Co-Ranking Authors and Documents in a Heterogeneous Network*

Ding Zhou[1] Sergey A. Orshanskiy[2] Hongyuan Zha[3] C. Lee Giles[4]

Computer Science and Engineering[1]
Department of Mathematics[2]
Information Sciences and Technology[4]
The Pennsylvania State University,
University Park, PA 16802

College of Computing[3]
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

The problem of evaluating scientific publications and their authors is important, and as such has attracted increasing attention. Recent graph-theoretic ranking approaches have demonstrated remarkable successes, 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. The proposed ranking approach has been tested using data collected from CiteSeer, and demonstrates a great improvement in author ranking quality compared with ranking by the number of publications, the number of citations and the PageRank calculated in the authors' social network.

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* [8, 14] are being widely used, although it is a subject of much controversy [22]. Accordingly, new metrics are constantly being proposed and questioned, leading to ever-increasing research efforts on bibliometrics [10, 14]. 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 considering the network structures among the entities in question (e.g. [19, 15]).

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 [5, 13, 21]. 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 [5], HITS [13]). Ranking social network actors, on the other hand, is employed for exploring scientific collaboration networks [23], understanding terrorist networks [16, 23], ranking scientific conferences [19] and mining customer networks for efficient viral marketing [7]. 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 connect-

*Accepted at IEEE ICDM 2007

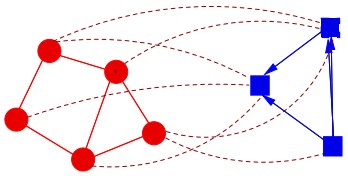


Figure 1. Three networks we use for co-ranking: a social network connecting authors, the citation network connecting documents, and the co-authorship network that ties the two together. Circles represent authors, rectangles represent documents.

ing 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. Further details will be given in § 3. A simple example of a such a heterogeneous network is shown in Fig. 1.

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 particular, 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 [15], 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.

We start from reviewing related work in § 2. We propose the new framework in § 3. We demonstrate the convergence of the ranking scores in § 4. We explain how we set up the framework in § 5. We present experimental results and give some comments in § 6 and conclude this work in § 7.

2. Related Work

The problem of ranking scientists and their work naturally belongs to at least two different fields: sociology [21] and bibliometrics [20]. An important step in bibliometrics was a paper by Garfield [8] in the early 70's, discussing the methods for ranking journals by Impact Factor. Within a few years, Gabriel Pinski and Francis Narin proposed several improvements [17]. Most importantly, they recognized that citations from a more prestigious journal should be given a higher weight [17]. They introduced a recursively defined weight for each journal. In particular, incoming citations from more authoritative journals, according to the weights computed during the previous iteration, contributed more weight to the journal being cited. Pinski and Narin stated it as an eigenvalue problem and applied to 103 journals in physics. However, their approach did not attract enough attention, so that simpler measures have remained in use.

It was 25 years later when Brin and Page, working on Google, applied a very similar method named PageRank to rank Web pages [5]. Independently, Kleinberg proposed the HITS algorithm [13], also intended for designing search systems, which is similar to PageRank in its spirit but used a mutual reinforcement principle. Since then, numerous papers on link analysis-based ranking have appeared, typically taking HITS or PageRank as the starting point (e.g. [1, 4]). There are several good introduction papers to the field (e.g. [4]). The mutual reinforcement principle has also been applied to text summarization and other natural language processing problems [24].

The Co-Ranking framework presented in this paper is another method based on PageRank and the mutual reinforcement principle, with its new focus on heterogeneous networks. Variations of PageRank have already been applied in many contexts. For example, Bollen et al. [3] ranked journals in their citation network, essentially by PageRank. They presented an interesting empirical comparison of this ranking with the ISI Impact Factor on journals in Physics, Computer Science, and Medicine. Their results clearly support that the Impact Factor measures popularity while the PageRank measures prestige. Another empirical study [6] ranked papers in Physics by PageRank. It turns out that famous but not so highly cited papers are ranked very high. Yet another study by Liu et al. focused on co-authorship networks [15]. They compared the rankings of scientists by PageRank and its natural variation with three other rankings by degree, betweenness centrality and closeness centrality. A recent work also looks into random walks for learning on the subgraph its relation with the complement of it [11]. Nevertheless, given all that, we are not aware of any attempts to correlate the rankings of two dif-

ferent kinds of entities included in a single heterogeneous network.

3 Co-Ranking Framework

3.1 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 $a_i, a_j, \cdots \in V_A$. $G_D = (V_D, E_D)$ 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 $d_i, d_j, \cdots \in V_D$. $G_{AD} = (V_{AD}, E_{AD})$ 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 includes three *random walks*, one on G_A, one on G_D and one on G_{AD}. A random walk on a graph is a Markov chain, its states being the vertices of the graph. It can be described by a square $n \times n$ matrix M, where n is the number of vertices in the graph. M prescribes the transition probabilities. That is, $0 \leq p(i, j) = M_{i,j} \leq 1$ is the conditional probability that the next state will be vertex j, given that the current state is vertex i. If there is no edge from vertex i to vertex j then $M_{i,j} = 0$, with the exception when there are no outgoing edges from vertex i at all. In that case we assume that $M_{i,j} = \frac{1}{n}$ for all vertices j. By definition, M is a *stochastic matrix*, i.e. its entries are nonnegative and every row adds up to one. A *simple random walk* on a graph goes equi-probably to any of the current vertex' neighbors.

In this paper, "Markov chain" and "random walk" are used interchangeably to mean "time-homogeneous finite state-space Markov chain". Unless otherwise stated, all Markov chains in question are ergodic, that is, irreducible and aperiodic. A *probability distribution* is a vector $\mathbf{v}$ with one entry for each vertex in the graph underlying a random walk, such that all its entries are nonnegative and add up to one, $\|\mathbf{v}\|_1 = 1$. After one step of a random walk, described by a stochastic matrix M, the probability distribution will be $M^T\mathbf{v}$, where M^T is the transpose of M. A *stationary probability distribution* $\mathbf{v_{st}} = \lim_{n\to\infty}(M^T)^n\mathbf{v}$ contains the limiting probabilities after a large number of steps of the random walk. It is a common convention that the PageRank ranking vector $\mathbf{r}$ satisfies $\|\mathbf{r}\|_1 = 1$, naturally, since $\mathbf{r}$ is a probability distribution. The co-ranking framework will produce two ranking vectors, $\mathbf{a}$ for authors and $\mathbf{d}$ for documents, also satisfying

$$\forall 1 \leq i \leq n_A, 1 \leq j \leq n_D, \;\; a_i, d_j \geq 0; \tag{1}$$

$$\|\mathbf{a}\|_1 = 1, \|\mathbf{d}\|_1 = 1 \tag{2}$$

As mentioned above, we will have three random walks. The random walk on G_A (respectively, G_D) will be described by a stochastic matrix A (respectively, D). We shall start from two random walks, described by stochastic matrices A and D, and then slightly alter them in § 3.2 to actually obtain $\widetilde{A}$ and $\widetilde{D}$. All 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 postponed until § 5.

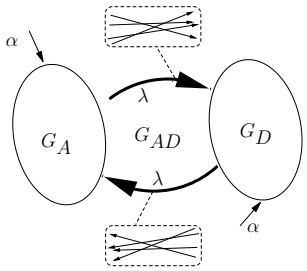


Figure 2. The framework for co-ranking authors and documents. G_A is the social network of authors. G_D is the citation network of documents. G_{AD} is the authorship network. α is the jump probability for the Intra-class random walks. λ is a parameter for coupling the random walks, quantifying the importance of G_{AD} versus that of G_A and G_D.

Before making everything precise, let us briefly sketch the co-ranking framework. The conceptual scheme is illustrated in Fig. 2. 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 other[1].

[1]This 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. Note, however, that in the latter case one would need a different normalization instead of (2), satisfying $\|\mathbf{a}\|_1\lambda_{AD} = \|\mathbf{d}\|_1\lambda_{DA}$.

3.2 PageRank: two random walks

First of all, we are going to rank the networks of authors and documents independently, according to the PageRank paradigm [5]. Consider a random walk on the author network G_A and let A be the transition matrix (A will be defined in § 5). Fix some α and say that at each time step with probability α we do not make a usual random walk step, but instead jump to any vertex, chosen uniformly at random. This is another random walk with the transition matrix

$$\widetilde{A} = (1-\alpha)A + \frac{\alpha}{n_A}\mathbf{1}\mathbf{1}^{\mathbf{T}} \tag{3}$$

Here $\mathbf{1}$ is the vector of n_A entries, each being equal to one. Let $\mathbf{a} \in \mathbf{R}^{n_A}, \|\mathbf{a}\|_1 = 1$ be the only solution of the equation

$$\mathbf{a} = \widetilde{A}^T\mathbf{a} \tag{4}$$

.

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 (4) follows from the random walk $\widetilde{A}$ being ergodic, and this is why we are using $\widetilde{A}$ instead of A. ($\alpha > 0$ guarantees irreducibility, because we can jump to any vertex in the graph.)

Documents can be ranked in the citation network G_D in a similar way. In particular,

$$\widetilde{D} = (1-\alpha)D + \frac{\alpha}{n_D}\mathbf{1}\mathbf{1}^{\mathbf{T}}, \tag{5}$$

For details regarding Markov chains, specifically that the stationary probabilities of an ergodic Markov chain can be computed by iterating the powers of the transition matrix, see any textbook on stochastic processes, such as [18].

3.3 (m, n, k, λ)–coupling of two Intra-class random walks

To couple these two random walks we construct a combined random walk on the heterogeneous graph $G = G_A \cup G_D \cup G_{AD}$. A probability distribution 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. In fact, we will multiply all of them by 2 to ensure that $\|\mathbf{a}\|_1 = \|\mathbf{d}\|_1 = 1$. Of course, the greater the stationary probability (ranking score), the higher the rank of an author or a document.

The coupling is parameterized by four parameters, m, n, k and λ. Ordinary PageRank score is sometimes viewed as the probability that a *random surfer* will be on this web page at some moment in the distant future. Similarly, we present 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 $v \in V_A$, then he can either make an *Intra-class step* (one step of the random walk parameterized by $\widetilde{A}$) 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 $\widetilde{D}$) 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 $(\widetilde{A}\mathbf{a}, \mathbf{0})$ or from $(\mathbf{0}, \mathbf{d})$ to $(\mathbf{0}, \widetilde{D}\mathbf{d})$, while one Inter-class step changes the probability distribution from $(\mathbf{a}, \mathbf{d})$ to $(DA^T\mathbf{d}, AD^T\mathbf{a})$.

Now, the combined random walk is defined as follows:

1. If the current state of RS is some 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 some 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.

It is convenient to write 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.

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: $\quad \mathbf{b} \leftarrow U^T\mathbf{c}$
4: $\quad \mathbf{c} \leftarrow V^T\mathbf{b}$
5: **end for**
6: $\mathbf{b} \leftarrow U^T\mathbf{c}$
7: return $\mathbf{b}$

Now, everything is ready to realize co-ranking in the following procedure, *CoupleWalk* (Algo. 2). It should be noted that the very recent work [11] of learning on subgraphs can be considered an implicit special version of our algorithm with infinite k and $m = n = 1$.

4. Convergence Analysis

We need to ensure that Algo. 2 converges. Fortunately, it is no more than an iterative computation of the stationary probabilities of a Markov chain that is the combined random walk. To see this, observe that $BiWalk(U, V, \mathbf{x}, k) = U^T(V^TU^T)^k\mathbf{x}$. Therefore, lines 6 and 7 in Algo. 2 can be rewritten as:

Algorithm 2 Coupling random walks for co-ranking

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

$\mathbf{a} \leftarrow \frac{1}{n_A}\mathbf{1}$
$\mathbf{d} \leftarrow \frac{1}{n_D}\mathbf{1}$
repeat
  $\mathbf{a}' \leftarrow \mathbf{a}$
  $\mathbf{d}' \leftarrow \mathbf{d}$
  $\mathbf{a} \leftarrow (1-\lambda)(\widetilde{A}^T)^m\mathbf{a}' + \lambda BiWalk(DA, AD, \mathbf{d}', k)$
  $\mathbf{d} \leftarrow (1-\lambda)(\widetilde{D}^T)^n\mathbf{d}' + \lambda BiWalk(AD, DA, \mathbf{a}', k)$
until $|\mathbf{a} - \mathbf{a}'| \leq \epsilon$
return $\mathbf{a}$, $\mathbf{d}$

$$\mathbf{a}^{t+1} = (1-\lambda)(\widetilde{A}^T)^m\mathbf{a}^t + \lambda DA^T(AD^TDA^T)^k\mathbf{d}^t \quad (6)$$

$$\mathbf{d}^{t+1} = (1-\lambda)(\widetilde{D}^T)^n\mathbf{d}^t + \lambda AD^T(DA^TAD^T)^k\mathbf{a}^t \quad (7)$$

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

$$M = \begin{bmatrix} (1-\lambda)(\widetilde{A}^T)^m & \lambda DA^T(AD^TDA^T)^k \\ \lambda AD^T(DA^TAD^T)^k & (1-\lambda)(\widetilde{D}^T)^n \end{bmatrix}. \quad (8)$$

Clearly, $\mathbf{v}^{t+1} = M\mathbf{v}^t$, and M is a stochastic matrix that parameterizes the combined random walk. It is also easy to see that for $0 < \alpha, \lambda < 1$, this Markov Chain is ergodic. Thus, the stationary probabilities can be found as $\lim_{n\to+\infty} M^n\mathbf{v}$, for any initial vector $\mathbf{v}$. In particular, $\mathbf{a}$ and $\mathbf{d}$ in Algo. 2 will converge to the ranking scores as we defined them. In practice, the convergence can be established numerically.

5. 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 $\widetilde{A}$), D (giving rise to $\widetilde{D}$), AD and DA.

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

The citation document network G_D is defined as follows: there is a directed edge from d_i to d_j, if document d_i cites document d_j at least once. The graph is not weighted; we ignore repeated citations from the same document to the same document. Self-citations are technically allowed, but, presumably, there are none.

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

$$P(j|i) = D_{i,j} = \frac{n^D_{i,j}}{n^D_i}, \quad (9)$$

where $n^D_{i,j}$ is the indicator of whether document i cites j; n^D_i 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 $\frac{1}{n_D}$.

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

Rather than taking G_A to be the social network, where two authors are connected by an edge, if they collaborated on a paper, we come up with a more general definition. This definition employs the notion of a *social event*. A social event could be any kind of activity, involving a group of authors. A co-occurrence of two authors in a social event is supposed to create or strengthen their social ties. In particular, we view collaborating on a paper or co-participating in a conference as such "co-occurrences". Let the set of social events be $\mathcal{E} = \{e_i\}$, where an event e_i is identified with the set of participating authors. We construct G_A as an unweighted graph, where two authors are connected by an edge if they co-occur in some social event $e \in \mathcal{E}$.

Intuitively, a paper of fewer authors infers stronger social ties among them on average (cf. [15]). To take this into account, we first make the graph G_A weighted. Define the social tie function $\tau(i, j, e_k) : \mathcal{A} \times \mathcal{A} \times \mathcal{E} \to [0, 1]$ 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. The strength of the social tie depends on the size of the corresponding social event. If there are only two people taking part in an event (say, collaborating on a paper), we say that it infers a *unit social tie*. Otherwise, the tie is somehow normalized by the size of the event. There are many ways to do that, we arbitrarily chose one that seemed promising to us:

$$\tau(i, j, e_k) = \frac{\mathbb{I}(i, j \in e_k)}{|e_k|(|e_k| + 1)/2} \quad (10)$$

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 $a_i = a_j$). $|e_k| \geq 2$ is the number of authors involved in event e_k. For $|e_k| = 1$, only a self social tie of that author is inferred. 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:

$$T_{i,j} = \sum_{e_k \in \mathcal{E}} \tau(i, j, e_k) \quad (11)$$

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, we consider two kinds of social events. The first kind is a collaboration on a paper (even if the paper has a single author), in this case the 'event' includes exactly all the authors of this paper. The second kind is the appearance of names in conference proceeding lists. Each conference instance (i.e. *ACM SIGMOD '01*) is a separate event, consisting of the authors who took part in it. We treat the two kinds equally, and we find it appropriate because of the normalization (10).

We proceed to define the Intra-class random walk on G_A in a natural way, namely, the next step is chosen according to the weights on the edges. 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:

$$P(j|i) = A_{i,j} = \frac{T_{i,j}}{\sum_j T_{i,j}}. \tag{12}$$

Here T is symmetric due to the design of τ. A is not necessarily symmetric because row sums can be different. $\widetilde{A}$ is defined accordingly.

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

The bipartite authorship graph G_{AD} is defined in the natural way. Namely, the entries in its adjacency matrix E_{AD} are the values of the indicator function of a document being written by an author, i.e.

$$E_{AD}(i,j) = \mathbb{I}(d_j \text{ is authored by } a_i). \tag{13}$$

Using the adjacency matrix E_{AD}, we define a weight matrix $W_{AD} = (w(i,j))$ as follows:

$$w(i,j) = \frac{E_{AD}(i,j)}{n_j^A}, \tag{14}$$

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

$$P(d_j|a_i) = AD_{i,j} = \frac{w(i,j)}{\sum_k w(i,k)}, \tag{15}$$

$$P(a_i|d_j) = DA_{j,i} = \frac{w(i,j)}{\sum_k w(k,i)}. \tag{16}$$

This completes the descriptions of networks and random walks [2]. Note that (14) implies $\sum_k w(k,j) = 1$. The design of the matrices AD and DA is asymmetric to reflect the asymmetric relationship between authors and documents. Indeed, it is better for an author to create many good documents; for a document it is better to have better authors, but not necessarily *more* authors.

6 Experiments

6.1 Data Preparation

For experiments, we use data from CiteSeer [9], a popular search engine and digital library 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 in the database. We also keep 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. In order to extract the information about conference proceedings, we perform a fuzzy matching of the titles of CiteSeer documents with the titles of documents listed by conferences in the manually prepared data from DBLP.

While performing the ranking on the full data collection is technically feasible, the bias in collection sizes towards certain domains can undermine the fairness of ranking scientists from different areas. Therefore, we start from categorizing the documents into domains. In particular, we apply the Latent Dirichlet Allocation (LDA) model [2] 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.

6.2 Author Subset Generation

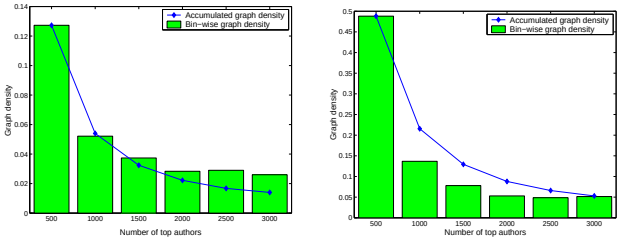
For a given topic (out of five listed above), LDA produces the 'topic weight' for each document. The sum of the topic weights over all documents of an author is the 'accumulated topic weight' for that author; very crudely, this is just the number of papers classified as belonging to a given topic.

We apply a two-step heuristic that further reduces the

[2] It should be noted that in this construction G_{AD} and G_A are strongly correlated, since G_{AD} intrinsically includes the information about co-authorship. Also, the co-occurrence in conference proceeding lists is correlated with co-authorship. We did not observe any difficulties from that.

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*



(a) Social network density (b) Citation network density

Figure 3. Density of author collaboration/citation networks vs. the number of top authors according to LDA accumulated weights, on the topic *data management***.**

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 see, how much information will be compromised when the problem is reduced in scale, we perform a simple statistical analysis of the graph densities (defined as $|E|/|V|^2$) of on author subsets with different sizes. Fig. 3(a) and Fig. 3(b) present the graph densities of social and citation networks for the subsets of top authors with respect to LDA accumulated topic weights, on the topic 48, *data management*. In the following experiments, for each topic we work with 500 authors with the highest topic weights. Once the author subset is generated, we work only on the documents by these authors.

6.3 Author Rankings

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;
- **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 § 5;
- **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) [12], 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, shuffled 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,

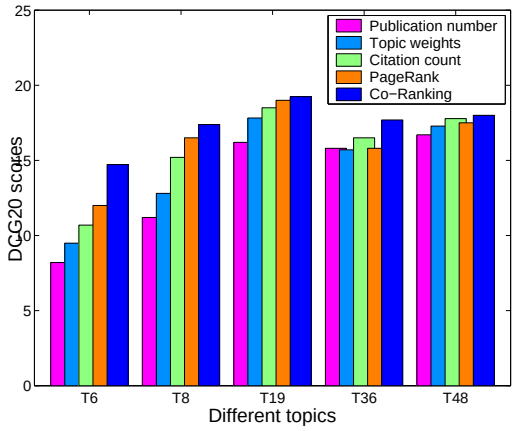


Figure 4. DCG_{20} scores for author rankings: number of papers, topic weights, number of citations, PageRank, and Co-Ranking.

being a fellowship of ACM/IEEE, etc. The judges' assessment scores are averaged. We observe a high agreement between the two judges.

The DCG_{20} scores obtained are presented in Fig. 4. 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 on the topics *data management* and *learning and classifications* in Table 2 and Table 3. Along with both tables, the ranks based on simple metrics are also presented. 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.

6.4 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 and Table 4 present the top documents according to Co-Ranking in the topics *data management* and *learning and classification*. For each document, we show the title, the first three authors (because of space constraints), the year of publication, and the number of citations. To get more information, follow 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

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# is the number of neighbors in the social network; p# is the number of papers; cite# is the number of citations; r denotes the ranks by the corresponding methods.

r	author names	con#	r	p#	r	cite#	r
1	Sebastian Thrun	178	6	293	8	782	4
2	Bernd Girod	72	180	217	10	313	33
3	Jurgen Schmidhuber	152	21	160	14	446	18
4	Stephen Muggleton	99	88	45	200	492	11
5	Robert E. Schapire	133	35	67	105	1093	1
6	Avrim Blum	102	82	295	7	239	58
7	Trevor Hastie	68	199	88	52	263	53
8	Rakesh Agrawal	68	197	129	22	843	2
9	Manuela Veloso	155	18	196	11	491	12
10	Thomas G. Dietterich	74	173	53	159	514	8
11	Alex Pentland	126	47	110	36	369	21
12	Michael I. Jordan	172	9	91	50	566	7
13	David J.C. MacKay	22	379	73	91	349	25
14	David Haussler	113	61	65	112	351	24
15	David Heckerman	77	163	56	150	491	14

Table 3. Top authors in the topic *learning and classifications* when $m = 2$, $n = 2$, $k = 1$. con# is the number of neighbors in the social network; p# is the number of papers; cite# is the number of citations; r denotes the ranks by the corresponding methods.

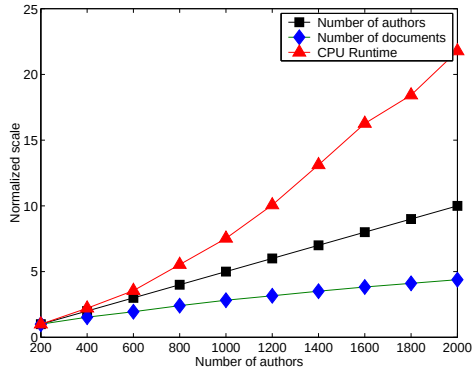
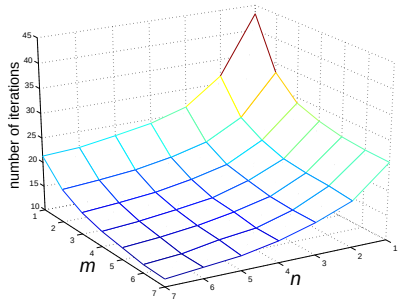
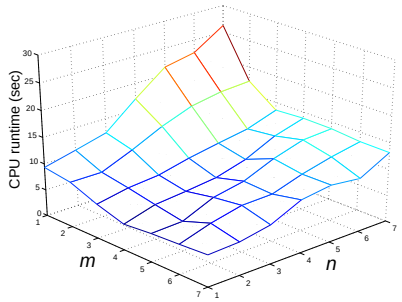


Figure 5. Average CPU runtime and number of documents w.r.t. the number of authors for five topics, where $m = 2$, $n = 2$, $k = 1$. Appropriate units have been chosen, so that a single normalized scale can be used. Everything is averaged over five topics.



(a) Number of iterations until convergence



(b) CPU runtime until convergence

Figure 6. Effect of m-n on convergence.

not produce any judgment on the document rankings we obtained due to the above concerns. In general, one can observe from Table 1 and Table 4 that top documents typically have many citations.

6.5 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.

6.6 Convergence and Runtime

Finally, we present some observations about the computational complexity: We observed that the algorithm converges faster for larger α. This is expected because a Markov chain takes a shorter time to reach the stationary status if the transition matrix is closer to uniform.

We fix $k = 1$, $\lambda = 0.2$, $\alpha = 0.1$ and vary m and n. Fig. 6(a) and Fig. 6(b) show the effect of m and n on the number of iterations before convergence and the runtime of the program. It can be seen that for large and increasing m and n the number of iterations decreases slowly. This is because the Intra-class random walks have enough steps to become nearly stationary before the next Inter-class step.

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. Fig. 5 shows the average CPU runtime w.r.t. to the number of authors. The Co-Ranking was implemented in Python and tested on Intel CoreDuo 1.66 GHz, 1G RAM, Windows O.S.

7. 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 PageR-

cs-id	title	authors	year	cite
364205	Learning Bayesian Networks: The Combination of Knowledge and Statistical Data	David Heckerman, Dan Geiger, David Chickering	1994	351
142690	Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension	David Haussler, Michael Kearns, Robert Schapire	1992	85
124084	Efficient Distribution-free Learning of Probabilistic Concepts	Michael J. Kearns, Robert E. Schapire	1993	115
25286	Bagging Predictors	Leo Breiman	1996	657
384587	Reinforcement Learning: Introduction	Richard Sutton	1998	614
48796	An Information-Maximization Approach to Blind Separation and Blind Deconvolution	Anthony J. Bell, Terrence J. Sejnowski	1995	491
41366	Stacked Generalization	David H. Wolpert	1992	367
527057	Optimization by Simulated Annealing	S. Kirkpatrick	1993	1527
25887	Mining Association Rules between Sets of Items in Large Databases	Rakesh Agrawal, Tomasz Imielinski, Arun Swami	1993	921
20336	Generalized Additive Models	Trevor Hastie, Robert Tibshirani	1995	450
123646	Experiments with a New Boosting Algorithm	Yoav Freund, Robert E. Schapire	1996	500
528249	Hierarchical Mixtures of Experts and the EM Algorithm	Michael I. Jordanand Robert A. Jacobs	1993	472
543817	The Strength of Weak Learnability	Robert E. Schapire	1990	273
63435	Systematic Nonlinear Planning	David McAllester and David Rosenblitt	1991	226
434739	Bayesian Interpolation	David J.C. MacKay	1991	244

Table 4. Top documents in the topic *learning and classification*

ank algorithm as applied to the social network only. We did not evaluate the ranking of documents due to the lack of any objective criteria.

Possible directions of future research include: (1) A larger empirical evaluation could be carried out to compare the Co-Ranking framework with other methods and find out, on which inputs it performs unsatisfactorily; (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, speed of convergence, stability, etc. It is also interesting to try to bring it into correspondence with the existing general frameworks for link based rankings (see e.g. [4]). We expect there to be interesting interconnections with the HITS algorithm and its variations, if authors are viewed as authorities and documents are viewed as hubs; (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. Studying the effect of introducing different $\lambda_{AD} \neq \lambda_{AD}$ may serve as a starting point; (4) Presumably, the framework can be generalized for co-ranking entities of several types. Even for the case of two types, its applications are not limited to co-ranking authors and documents either.

References

[1] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Trans. Inter. Tech.*, 5(1):92–128, 2005.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[3] J. Bollen, M. A. Rodriguez, and H. Van de Sompel. Journal Status. *arXiv.org:cs/0601030*, 2006.

[4] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, 5(1):231–297, 2005.

[5] 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.

[6] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding Scientific Gems with Google. *J.INFORMET.*, 1:8, 2007.

[7] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM Press, 2001.

[8] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(60):471–479, November 1972.

[9] 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.

[10] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102:16569, 2005.

[11] J. Huang, T. Zhu, R. Greiner, D. Zhou, and D. Schuurmans. Information marginalization on subgraphs. In *PKDD*, pages 199–210, 2006.

[12] 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.

[13] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[14] S. Lehmann, A. D. Jackson, and B. E. Lautrup. Measures and mismeasures of scientific quality, 2005.

[15] 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.

[16] S. A. Macskassy and F. J. Provost. Suspicion scoring based on guilt-by-association, collective inference, and focused data access. In *NAACSOS conference proceedings*, June 2005.

[17] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Inf. Process. Manage.*, 12(5):297–312, 1976.
[18] S. M. Ross. *Stochastic Processes*. Wiley Press, 1995.
[19] A. Sidiropoulos and Y. Manolopoulos. A new perspective to automatically rank scientific conferences using digital libraries. *Inf. Process. Manage.*, 41(2):289–312, 2005.
[20] R. Todorov and W. Gilazel. Journal citation measures: a concise review. *J. Inf. Sci.*, 14(1):47–56, 1988.
[21] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
[22] P. Weingart. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1):117–131, 2005.
[23] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–275. ACM Press, 2003.
[24] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 113–120, New York, NY, USA, 2002. ACM Press.