioned in the Background section [10]. We refer to these symmetrized methods as ***Auth*** and ***Hub*** respectively.

One can partition the directed graph into two clusters by examining the values in the eigenvector thresholding at zero. Pages within an initial grouping can then be partitioned further after the first partitioning [1], and so on. In addition to just partitioning the vertices, however, the eigenvector values can also be used to assign a *weight* or *confidence* that each Web page belongs to its assigned cluster. That is, the greater the eigenvector value at a page, the more likely the page is to belong to the given cluster. We will therefore refer to these values as the *weights* of pages below. We visualize the partitioning by assigning each vertex on a solid line as shown in Figure 1.

[1] We briefly note that [11] uses the stationary distribution proportional to vertex in-degrees to perform a simple ranking method and showed similar derivations of stationary distributions.

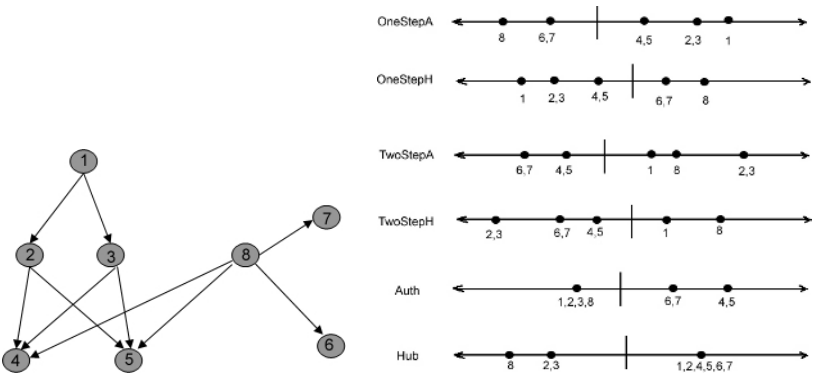

Fig. 1. Left: A toy example of a directed graph. Right: Illustrating partitioning by sorted values. Here, "|" indicates the threshold value (zero) such that vertices on each sides are grouped into separate clusters.

In the toy example, the partitions are the same for OneStepA and OneStepH, which tend to extract highly correlated clusters via direct connections. Moreover the vertices that have large values (e.g., 1 and 8) are also the vertices that have the highest stationary distributions under the random walks. It is known that PageRank ranks Web pages by their stationary distribution, but pages with high stationary probabilities might be of dissimilar topics. However, besides clustering, this method can provide rearranged rankings within each cluster which is very useful to current search engines.

TwoStepA tends to group strong authorities (4, 5, 6, 7) together that are linked by common pages. TwoStepH extracts the hub vertices (1, 8) that link to similar vertices directly and/or indirectly, e.g., vertex 1 points to vertices 4 and 5 after passing 2 and 3. Vertex 8 points to vertices 4 and 5 directly. This random walk tends to group good hubs that link to common pages either implicitly or explicitly. The partition using the symmetrized authority score is similar to TwoStepA, but it does not distinguish among the vertices 1, 2, 3 and 8. The partition using the symmetrized hub score also ignores any differences among the vertices in each group, and is thereby less meaningful.

Random walks are able to effectively capture the differences between direct hyperlink and indirect second order hyperlink topologies that have different co-citation and co-reference patterns in directed spectral clustering. All of these can be exploited to efficiently identify vertex communities via directed spectral clustering.

Table 1. Web graphs statistics

Root queries	vertex num	edge num
1. "waterloo"	2130	4688
2. "movies"+"olympics"	6634	65536
3. "risk analysis"+"bussiness optimization"	3357	10490
4. "differential geometry"+ "parallel computing"	2575	6844
5. "data mining"+"computer vision"	3907	12416
6. "body arts"+"fashion design"	3091	4122

5 Empirical Results

5.1 Experimental Design

We construct Web graphs of varying degrees of difficulty by either building the graph from a single topic query, which results in multiple topics that can be hard to distinguish, or building the graph from multiple queries, which results in a few more easily distinguishable topics. To obtain Web graphs, we first chose some *root queries*, submited these to Google, and retrieved the first t html pages (not including pdf or ps files). For a given query or set of queries, we then combined the retrieved pages as *roots* and perform a one level expansion by adding pages that are linked from or link to the root pages. Finally, we filtered out non-informative links that exist among Web pages as follows. We restrict the number of pages that link to or are pointed to by every root URL to be at most d pages. This operation was first proposed in [10]. We also filter out all *cgi scripts* links. We set t and d equal to 100 and 50 respectively. The collections we finally obtained were relatively sparse graphs. In our experiments, we use several groups of root queries. Their statistics are listed in Table 1. The root queries focus on a variety of interests. Pages retrieved from queries that have significant overlap intuitively should increase the difficulty of Web page clustering.

5.2 Results

Choosing Parameters. Practically, two parameters need to be selected when defining the random walks on the Web graphs: the damping factor ϵ in the one-step and two-step random walks, and the tuning parameter β in the two-step random walks. We test with 2 root queries using the damping factor ϵ set to 0.75, 0.85 and 0.95. Clustering performance is evaluated by counting the correctly classified pages that have the 30 greatest weights among those ranked within top 100 by Google.

Figures 2 plot the confusion matrix values corresponding to the numbers of pages among the 30 with the greatest weight that are classified as "movie" (class 1) and "olympics" (class 2). Ideally, the best result should have corresponding numbers of 30, 0, 0, 30. Since OneStepA and OneStepH give very similar results in this experiment, we only show the results of OneStepA. One can see from these figures that the directed spectral method with OneStepA obtains the best performance when ϵ equals 0.85. Thus, we fix this value for OneStepA in later experiments. For TwoStepA, the results are competitive when ϵ takes value 0.85 and 0.95. Since each result has a better performance for one of the communities, we choose $\epsilon = 0.90$ as an compromise value in the following experiments.

Next, we consider the tuning parameter β that balances between P^A and P^H in the two-step random walk. Figure 3 (left) shows the results when β changes from 1 to 0 in the "movies+olympic" Web graph. Instead of reporting the confusion matrix values in detail, we summarize it by the F *measure*, which can be derived from the confusion matrix as $\frac{2(precision \times recall)}{(precision + recall)}$ where $precision = C_{11}/(C_{11} + C_{21})$ and $recall = C_{11}/(C_{11} + C_{12})$. The Figure shows that the best

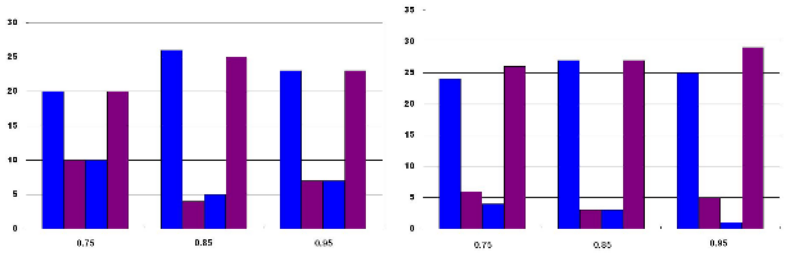

Fig. 2. OneStepA results(left) and TwoStepA results(right). Plot of confusion matrix values $C_{11}, C_{12}, C_{21}, C_{22}$(from left to right of each column block) for $\epsilon = 0.75, 0.85, 0.95$.

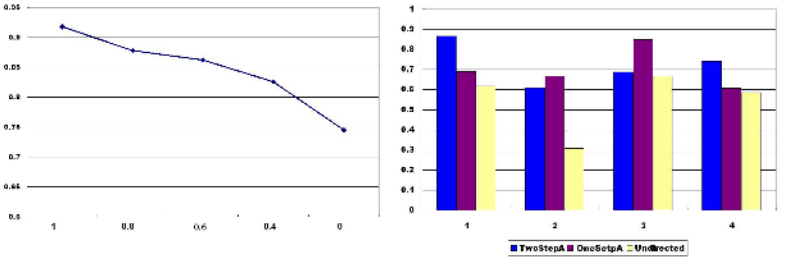

Fig. 3. Left: F scores when β changes in two-step random walk, $\epsilon = 0.90$. Right: F score for 4 binary clustering tasks. Blue: TwoStepA, Red: OneStepA, Yellow: Undirected.

performance is obtained when $\beta = 1$. This means that the Web page similarities are most correctly assessed when the transition matrix is P^A for this Web graph. Not surprisingly, this result is consistent with the ranking methods that consider inlink degree and authority scores from $A^T A$ [6,11]. These studies have already shown that important pages can be found by evaluating their authority scores only. Thus, we set $\beta = 1$ in our following experiments, although one should point out that this is not a globally optimal choice.

Single Broad-Topic Query. Table 2 lists the communities detected by the directed spectral method using OneStepA. For each community, we list the URLs with significant PageRanks. We can visulize that using only hyperlink structure, one can still identify reasonable communities from a Web graph constructed by a single broad topic query. The weights of pages in clusters 1 to 4 are closer to each other than to the pages in other clusters. This discloses that the first 4 clusters are related within a broader scope: they are mainly pages from Waterloo, Canada, including academic institutions, social communities and living. The observation that the weights of clusters 5 and 6 are closer to each other than to the others identifies they are the pages of Waterloo locales in the US. The sub-topics are generalized upward to larger common topics. Cluster 9 identifies the pages from Wikipedia, even though we eliminated links among pages from the same domain.

Table 2. Communities from query "waterloo"

Cluster 1: Pages from universities and schools at Waterloo, Canada	Cluster 2: Pages for the public community service in Waterloo, Canada
www.uwaterloo.ca/ www.wlu.ca/ www.lib.uwaterloo.ca/ www.math.uwaterloo.ca/ www.cs.uwaterloo.ca/ www.wcdsb.edu.on.ca/	www.city.waterloo.on.ca/ www.waterloorecords.com/ www.therecord.com/ www.wpl.ca/ www.wrps.on.ca/ www.oktoberfest.ca/
Cluster 3: Pages for living at Waterloo, Canada	Cluster 4: Pages for life at Waterloo, Canada
www.waterlooinn.com/ www.waterloochamber.org/ www.kwhumane.com/ www.kwsymphony.on.ca/	www.kwymca.org/ www.waterloo.ca/ www.kwag.on.ca/ www.uptownwaterloojazz.ca/ www.kwsc.org/ www.waterloo-biofilter.com/ www.wnhydro.com/
Cluster 5: Pages for Waterloo, Iowa, USA	Cluster 6: Pages for Waterloo in the USA
www.wplwloo.lib.ia.us/waterloo/ www.wcfsymphony.org/ www.waterloocvb.org/ www.waterlooindustries.com/	www.waterloobucks.com/ www.waterloo.k12.ia.us/ www.waterloo.il.us/ www.waterlooindustries.com/
Cluster 7: Pages for Waterloo in Europe	Clusters 8 and 9: Pages for the history of Waterloo from public pages and from wiki
www.trabel.com/waterloo/ waterloo-thebattle.htm/ www.waterloo.org.uk/ www.trabel.com/waterloo/waterloo.htm/ www.napoleonguide.com/ battle_waterloo.htm/ www.waterloo.co.uk/	www.garywill.com/waterloo/ history.htm/ www.bbc.co.uk/history/war/ trafalgar_waterloo/ en.wikipedia.org/wiki/ Battle_of_Waterloo/ en.wikipedia.org/wiki/Waterloo_station/

Multiple Topic Related Queries. We also evaluated clustering performance for 4 Web graphs that are obtained from multiple root queries. We compare the directed spectral methods using one-step random walk and two-step random walks to the undirected method that uses the symmetrized authority scores from A^TA (referred to as the undirected method in the results). This undirected method is more efficient than performing SVD on A^TA in undirected graph clustering which is essentially the method in [6] as been explained in Background. Therefore our comparison is more challenging.

Figure 3–Right shows the clustering results for 4 Web graphs obtained from root queries 3, 4, 5 and 6. Not surprisingly, both of the directed spectral methods outperformed the undirected method in all cases.

Table 3. Pages with the top 10 significant weights for Queries of "computer vision" + "data mining"

Directed spectral method with OneStepA		**Undirected method.**	
URL	Cat	URL	Cat
cmp.felk.cvut.cz/eccv2004/	1	dms.irb.hr/index.php	2
iris.usc.edu/Vision-Notes/bibliography/contents.html	1	www.comp.leeds.ac.uk/vision/	2
www.intel.com/research/mrl/research/opencv/	1	www.comp.leeds.ac.uk/vision/	1
marathon.csee.usf.edu/	1	www.statsoft.com/textbook/stdatmin.html	2
vis-www.cs.umass.edu/	1	lear.inrialpes.fr/people/triggs/events/iccv03/	1
www.cs.cmu.edu/ cil/vision.html	1	dir.groups.yahoo.com/group/datamining2/	2
www.sciencedirect.com/science/journal/10773142	1	www.acv.ac.at/	1
www.cs.cmu.edu/ cil/v-source.html	1	www-ai.ijs.si/SasoDzeroski/RDMBook/	2
iris.usc.edu/Information/Iris-Conferences.html	1	www.autonlab.org/tutorials/	2
homepages.inf.ed.ac.uk/rbf/CVonline/	1	www.cs.columbia.edu/ sal/hpapers/USENIX/ usenix.html	2
itmanagement.webopedia.com/TERM/D/ data_mining.html	2	www.scd.ucar.edu/hps/GROUPS/dm/dm.html	2
www.ncdm.uic.edu/	2	www.kdnuggets.com/	2
www.kdnuggets.com/	2	www.spss.com/	2
www.dmg.org/	2	www.eco.utexas.edu/ norman/BUS.FOR/course.mat/ Alex/	2
www.salforddatamining.com/	2	www.acm.org/sigkdd/	2
www.spss.com/	2	www.infogoal.com/dmc/dmcdwh.htm	2
www.acm.org/sigkdd/	2	www.the-data-mine.com/	2
www.megaputer.com/	2	www.thearling.com/text/dmwhite/dmwhite.htm	2
www.cacs.louisiana.edu/ icdm05/	2	www.ncdm.uic.edu/	2

Table 4. Pages with top 15 significant weights for Queries "movies"+ "olympics"

Directed spectral method with TwoStepA		Undirected method	
URL	Cat	URL	Cat
www.saltlake2002.com/	1	www.fhw.gr/olympics/ancient/	1
www.specialolympics.org/	1	cityguide.aol.com/main.adp	1
www.olympic.org/	1	www.dallasnews.com/sharedcontent/dws/spt/olympics/vitindex.html	1
www.torino2006.it	1	www.baltimoresun.com/sports/olympics/	1
sports.espn.go.com/oly/index	1	www.latimes.com/sports/olympics/	1
www.athens2004.com/athens2004/	1	diveintomark.org/howto/ipod-dvd-ripping-guide/	2
www.perseus.tufts.edu/Olympics/	1	movies.nytimes.com/pages/movies/	2
www.perseus.tufts.edu/Olympics/sports.html	1	news.bbc.co.uk/sport1/hi/other_sports/olympics_2012/default.stm	1
news.bbc.co.uk/sport1/hi/olympics_2004/default.stm	1	www.austin360.com/movies/content/movies/	2
www.nbcolympics.com/	1	www.musicfromthemovies.com/default.asp	2
www.olympics.com.au/	1	sports.yahoo.com/olympics	1
www.fhw.gr/projects/olympics/	1	movies.yahoo.com/mv/upcoming/	2
www.london2012.org/	1	www.fairolympics.org/en/	2
en.beijing-2008.org/	1	cbs.sportsline.com/u/olympics/2002/	1
www.imdb.com/	2	www.imdb.com/	2
us.imdb.com/	2	us.imdb.com/	2
www.imdb.com/search	2	rogerebert.suntimes.com/	2
movies.go.com/	2	www.lordoftherings.net/	2
www.usatoday.com/life/movies/front.htm	2	www.allmovie.com/	2
movies.aol.com/	2	www.rottentomatoes.com/	2
movies.yahoo.com/	2	www.infonegocio.com/xeron/bruno/olympics.html	1
movies.guide.real.com	2	www.brainpop.com/	2
www.rottentomatoes.com/	2	www.foxmovies.com/	2
www.hollywood.com/	2	www.hollywood.com/	2
www.boxofficemojo.com/	2	www.reel.com/	2
www.movieflix.com/	2	www.perseus.tufts.edu/Olympics/	1
www.ifilm.com/	2	www.ucmp.berkeley.edu/geology/tectonics.html	1

We also show some of the clustering results by listing the highly ranked URLs with the most significant weights in corresponding communities in Tables 3 and 4."Cat" denotes the ture category for each URL. Once again, we can see that the directed spectral methods work better than the undirected method by tending to group pages more correctly. For example, in Table 3, the pages correctly clustered in the data mining community are about major conferences, term explanations, and companies in data mining. In Table 4, we see in the olympics community, multiple homepages from the olympic game hosts were obtained. Although these pages do not have hyperlinks between them, they all are pointed to by the Olympic Games organization (olympic.org). Thus, the two-step random walk was able to detect their similarity by identifying a common hub. Similar observations can be made about the pages classified in the movies community. In each of these tasks, the undirected method failed to identify pages from same communities, and tended to mix pages from the different communities.

6 Conclusion

To automatically identify Web communities from hyperlink topology, we addressed a key component in directed spectral clustering: the random walk model that should be used to infer relationships between Web pages. In addition to one-step random walks, we also proposed variations of two-step random walk models that can detect higher order similarities between pages. The linear combination of two-step random walks suggests a practical approach to inferring the relationship between link structure and topic similarity by inspecting the clustering results. The experiments show that the different random walk models

can capture different relationships based on the hyperlink topology in directed spectral method.

Acknowledgments

Thanks to Dengyong Zhou for helpful discussions. Work supported by the Alberta Ingenuity Centre for Machine Learning, NSERC, and MITACS.

References

1. S. Chakrabarti, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Spectral filtering for resource discovery. In *ACM SIGIR workshop on Hypertext Information Retrieval on the Web*, 1998.
2. F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.
3. C. Ding, X. He, P. Husbands, H. Zha, and H. Simon. PageRank, HITS and a unified framework for link analysis. Technical report, LBNL, 2002.
4. G. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *ACM SIGKDD*, 2000.
5. G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 2002.
6. D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, 1998.
7. M. Henzinger. Hyperlink analysis for the web. In *IEEE Internet Computing*, 2001.
8. H. Ino, M. Kudo, and A. Nakamura. Partitioning of web graphs by community topology. In *WWW*, 2005.
9. M. Kessler. Bibliographic coupling between scientific papers. In *American Documentation*, 1963.
10. J. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 1999.
11. R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. In *WWW*, pages 387–401, 2000.
12. L. Page, S. Brin, R. Motwani, and T.Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
13. P. Perona and W. Freeman. A factorization approach to grouping. In *ECCV*, pages 655–670, 1998.
14. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE PAMI*, 2000.
15. H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS*, 1973.
16. D. Zhou, J. Huang, and B. Scholkopf. Learning from labeled and unlabeled data on a directed graph. In *ICML*, 2005.