

Co-Ranking Authors and Documents in a Heterogeneous Network

Ding Zhou¹ Sergey A. Orshanskiy² Hongyuan Zha³ C. Lee Giles⁴

Computer Science and Engineering¹
Department of Mathematics²
Information Sciences and Technology⁴
The Pennsylvania State University,
University Park, PA 16802

College of Computing³
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

Recent graph-theoretic approaches have demonstrated remarkable successes for ranking networked entities, but most of their applications are limited to homogeneous networks such as the network of citations between publications. This paper proposes a novel method for co-ranking authors and their publications using several networks: the social network connecting the authors, the citation network connecting the publications, as well as the authorship network that ties the previous two together. The new co-ranking framework is based on coupling two random walks, that separately rank authors and documents following the PageRank paradigm. As a result, improved rankings of documents and their authors depend on each other in a mutually reinforcing way, thus taking advantage of the additional information implicit in the heterogeneous network of authors and documents.

1. Introduction

Quantitative evaluation of researchers' contributions has become an increasingly important topic since the late 80's due to its practical importance for making decisions concerning matters of appointment, promotion and funding. As a result, bibliometric indicators such as citation counts and different versions of the *Journal Impact Factor* [3] are being widely used, although it is a subject of much controversy [11]. Accordingly, new metrics are constantly being proposed and questioned, leading to ever-increasing research efforts on bibliometrics [5]. These simple counting metrics are attractive, because it is convenient to have a single number that is easy to interpret. However, it has become evident in recent research that the evaluation of the scientific output of individuals can be performed better by consider-

ing the network structures among the entities in question (e.g. [9]).

Recently, a great amount of research has been concerned with ranking networked entities, such as social actors or Web pages, to infer and quantify their relative importance, given the network structure. Several *centrality* measures have been proposed for that purpose [2, 8, 10]. For example, a journal can be considered influential if it is cited by many other journals, especially if those journals are influential, too. Ranking networked documents received a lot of attention, particularly because of its applications to search engines. (e.g. PageRank [2], HITS [8]). Ranking social network actors, on the other hand, is employed for exploring scientific collaboration networks, understanding terrorist networks, ranking scientific conferences and mining customer networks for efficient viral marketing. While centrality measures are finding their way into traditional bibliometrics, let us point out that the evaluations of the relative importance of networked documents have been carried *independently*, in the similar studies, from social network actors, where the natural connection between researchers and their publications *authorship* and the social network among researchers are not fully leveraged.

This paper proposes a framework for co-ranking entities of different kinds in a heterogeneous network connecting the researchers (*authors*) and publications they produce (*documents*). The heterogeneous network is comprised of G_A , a social network connecting authors, G_D , the citation network connecting documents, and G_{AD} , the bipartite authorship network that ties the previous two together.

We propose a co-ranking method in a heterogeneous network by coupling two random walks on G_A and G_D using the authorship information in G_{AD} . We assume that there is a mutually reinforcing relationship between authors and documents that could be reflected in the rankings. In par-

ticular, the more influential an author is, the more likely his documents will be well-received. Meanwhile, well-known documents bring more acknowledgments to their authors than those that are less cited. While it is possible to come up with a ranking of authors based solely on a social network and obtain interesting and meaningful results [9], these results are inherently limited, because they include no direct consideration neither of the number of publications of a given author (encoded in the authorship network) nor of their impact (reflected in the citation network).

The contributions of this paper include: (1) A new framework for co-ranking entities of two types in a heterogeneous network is introduced; (2) The framework is adapted to ranking authors and documents: a more flexible definition of the social network connecting authors is used and random walks that are part of the framework are appropriately designed for this particular application; (3) Empirical evaluations have been performed on a part of the CiteSeer data set allowing to compare co-ranking with several existing metrics. Obtained results suggest that co-ranking is successful in grasping the mutually reinforcing relationship, therefore making the rankings of authors and documents depend on each other.

2 Co-Ranking Framework

2.1 Notations and preliminaries

Denote the heterogeneous graph of authors and documents as $G = (V, E) = (V_A \cup V_D, E_A \cup E_D \cup E_{AD})$. There are three graphs (networks) in question. $G_A = (V_A, E_A)$ is the unweighted undirected graph (social network) of authors. V_A is the set of authors, while E_A is the set of bidirectional edges, representing social ties. The number of authors $n_A = |V_A|$ and authors are denoted as $a_i, a_j, \dots \in V_A$. $G_D = (V_D, E_D)$ is the unweighted directed graph (citation network) of documents, where V_D is the document set, E_D is the set of links, representing citations between documents. The number of documents $n_D = |V_D|$. Individual documents are denoted as $d_i, d_j, \dots \in V_D$. $G_{AD} = (V_{AD}, E_{AD})$ is the unweighted bipartite graph representing authorship. $V_{AD} = V_A \cup V_D$. Edges in E_{AD} connect each document with all of its authors.

The framework we propose includes three *random walks*, one on G_A , one on G_D and one on G_{AD} . We shall start from two random walks, described by stochastic matrices A and D , and then slightly alter them in § 2.2 to actually obtain \tilde{A} and \tilde{D} . Both of them are called *Intra-class random walks*, because they walk either within the authors' or the documents' network. The third random walk on G_{AD} is called the *Inter-class random walk*. It will suffice to describe it by an $n_A \times n_D$ matrix AD and an $n_D \times n_A$ matrix DA , since G_{AD} is bipartite. The design of A , D , AD and DA is in § 3.

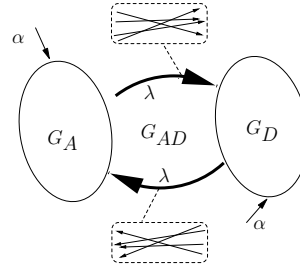


Figure 1. The framework for co-ranking authors and documents. G_A is the social network of authors. G_D is the citation network of documents. G_{AD} is the authorship network. α is the jump probability for the Intra-class random walks. λ is a parameter for coupling the random walks, quantifying the importance of G_{AD} versus that of G_A and G_D .

We briefly sketch the conceptual scheme of our co-ranking framework in Fig. 1. Two Intra-class random walks incorporate the *jump probability* α , which has the similar meaning to the damping factor as used in PageRank. They are coupled using the Inter-class random walk on the bipartite authorship graph G_{AD} . The coupling is regulated by λ . In the extreme case $\lambda = 0$ there is no coupling; this amounts to separately ranking authors and documents by PageRank. In general, λ represents the extent to which we want the rankings of documents and their authors depend on each other¹

2.2 PageRank: two random walks

First of all, let us rank the networks of authors and documents independently, according to the PageRank paradigm [2]. Consider a random walk on the author network G_A and let A be the transition matrix (A will be defined in § 3). Fix some α as the damping factor, we introduce another random walk (corresponding to an ergodic Markov chain) with the transition matrix

$$\tilde{A} = (1 - \alpha)A + \frac{\alpha}{n_A} \mathbf{1}\mathbf{1}^T \quad (1)$$

where $\mathbf{1}$ is all-one vector of size n_A . Let $\mathbf{a} \in \mathbf{R}^{n_A}$, $\|\mathbf{a}\|_1 = 1$ be the only solution of the equation $\mathbf{a} = \tilde{A}^T \mathbf{a}$. Here Vector \mathbf{a} contains the ranking scores for the vertices in G_A . It is a standard fact that the existence and uniqueness of the solution of the ranking vector \mathbf{a} which follows from the random walk \tilde{A} being ergodic. Similarly, documents can be ranked

¹This is a symmetric setting of parameters. An asymmetric setting of parameters can introduce $\alpha_A \neq \alpha_D$ and $\lambda_{AD} \neq \lambda_{DA}$. We do not expect that different α can make any difference. We do expect that different λ can make a difference, but we did not investigate that.

in the citation network G_D . In particular, the transition matrix is setup as:

$$\tilde{D} = (1 - \alpha)D + \frac{\alpha}{n_D} \mathbf{1}\mathbf{1}^T, \quad (2)$$

Next, we show how to combine the two random walks in authors and documents, which are two graphs of different types of vertices.

2.3 (m, n, k, λ) -coupling of random walks

To couple the two random walks defined before, we construct a combined random walk on the heterogeneous graph $G = G_A \cup G_D \cup G_{AD}$. A probability distribution on such a graph will have the form (\mathbf{a}, \mathbf{d}) , satisfying $\|\mathbf{a}\|_1 + \|\mathbf{d}\|_1 = 1$. We will use the stationary probabilities of the vertices in V_A to rank authors and the stationary probabilities of the vertices in V_D to rank documents. The coupling is parameterized by four parameters, m, n, k and λ .

Consider the the combined random walk in terms of a random surfer (RS) who is capable of browsing over documents and their authors as well. If at any given moment RS finds himself on the author side, the current vertex $v \in V_A$, he can either make an *Intra-class step* (one step of the random walk parameterized by \tilde{A}) or an *Inter-class step* — one step of the Inter-class random walk. Similarly, if RS finds himself on the document side, the current vertex $v \in V_D$, then one option is to make an *Intra-class step* (one step of the random walk parameterized by \tilde{D}) while another option is to make one step of the Inter-class random walk. In general, one Intra-class step changes the probability distribution from $(\mathbf{a}, \mathbf{0})$ to $(\tilde{A}\mathbf{a}, \mathbf{0})$ or from $(\mathbf{0}, \mathbf{d})$ to $(\mathbf{0}, \tilde{D}\mathbf{d})$, while one Inter-class step changes the probability distribution from (\mathbf{a}, \mathbf{d}) to $(DA^T\mathbf{d}, AD^T\mathbf{a})$.

More precisely, define the combined random walk as follows:

1. If the current state of RS is an author, $v \in V_A$, then with probability λ take $2k + 1$ Inter-class steps, while with probability $1 - \lambda$ take m Intra-class steps on G_A .
2. If the current state of RS is a document, $v \in V_D$, then with probability λ take $2k + 1$ Inter-class steps, while with probability $1 - \lambda$ take n Intra-class steps on G_D .

First, we show a subroutine *BiWalk* (Algo. 1) that takes \mathbf{x} , the probability distribution on one side of a bipartite graph and returns the distribution on the other side after taking $2k + 1$ Inter-class steps. U is the transition matrix from the current side to the other and V is the transition matrix from the other side back to the current side.

Next, we present co-ranking in *CoupleWalk* (Algo. 2). The returned \mathbf{a}, \mathbf{d} are the ranking vectors for authors and documents. It should be noted that the very recent work [6] of learning on subgraphs can be considered an implicit special version of our algorithm with infinite k and $m = n = 1$.

Algorithm 1 Random walk on a Bipartite Graph

procedure *BiWalk*(U, V, \mathbf{x}, k)

- 1: $\mathbf{c} \leftarrow \mathbf{x}$
 - 2: **for** $i = 1$ to k **do**
 - 3: $\mathbf{b} \leftarrow U^T \mathbf{c}$
 - 4: $\mathbf{c} \leftarrow V^T \mathbf{b}$
 - 5: **end for**
 - 6: $\mathbf{b} \leftarrow U^T \mathbf{c}$
 - 7: **return** \mathbf{b}
-

Algorithm 2 Coupling random walks for co-ranking

procedure *CoupleWalk*($\tilde{A}, \tilde{D}, AD, DA, m, n, k, \lambda, \epsilon$)

- 1: $\mathbf{a} \leftarrow \frac{1}{n_A} \mathbf{1}, \mathbf{d} \leftarrow \frac{1}{n_D} \mathbf{1}$
 - 2: **repeat**
 - 3: $\mathbf{a}' \leftarrow \mathbf{a}, \mathbf{d}' \leftarrow \mathbf{d}$
 - 4: $\mathbf{a} \leftarrow (1 - \lambda)(\tilde{A}^T)^m \mathbf{a}' + \lambda \text{BiWalk}(DA, AD, \mathbf{d}', k)$
 - 5: $\mathbf{d} \leftarrow (1 - \lambda)(\tilde{D}^T)^n \mathbf{d}' + \lambda \text{BiWalk}(AD, DA, \mathbf{a}', k)$
 - 6: **until** $\|\mathbf{a} - \mathbf{a}'\| \leq \epsilon$
 - 7: **return** \mathbf{a}, \mathbf{d}
-

Finally, we show that the Algo. 2 converges. To see this, observe that $\text{BiWalk}(U, V, \mathbf{x}, k) = U^T(V^T U^T)^k \mathbf{x}$. Therefore, lines 4 and 5 in Algo. 2 can be rewritten as:

$$\mathbf{a}^{t+1} = (1 - \lambda)(\tilde{A}^T)^m \mathbf{a}^t + \lambda DA^T (AD^T DA^T)^k \mathbf{d}^t \quad (3)$$

$$\mathbf{d}^{t+1} = (1 - \lambda)(\tilde{D}^T)^n \mathbf{d}^t + \lambda AD^T (DA^T AD^T)^k \mathbf{a}^t \quad (4)$$

where \mathbf{a}^t and \mathbf{d}^t are the ranking vectors for authors and documents from the previous iteration; m, n are prescribed parameters. Now we concatenate \mathbf{a} and \mathbf{d} into a vector \mathbf{v} such that $\mathbf{v} = [\mathbf{a}^T, \mathbf{d}^T]^T$. In particular, $\mathbf{v}^t = [(\mathbf{a}^t)^T, (\mathbf{d}^t)^T]^T$, is composed of \mathbf{a} and \mathbf{d} as in Algo. 2 after t iterations. Construct a matrix M , where

$$M = \begin{bmatrix} (1 - \lambda)(\tilde{A}^T)^m & \lambda DA^T (AD^T DA^T)^k \\ \lambda AD^T (DA^T AD^T)^k & (1 - \lambda)(\tilde{D}^T)^n \end{bmatrix}. \quad (5)$$

Clearly, $\mathbf{v}^{t+1} = M\mathbf{v}^t$, and M is a stochastic matrix that parameterizes the combined random walk. When $0 < \alpha, \lambda < 1$, this Markov Chain is ergodic. Thus, the stationary probabilities can be found as $\lim_{n \rightarrow +\infty} M^n \mathbf{v}$, for any initial vector \mathbf{v} .

3. Random Walks in a Scientific Repository

This section sets up the co-ranking framework to be applied to co-ranking scientists and their publications. It includes defining three networks and the three corresponding random walks, parameterized by four stochastic matrices: A (giving rise to \tilde{A}), D (giving rise to \tilde{D}), AD and DA .

3.1 G_D : document citation network, and D : the Intra-class random walk on G_D

The design of D is straightforward. Namely, the Intra-class random walk on G_D is just a simple random walk on it. The transition probability is $P(j|i) = D_{i,j} = \frac{n_{i,j}^D}{n_i^D}$, where $n_{i,j}^D$ is the indicator of whether document i cites j ; n_i^D is the total number of citations document i makes. If a document does not cite anything (which effectively means that the citations of this documents are not in the corpus), let the transition probabilities from this document be $\frac{1}{n_D}$.

3.2 G_A : author social network, and A : the Intra-class random walk on G_A

We define the social network among authors using a newly introduced notion called *social event*. A social event could be any kind of activity, involving a group of authors. In particular, we view collaborating on a paper or co-participating in a conference as such social events.

We start by giving a weighted graph G_A . Let the set of social events be $\mathcal{E} = \{e_k\}$, where an event e_k contains a set of participating authors. Define the social tie function $\tau(i, j, e_k) : \mathcal{A} \times \mathcal{A} \times \mathcal{E} \rightarrow [0, 1]$ representing the strength of a social tie between actor a_i and actor a_j resulting from their co-occurrence in the event e_k : $\tau(i, j, e_k) = \frac{\mathbb{I}(i, j \in e_k)}{|e_k|(|e_k|+1)/2}$, where $\mathbb{I}(i, j \in e_k)$ is the indicator function of whether authors i and j co-occur in the event e_k (that is, if $a_i \in e_k$ and $a_j \in e_k$; it can be that $a_i = a_j$). Adding up social ties inferred from all events, we obtain a cumulative matrix $T = (T_{i,j}) \in \mathbb{R}^{n_A \times n_A}$, by definition: $T_{i,j} = \sum_{e_k \in \mathcal{E}} \tau(i, j, e_k)$, where \mathcal{E} is the set of social events. Now G_A can be viewed as a weighted graph, with the weight on the edge connecting a_i and a_j being $T_{i,j}$. In this paper, treat the different kinds of events similarly thanks to the definition of $\tau(i, j, e_k)$.

We proceed to define the Intra-class random walk on G_A in a natural way. Technically, it amounts to normalizing T by rows. The transition probabilities from author a_i to author a_j (i.e. of the author a_j given a_i) can then be found as: $P(j|i) = A_{i,j} = \frac{T_{i,j}}{\sum_j T_{i,j}}$. Here T is symmetric due to the design of τ . A is not necessarily symmetric because row sums can be different. \tilde{A} is defined accordingly.

3.3 G_{AD} : the bipartite authorship network, and AD, DA : the Inter-class random walk on G_{AD}

We first define the bipartite authorship graph G_{AD} is defined in the natural way: $E_{AD}(i, j) = \mathbb{I}(d_j \text{ is authored by } a_i)$. Using the adjacency matrix E_{AD} , we define a weight matrix $W_{AD} = (w(i, j))$ as follows: $w(i, j) = \frac{E_{AD}(i,j)}{n_j^A}$, where n_j^A is the number of authors of the document d_j .

Then we proceed to define AD and DA , containing the conditional transition probabilities of a random surfer moving from author i to document j and vice versa, respectively, given that the next step is taken in the bipartite graph G_{AD} . That is, let $P(d_j|a_i) = AD_{i,j} = \frac{w(i,j)}{\sum_k w(i,k)}$, $P(a_i|d_j) = DA_{j,i} = \frac{w(i,j)}{\sum_k w(k,i)}$.

This completes the descriptions of networks and random walks. Note that the design of W_{AD} implies $\sum_k w(k, j) = 1$. The matrices AD and DA are asymmetric to reflect the asymmetric relationship between authors and documents.

4 Experiments

4.1 Data Preparation

For experiments, we use data from CiteSeer [4], which currently has a collection of over 739,135 scientific documents in Computer Sciences. The documents have 418,809 distinct authors after name disambiguation. Since the data in CiteSeer are collected automatically by crawling the Web, we may not have enough information about certain authors. Accordingly, we concentrate on the subset of those authors who have at least five co-authored publications and all documents that have at least one author from this selected subset. Presumably, this gives us a more informative sample including 7,488 authors and 182,662 documents from 1991 to 2004. We performed a fuzzy matching of the titles of CiteSeer documents with the document titles from DBLP data to extract document proceedings.

We further categorize the documents into topics using the Latent Dirichlet Allocation (LDA) model [1] with the desired number of topics set to $T = 50$. We selected five topics that are well-represented in the database: T6: stochastic and Markov processes, T8: WWW and information retrieval, T19: learning and classification, T36: statistical learning, and T48: data management. All experiments were carried out for each of these five topics.

4.2 Author Rankings

We apply a two-step heuristic that further reduces the problem scale. Once the topic is fixed, we sort all authors by their accumulated topic weights. Then we choose a subset of top authors and all their documents, and re-rank them. This is similar to the approach used by search engines: take a subset of pages with large in-degrees and rank them by PageRank.

To evaluate the co-ranking approach, we perform a ranking of authors in each topic t by the methods listed below:

- **Publication count**, the number of papers (on the topic t) an author has in the document subset;
- **Topic weight**, the sum of topic weights of all documents, produced or co-authored by an author;

cs-id	title	authors	year	cite
116523	The Well-Founded Semantics for General Logic Programs	Allen Van Gelder, Kenneth A. Ross, John S. Schlipf	1991	312
25887	Mining Association Rules between Sets of Items in Large Databases	Rakesh Agrawal, Tomasz Imielinski, Arun Swami	1993	921
35061	Answering Queries Using Views	Alon Levy, Alberto Mendelzon, Yehoshua Sagiv, et.al.	1995	296
440364	Competitive Paging Algorithms	Amos Fiat, Richard M. Karp, Michael Luby, et.al.	1991	147
70633	Efficient Similarity Search In Sequence Databases	Rakesh Agrawal, Christos Faloutsos, Arun Swami	1993	205
229795	On The Power Of Languages For The Manipulation Of Complex Objects	Serge Abiteboul, Catriel Beeri	1993	129
24123	Implementing Data Cubes Efficiently	Venky Harinarayan, Anand Rajaraman, Jeffrey Ullman	1996	248
6606	The Design Of Postgres	Michael Stonebraker, Lawrence Rowe	1986	152
142235	Objects and Views	Serge Abiteboul, Anthony Bonner	1991	196
118598	Database Mining: A Performance Perspective	Rakesh Agrawal, Tomasz Imielinski, Arun Swami	1993	100
16843	An Interval Classifier for Database Mining Applications	Rakesh Agrawal, Sakti Ghosh, Tomasz Imielinski, et.al.	1992	95
88311	Querying Semi-Structured Data	Serge Abiteboul	1997	373
84227	Object Exchange Across Heterogeneous Information Sources	Yannis P., Hector Garcia-Molina, Jennifer Widom	1995	316
65646	Mediators in the Architecture of Future Information Systems	Gio Wiederhold	1992	460
9685	The Object-Oriented Database System Manifesto	M. Atkinson, Francois Bancilhon, David DeWitt, et.al.	1989	298

Table 1. Top documents in the topic *data management*

r	author names	con#	r	p#	r	cite#	r
1	Rakesh Agrawal	171	44	129	32	1915	1
2	Serge Abiteboul	209	12	115	42	1300	3
3	Jennifer Widom	234	5	113	44	1617	2
4	Jiawei Han	271	2	142	22	720	10
5	Hector Garcia-Molina	232	7	169	16	1247	4
6	Ian Foster	142	79	215	12	513	19
7	Azer Bestavro	97	198	174	14	354	42
8	Deborah Estrin	134	100	186	13	471	23
9	Subbarao Kambhampati	118	130	275	8	173	132
10	Michael Stonebraker	59	322	144	21	299	66
11	Christos Faloutsos	218	11	98	58	770	9
12	Moshe Y. Vardi	184	29	148	20	415	30
13	Rajeev Motwani	145	75	127	33	579	15
14	Richard T. Snodgrass	125	115	68	131	330	50
15	Joseph Hellerstein	63	305	75	103	132	208

Table 2. Top authors in the topic *data management* when $m = 2, n = 2, k = 1$. **con#**: number of neighbors in the social network; **p#** : number of papers; **cite#** : number of citations; **r**: ranks by the corresponding methods.

- **Number of citations**, the total number of citations to the documents of an author from the other documents on the same topic;
- **PageRank in the social network**, ranking by PageRank on the graph G_A , constructed as outlined in § 3;
- **Co-Ranking**, co-ranking authors and documents by the new method.

The parameter values we used in the Co-Ranking framework are $m = 2, n = 2, k = 1, \lambda = 0.2, \alpha = 0.1$. For different settings of m, n, k the top 20 authors and papers varied slightly, even less for different α .

We used a well-known metric, the Discounted Cumulated Gain (DCG) [7], in order to compare the five different rankings of authors. Top 20 authors according to each ranking (publication count, etc.) are merged in a single list, shuf-

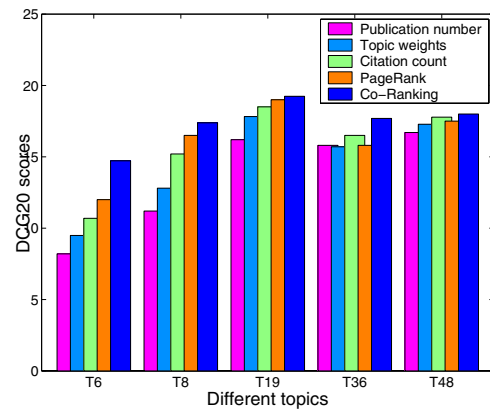


Figure 2. DCG₂₀ scores for author rankings: number of papers, topic weights, number of citations, PageRank, and Co-Ranking.

fled and submitted for judgment. Two human judges, one an author of this paper and the other one from outside, provide feedback. Numerical assessment scores of 0, 1, 2, and 3 are collected to reflect the judges' opinion with regard to whether an author is ranked top 20 in a certain field, which respectively means *strongly disagree*, *disagree*, *agree*, and *strongly agree*, with the fact that these authors are ranked top 20 in the corresponding field. As suggested, assessments were carried out based on professional achievement of the authors such as winning of prestigious awards, being a fellowship of ACM/IEEE, etc. The judges' assessment scores are averaged. We observe a high agreement between the two judges.

The DCG₂₀ scores obtained are presented in Fig. 2. The figure shows five groups of bars corresponding to five topics. This evaluation shows that the new co-ranking method outperforms the other four ranking methods, achieving an

average improvement of 27.8%, 19.1%, 10.6%, and 7.7% over rankings by the number of papers, the topic weights, the number of citations, and the PageRank.

We list the top 15 authors ordered by the Co-Ranking scores for the topic *data management* only due to space limit. Note that in the top author lists, we observe a mix of famous scientists from different fields. This is due to the imperfect automatic categorization performed by LDA; manual categorization labels can be used instead.

4.3 Document Rankings

For each topic, we obtained the Co-Ranking scores for the documents. For comparison, we also found the number of citations to each document within the same document subset. Table 1 presents the top documents according to Co-Ranking in the topic *data management*. More information regarding a document can be found at the URL `http://citeseer.ist.psu.edu/x` where x are the cs-id.

The quality of ranking documents is hard to quantify, there are few objective criteria to rely on, and also domain-specific knowledge is required for an assessment. We did not produce any judgment on the document rankings we obtained due to the above concerns.

4.4 Parameter Effect

We ran Co-Ranking on 50 synthetic datasets with various settings of m , n , k , λ , and α and arrived at the following conclusions: (1) Large λ introduces more mutual dependence of the rankings between authors and documents. In particular, as λ increases, the ranking of authors becomes closer to the ranking by the number of publications; (2) In case of large α such as 0.5, the ranking of authors becomes more uniform, so that the documents of productive authors are neglected, and also generally benefiting the documents with many authors. Since both effects are undesirable, keep α small; (3) For small m , especially $m = 1$, the weight of edges in G_A is not fully taken into account, but only the local differences in weights matter; (4) Prevent large k . It completely eliminates the effect of authors on documents and vice versa, except for the authorship information: the bipartite random walk forgets everything, as expected from a Markov chain after many steps; (5) For small n , the structure of the citation network is less important, making the Co-Ranking more like a citation counting.

The computational complexity of Algo. 1 is $O(k \times n_A \times n_D)$. The complexity of Algo. 2 is $O(t \times n_A \times n_D \times (n + m + 2k + 1))$, where n , m , k are parameters and t is the number of steps before convergence.

5. Conclusions and Future Research

This paper proposes a new link analysis ranking approach for co-ranking authors and documents respectively

in their social and citation networks. Starting from the PageRank paradigm as applied to both networks, the new method is based on coupling two random walks into a combined one, presumably exploiting the mutually reinforcing relationship between documents and their authors: good documents are written by reputable authors and vice versa. Experiments on a real world data set suggest that Co-Ranking is more satisfactory than counting the number publications or the total number of citations a given scientist has received. Also, it appears competitive with the PageRank algorithm as applied to the social network only. Possible directions of future research are: (1) A larger empirical evaluation could be carried out; (2) A formal analysis of the properties of the new Co-Ranking framework is required, including the effect of parameters m , n , k , λ on the ranking results. We expect there to be interesting interconnections with the HITS algorithm; (3) Other ways shall be explored for coupling random walks other than the one suggested in this paper. Several possibilities have been deemed unsatisfactory, however, presumably, the (m, n, k, λ) - setting does not exhaust all meaningful ways to do that.

References

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.
- [2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web* 7, pages 107–117. Elsevier Science Publishers B. V., 1998.
- [3] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(60):471–479, November 1972.
- [4] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: an automatic citation indexing system. In *DL '98: Proceedings of the third ACM conference on Digital libraries*, pages 89–98, 1998.
- [5] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102:16569, 2005.
- [6] J. Huang, T. Zhu, R. Greiner, D. Zhou, and D. Schuurmans. Information marginalization on subgraphs. In *PKDD*, pages 199–210, 2006.
- [7] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 41–48, 2000.
- [8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.
- [9] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *arXiv.org:cs/0502056*, 2005.
- [10] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- [11] P. Weingart. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1):117–131, 2005.