Electronic Transactions on Numerical Analysis.
Volume 33, pp. 1-16, 2008.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

EVALUATING SCIENTIFIC PRODUCTS BY MEANS OF CITATION-BASED MODELS: A FIRST ANALYSIS AND VALIDATION*

DARIO A. BINI†, GIANNA M. DEL CORSO‡, AND FRANCESCO ROMANI‡

Dedicated to Gérard Meurant on the occasion of his 60th birthday

Abstract. Some integrated models for ranking scientific publications together with authors and journals are presented and analyzed. The models rely on certain adjacency matrices obtained from the relationships between citations, authors and publications, which together give a suitable irreducible stochastic matrix whose Perron vector provides the ranking. Some perturbation theorems concerning the Perron vectors of nonnegative irreducible matrices are proved. These theoretical results provide a validation of the consistency and effectiveness of our models. Several examples are reported together with some results obtained on a real set of data.

Key words. Page rank, Perron vector, perturbation results, impact factor

AMS subject classifications. 65F15

1. Introduction.

Ranking scientific publications independently of their contents is a problem of great practical importance and of particular theoretical interest. Most of the attempts to evaluate the quality of a scientific publication are based on the analysis of the citations received.

Recently, a certain interest has been given to citation analysis and to related models, mainly because they enable one to rigorously measure delicate concepts that otherwise would be difficult to capture, such as the quality of the research performed by scholars or the reputation and the influence of researchers. Indeed, only a careful reading of a paper can tell what is the real nature of a citation; in fact, an analysis independent of the context cannot distinguish between critical and positive citations. However, it is interesting to point out that in all of the models presented in the literature, receiving a citation is considered a positive fact regardless of the nature of the citation.

A common measure to assess the importance of a scientific journal is the well known *Impact Factor* calculated by the Institute of Scientific Information (ISI) and introduced by Garfield [9]. However, not all the scientific community agrees about the effectiveness of this measure, because regarding all the citations with the same weight is essentially a metric of popularity, and it does not seem to capture criteria such as prestige or importance [4]. Many other proposals have been made over the years, starting with the one by Pinski and Narin [16], where the authors anticipated of many years the Google model [5]. This same model recently has been reconsidered [15] and it has been proved that this kind of approach is the only one satisfying a number of very reasonable requirements. Another proposal is the *Eigenfactor* method [2], which combines a Google-like approach with a time-aware mechanism.

Though most of the related literature addresses the problem of ranking journals [4, 15, 16], some other authors propose strategies for ranking scholars [11, 14] and scientific institutions [17, 18]. In our study we aim to present and analyze an integrated model, in which we consider relationships among more subjects, such as authors, papers, journals, fields, and institutions.

* Received January 31, 2008. Accepted April 21, 2008. Published online on October 17, 2008. Recommended by Lothar Reichel.
†Dipartimento di Matematica, Università di Pisa, Largo B. Pontecorvo 5, 56127 Pisa, Italy (bini@dm.unipi.it).
‡Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy ({delcorso,romani}@di.unipi.it).

ETNA
Kent State University
etna@mcs.kent.edu

In particular, the idea is that in order to determine the importance of a journal, one has to take into account not only the "quality" of the citations from other papers (as done by the ranking schemes in the literature), but also the "quality" of papers and of their authors. Similarly, an author is important if he/she publishes important papers in important journals and maybe with important co-authors. A paper is important if it receives citations from other important papers, but also if it is published in an important journal and is written by important authors. This leads to an integrated model where each player —journals, authors, and papers— contributes to the determination of the score of the others. Throughout, we refer to these players as *subjects*.

The basic principle that we follow is that the importance of a subject is the weighted sum of the importances of all the subjects that are related to it in a sense that will be made clear later on. In this model, the sum of the weight coefficients must be one, so that the overall amount of importance is neither destroyed nor created.

We start with the simple *one-class model*, in which only the class of *Papers* is taken into account, and where the importance is given on the basis of *citations*. Then we consider more general models, in which other actors are involved as well. The *two-class model* takes into account the class of *Authors* as well as Papers, and the importance is given on the basis of citations and of *authorship*. The *three-class model* adds to the latter the class of *Journals*. More elaborate models involving, say, research areas and institutions can be introduced and are left to future work.

In all these models, the vector with the rating of all the involved subjects is obtained as the positive invariant vector of an irreducible row-stochastic matrix, normalized so that the sum of its components is one (the Perron vector).

We perform a consistency analysis of the introduced models and prove new perturbation theorems concerning the Perron vector of the stochastic matrices involved which extend some result given in [7]. These perturbation results are the matrix formulation of the desired properties which are consistent with our models. In particular, in the one-class model, we prove that a paper which receives a new citation has an increase of its rank which is larger than the increase received by the other papers. Similarly, we prove that if a new paper is introduced and this paper contains a citation to a given paper, then the importance of the latter has an increase larger than the ones received by the other papers. These properties keep their validity in the two-class and in the three-class models. Several examples are given which confirm the expected properties.

The paper is organized as follows. In Section 2, we introduce and analyze the one-class model; and in Section 3, we describe the two-class model and its variants. Section 4 contains a brief description of the three-class model. In Section 5, we report results of some experiments performed on the Citeseer.IST [6] database. In Section 6, we draw conclusions and discuss some open issues.

2. One-class model. Assume that we are given n papers numbered from 1 to n together with the $n \times n$ adjacency matrix $H = (h_{i,j})$, such that $h_{i,j} = 1$ if paper i cites paper j, $h_{i,j} = 0$ otherwise. Following a model similar to Google [5], we assume that the importance p_j of paper j is given by the importances of the papers p_i that cite paper j, scaled by the factor d_i which is the number of citations contained in paper i. In this way, the importance given by paper i is uniformly distributed among all the papers cited therein, and the principle that the importance of a subject is neither destroyed nor created is respected.

Here and below, we denote by $\boldsymbol{e}$ the vector of appropriate length with all components equal to one. We denote by $\boldsymbol{e}_k$ the kth column of the identity matrix of appropriate size. The size of vectors and matrices, if not specified, is deduced by the context. Given a vector $\boldsymbol{v} = (v_i)$ of n components, with the expression $\operatorname{diag}(\boldsymbol{v})$ we denote the $n \times n$ diagonal matrix

ETNA
Kent State University
etna@mcs.kent.edu

having diagonal entries v_i, $i = 1, \ldots, n$.

The scaling factors $d_i = \sum_j h_{i,j}$ define the vector $\boldsymbol{d} = (d_i)$, which satisfies the equation $\boldsymbol{d} = H\boldsymbol{e}$. Moreover, if $d_i \neq 0$ for all i, the matrix

$$P = (p_{i,j}) = \mathrm{diag}(\boldsymbol{d})^{-1}H$$

is row-stochastic, that is, $\sum_j p_{i,j} = 1$.

Since in principle there might be papers with an empty set of citations, the matrix H might have some null rows and some factors d_i might be zero. This fact may make the model inconsistent. We cure this drawback by introducing a *dummy paper*, paper $n+1$, which cites and is cited by all the existing papers except itself. In this way the new adjacency matrix of size $n+1$, which with an abuse of notation we still denote by H, has no null row and is irreducible. From the modeling point of view, the dummy paper collects the importance of all the papers and redistributes it uniformly to all the subjects.

Note that the introduction of the dummy paper guarantees that the matrix P is stochastic, acyclic and aperiodic. This provides important computational advantages in the numerical solution of the model. It is interesting to observe that a similar technique is used in the Google model where, unlike in our case, a damping factor is also introduced.

The equation that we obtain in this way is

$$\boldsymbol{p}^T = \boldsymbol{p}^T P, \qquad P = \mathrm{diag}(H\boldsymbol{e})^{-1}H \tag{2.1}$$

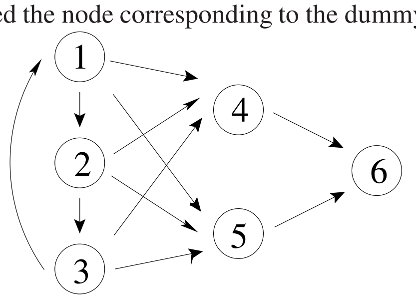
and, since the matrix $\mathrm{diag}(H\boldsymbol{e})^{-1}H$ is nonnegative and irreducible, from the Perron-Frobenius theorem there exists a unique vector $\boldsymbol{p} = (p_i)$ such that $p_i > 0$, $\sum_i p_i = 1$, which solves (2.1). We call $\boldsymbol{p}$ the *Perron vector* of P.

Equation (2.1) states that the importance of paper j is given by the sum of the importances received by all the other papers, that is, by the values p_i scaled by the factors $h_{i,j}/\sum_s h_{i,s}$, $i = 1, \ldots, n+1$, i.e.,

$$p_j = \sum_{i=1}^{n+1} p_i \frac{h_{i,j}}{\sum_{s=1}^{n+1} h_{i,s}}, \qquad j = 1, 2, \ldots, n-1.$$

In fact, each paper i uniformly distributes its importance to all the $\sum_s h_{i,s}$ papers that it cites.

EXAMPLE 2.1. Consider the case of 6 papers, where citations are given by the following graph. We have not reported the node corresponding to the dummy paper.



The adjacency matrix, including the dummy paper, is

$$H = \left[\begin{array}{cccccc|c} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}\right].$$

Papers 1, 2, and 3 are on the same rank level: except for the dummy paper, they receive one citation each and are inside a cycle. Papers 4 and 5 receive three citations by papers 1, 2, 3 and are on the same level but in a higher position with respect to papers 1, 2, and 3. Paper 6 receives only two citations by papers 4 and 5. Therefore, in a model based only on the number of citations, the rank of paper 6 should be inferior to the rank of papers 4 and 5. However, since paper 6 is cited by two papers which are more important than papers 1, 2, and 3, one should expect that in our model its rank is higher. In fact, the left eigenvector of $\mathrm{diag}(H\boldsymbol{e})^{-1}H$ is

$$\boldsymbol{p}^T = (0.0784314, 0.0784314, 0.0784314, 0.117647, 0.117647, 0.176470, 0.352941)$$

where $p_1 = p_2 = p_3 < p_4 = p_5 < p_6$ and paper 6 reaches the highest rank as expected. Modifying the data by adding a citation from paper 5 to paper 4 yields the vector

$$\boldsymbol{p}^T = (0.075472, 0.075472, 0.075472, 0.150943, 0.113208, 0.169811, 0.339623)$$

where $p_1 = p_2 = p_3 < p_5 < p_4 < p_6$ and paper 4 gets an higher rank than paper 5.

An interesting question is to figure out what happens to the Perron vector $\boldsymbol{p}$ of the matrix P if P is perturbed in the following way: a new link is inserted in the graph connecting node r to node s, where in the original adjacency matrix $h_{r,s} = 0$. That is, the new matrix $\widehat{H}$ is constructed in such a way that $\widehat{h}_{r,s} = 1$ while $\widehat{h}_{i,j} = h_{i,j}$ for the remaining entries and $\widehat{P} = \mathrm{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$.

One would expect that the paper receiving the new citation increases its value more than the other papers do, i.e., the component $\widehat{p}_s$ of the Perron vector $\widehat{\boldsymbol{p}}$ of the matrix $\widehat{P}$ constructed from $\widehat{H}$ increases more than the remaining components. Formally, $\widehat{p}_s/p_s \geq \widehat{p}_i/p_i$ for any i.

The following result of [7], which extends the result of [8], is useful for formally proving this fact.

THEOREM 2.2. *Let A and B be $n \times n$ nonnegative irreducible matrices having the same spectral radius ρ. Let $\boldsymbol{x} = (x_i)$ and $\boldsymbol{y} = (y_i)$ be their positive Perron vectors such that $A\boldsymbol{x} = \rho\boldsymbol{x}$, $B\boldsymbol{y} = \rho\boldsymbol{y}$. Assume that A and B differ only in the rows having index in the set $\Omega \subset \{1, 2, \ldots, n\}$. Assume that the set Ω and its complement are nonempty. Then*

$$\min_{i\in\Omega} \frac{x_i}{y_i} \leq \frac{x_j}{y_j} \leq \max_{i\in\Omega} \frac{x_i}{y_i}, \quad j = 1, \ldots, n.$$

The above result yields information about the variation of the right Perron vector under perturbation of rows. Here we need a sort of dual result concerning the variation of the left Perron vector. The following theorem provides this extension under specific perturbations.

THEOREM 2.3. *Let H be an irreducible adjacency matrix, let (r, s) be a pair of integers such that $h_{r,s} = 0$, and let q be the number of nonzero entries in the rth row. Define $\widehat{H} = (\widehat{h}_{i,j})$ such that $\widehat{h}_{r,s} = 1$ and $\widehat{h}_{i,j} = h_{i,j}$ otherwise. Let $P = \mathrm{diag}(H\boldsymbol{e})^{-1}H$ and $\widehat{P} = \mathrm{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$, and denote by $\boldsymbol{p}$ and $\widehat{\boldsymbol{p}}$ their corresponding left Perron vectors. Then*

$$\sigma\frac{\widehat{p}_r}{p_r} \leq \frac{\widehat{p}_j}{p_j} \leq \frac{\widehat{p}_s}{p_s}, \quad j = 1, \ldots, n, \tag{2.2}$$

and $\sigma\widehat{p}_r/p_r < \widehat{p}_s/p_s$, for $\sigma = q/(q + 1)$. Moreover,

$$\frac{\widehat{p}_s}{p_s} > 1, \tag{2.3}$$

ETNA
Kent State University
etna@mcs.kent.edu

and

$$\frac{\widehat{p}_j}{p_j} < \frac{\widehat{p}_s}{p_s}, \quad \textit{if } h_{r,j} \neq 0. \tag{2.4}$$

Proof. Let D be the diagonal matrix having one on the main diagonal except for the rth diagonal entry which is $\sigma = q/(q+1)$ and observe that

$$\widehat{P} = DP + \frac{1}{q+1}\boldsymbol{e}_r\boldsymbol{e}_s{}^T.$$

Define $C = D^{-1}\widehat{P}D$. Then $\boldsymbol{z} = D\widehat{\boldsymbol{p}}$ is a left eigenvector of C, i.e., $\boldsymbol{z}^T C = \boldsymbol{z}^T$. Moreover, $z_i = \widehat{p}_i$ for $i \neq r$, $z_r = \sigma\,\widehat{p}_r$. Since

$$C = PD + \frac{1}{q}\boldsymbol{e}_r\boldsymbol{e}_s{}^T,$$

the matrix C differs from the matrix P only in the columns r and s. Applying Theorem 2.2 with $A = C^T$ and $B = P^T$ yields

$$\min\left\{\frac{z_r}{p_r}, \frac{z_s}{p_s}\right\} \leq \frac{z_j}{p_j} \leq \max\left\{\frac{z_r}{p_r}, \frac{z_s}{p_s}\right\}, \quad j = 1, \ldots, n,$$

and since $z_r = \sigma\,\widehat{p}_r$, $z_j = \widehat{p}_j$ for $j \neq r$, one gets

$$\min\left\{\sigma\frac{\widehat{p}_r}{p_r}, \frac{\widehat{p}_s}{p_s}\right\} \leq \frac{\widehat{p}_j}{p_j} \leq \max\left\{\sigma\frac{\widehat{p}_r}{p_r}, \frac{\widehat{p}_s}{p_s}\right\}, \quad j = 1, \ldots, n. \tag{2.5}$$

Now, it suffices to prove that $\sigma\widehat{p}_r/p_r < \widehat{p}_s/p_s$ in order to deduce (2.2) from (2.5). Assume, by contradiction, that $\sigma\widehat{p}_r/p_r \geq \widehat{p}_s/p_s$, and deduce from (2.5) that $\widehat{p}_j/p_j \leq \sigma\widehat{p}_r/p_r$. For $j = r$ this implies that $\sigma \geq 1$, which is a contradiction.

The inequality (2.4) follows from the fact that $\sigma < 1$ and $p_{r,j} \neq 0$ since

$$\widehat{p}_j = \sum_i \widehat{p}_{i,j}\widehat{p}_i = \sum_{i\neq r} p_{i,j}\widehat{p}_i + \sigma p_{r,j}\widehat{p}_r < \sum_i p_{i,j}\widehat{p}_i = \sum_i p_{i,j}p_i\frac{\widehat{p}_i}{p_i} \leq p_j\frac{\widehat{p}_s}{p_s},$$

where the last inequality is obtained by replacing $\widehat{p}_i/p_i$ by $\widehat{p}_s/p_s$, in view of (2.2), and using the fact that $\sum_i p_{i,j}p_i = p_j$.

Concerning (2.3), if $\widehat{p}_s/p_s \leq 1$, then (2.2) would imply $\widehat{p}_j/p_j \leq 1$. Since H is irreducible, there exists an integer $j \neq r$ such that $h_{r,j} \neq 0$; that is, in view of (2.4) one obtains $\widehat{p}_j/p_j < 1$. Hence, $1 = \sum_j \widehat{p}_j < \sum_j p_j = 1$, which is a contradiction. □

The above theorem says that if we introduce a new citation from paper r to paper s, paper s which receives the citation has an increase of importance greater than or equal to the increase received by any other paper. Moreover, if paper j is cited by paper r, i.e., if $h_{r,j} \neq 0$, then the increase of importance of paper j is strictly less than that of paper s.

Another interesting issue concerns the variation of the Perron vector when a new node is introduced in the graph with a single link to another node. One would expect that the paper that receives the new citation should improve its rank with respect to the other papers. We can provide a formal proof of this fact.

Let V be an $n \times n$ adjacency matrix and denote by $\widetilde{V}$ the $(n+1)\times(n+1)$ matrix having V as its leading principal submatrix and zeros in the last row and in the last column. Let H be the $(n+1)\times(n+1)$ matrix having V as its leading principal submatrix and having ones in the last row and last column except for the last diagonal entry which is zero. Similarly, define $\widetilde{H}$ the $(n+2)\times(n+2)$ matrix having $\widetilde{V}$ as its leading principal submatrix and having ones in the last row and last column except for the last diagonal entry which is zero. Thus,

$$H = \left[\begin{array}{ccc|c} & V & & 1 \\ & & & \vdots \\ & & & 1 \\ \hline 1 & \dots & 1 & 0 \end{array}\right], \quad \widetilde{H} = \left[\begin{array}{ccc|c|c} & V & & 0 & 1 \\ & & & \vdots & \vdots \\ & & & 0 & \\ \hline 0 & \dots & 0 & 0 & 1 \\ \hline 1 & \dots & \dots & 1 & 0 \end{array}\right].$$

Observe that H represents the adjacency matrix of the citation graph associated with V where the dummy paper is added, while $\widetilde{H}$ represents the citation graph associated with the matrix obtained by adding a new paper with no citations, where once again the dummy paper is added after the new insertion.

Both H and $\widetilde{H}$ are irreducible and we can scale their rows to get the stochastic matrices

$$P = \mathrm{diag}(H\boldsymbol{e})^{-1}H, \quad \widetilde{P} = \mathrm{diag}(\widetilde{H}\boldsymbol{e})^{-1}\widetilde{H}.$$

We have the following lemma.

LEMMA 2.4. *Let $\boldsymbol{p}^T = (p_i)$ be the left Perron vector of P. Then the vector*

$$\widetilde{\boldsymbol{p}}^T = \theta\left(p_1, \dots, p_n, \frac{1}{n}p_{n+1}, \frac{n+1}{n}p_{n+1}\right), \quad \theta = 1/(1+\frac{2}{n}p_{n+1}), \tag{2.6}$$

is the left Perron vector of $\widetilde{P}$.

Proof. Note that the vectors $H\boldsymbol{e}$ and $\widetilde{H}\boldsymbol{e}$ coincide for the first n components; in fact

$$H\boldsymbol{e} = \begin{bmatrix} V\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}} \\ n \end{bmatrix}, \quad \widetilde{H}\boldsymbol{e} = \begin{bmatrix} V\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}} \\ 1 \\ n+1 \end{bmatrix},$$

where $\hat{\boldsymbol{e}}$ is the unit n-vector. Denoting by D the $n\times n$ diagonal matrix whose entries are the first n entries of $H\boldsymbol{e}$, we have

$$P = \left[\begin{array}{c|c} D^{-1}V & D^{-1}\hat{\boldsymbol{e}} \\ \hline \frac{1}{n}\hat{\boldsymbol{e}}^T & 0 \end{array}\right], \quad \widetilde{P} = \left[\begin{array}{c|c|c} D^{-1}V & \boldsymbol{0} & D^{-1}\hat{\boldsymbol{e}} \\ \hline \boldsymbol{0} & 0 & 1 \\ \hline \frac{1}{n+1}\hat{\boldsymbol{e}}^T & \frac{1}{n+1} & 0 \end{array}\right].$$

It is a simple matter to verify that the vector in equation (2.6) is indeed the Perron vector of $\widetilde{P}$, that is $\widetilde{\boldsymbol{p}}^T = \widetilde{\boldsymbol{p}}^T\widetilde{P}$. □

Now, suppose that the new added paper has a citation to paper $s \le n$. The new adjacency matrix is obtained by setting $\widetilde{h}_{n+1,s} = 1$ in the matrix $\widetilde{H}$. Let us denote by $\widehat{H}$ the matrix obtained in this way and by $\widehat{P} = \mathrm{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$ the stochastic matrix obtained by scaling the rows of $\widehat{H}$. Applying Theorem 2.3 to $\widetilde{H}$ and to $\widehat{H}$ enables us to prove the following result.

THEOREM 2.5. *For the Perron vectors $\boldsymbol{p}$ and $\widehat{\boldsymbol{p}}$ of the matrices $P = \mathrm{diag}(H\boldsymbol{e})^{-1}H$ and $\widehat{P} = \mathrm{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$, respectively,*

$$\frac{\widehat{p}_j}{p_j} \le \frac{\widehat{p}_s}{p_s}, \quad j = 1, \dots, n.$$

Moreover, $\widehat{p}_s/p_s > 1 + \frac{2}{n}p_{n+1}$.

Proof. From Theorem 2.3 applied to $\widetilde{P}$ and $\widehat{P}$, with $r = n+1$, one obtains $\widehat{p}_j/\widetilde{p}_j \le \widehat{p}_s/\widetilde{p}_s$, $j = 1,\dots,n$. The theorem holds since by Lemma 2.4 $\widetilde{p}_i = p_i/(1+\frac{2}{n}p_{n+1})$, $i = 1,\dots,n$. Using (2.3) we obtain that $\widehat{p}_s/p_s > 1+\frac{2}{n}p_{n+1}$. □

The above theorem says that if we introduce a new paper which contains a citation to paper s, then paper s has an increase of importance which is greater than or equal to the increase of importance for the other papers. Perturbation analysis of the Perron vector for a stochastic irreducible matrix has been addressed in [13] with a specific attention to PageRank. However, our results have a different flavor since we are interested in the rank index of the subjects rather than in the values of the entries of the eigenvector.

When we have a new paper that cites d papers, it follows, from Theorem 2.2 that *at least* one of the cited papers will have an increase of importance greater than that of the other papers. However, we cannot say that *all* the cited papers will increase their rank more than the non-cited ones.

3. Two-class model.

Consider the case when besides papers we also would like to rank authors. We can do this in an integrated model in which a paper, besides giving importance to the papers that it cites, gives importance to its authors, and in which an author gives importance to the papers that he/she has written and to his/her co-authors. This approach is similar to Kleinberg's idea [12] of Hub and Authorities for ranking Web pages, which can be reformulated in terms of a symmetric block matrix as described in [3].

As in the one-class model, we require that the amount of importance given by each subject to all the others is equal to the importance of the subject itself. That is, importance is neither destroyed nor created. This gives rise to row-stochastic nonnegative matrices.

Assume we have m authors numbered from 1 to m. Besides the adjacency matrix H concerning paper citations, we introduce the $m\times n$ matrix $K = (k_{i,j})$ concerning authorship, such that $k_{i,j} = 1$ if the author i is (co)author of the paper j; $k_{i,j} = 0$ otherwise. Define the matrix $A = KK^T = (a_{i,j})$. By simple inspection, it turns out that $a_{i,j}$ is the number of papers which are co-authored by authors i and j.

Observe that, by definition, any author has at least one paper, so that the matrix K cannot have null rows and it can be made row-stochastic. As in the case of the one-class model, the matrix H might have null rows. Therefore we proceed as we did in Section 2 by introducing a dummy paper with the same features as before. In addition, we assume that this paper is co-authored by all the existing authors. The introduction of this new paper favors neither any specific author nor any specific paper.

Now, let us still denote with n the number of papers, including the dummy paper, and introduce the $(m+n)\times(m+n)$ matrix S, which collects the information about citation and co-authorship:

$$S = \begin{bmatrix} KK^T & K \\ K^T & H \end{bmatrix},$$

where H is the $n\times n$ adjacency matrix of papers introduced in Section 2, and the $m\times n$ matrix K contains the information about the co-authorship. Recall that, due to the dummy paper, the last column of K is made up of all ones, i.e., all the authors are co-authors of the dummy paper. Moreover, the matrix S is irreducible.

The role of K^T in the lower left block of S is that, for $i > m$ and $j \le m$, $s_{i,j}$ is an entry of K^T, and this entry is nonzero if and only if the corresponding paper $i-m$ has author j as (co)author. In other words, the matrix S captures the relationships of authorship and citation among the different subjects (authors and papers) of this model, so that $s_{i,j} = 0$ if there exists no relationship between subject i and subject j. The kind of relationship, i.e., either citation

ETNA
Kent State University
etna@mcs.kent.edu

or authorship, is determined by the kind of classes the subjects i and j belong to.

EXAMPLE 3.1. Consider Example 2.1 when four different authors are added with the following authorship: author 1 has written papers 1 and 4, author 2 has written papers 2 and 4, author 3 has written papers 3 and 4, author 4 has written papers 5 and 6. In this way, the matrix K is given by

$$K = \left[\begin{array}{cccccc|c} 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right]$$

including the dummy paper, and the full matrix S is

$$S = \left[\begin{array}{cccc|ccccccc} 3 & 2 & 2 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 2 & 3 & 2 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 2 & 2 & 3 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}\right].$$

The matrix S is a generalized adjacency matrix in which no block (KK^T, K, K^T, or H) can have null rows. It is a simple matter to scale S by rows in order to obtain a stochastic matrix to be used as weight matrix for distributing the importance from one subject to another by means of citation or authorship. However, due to the different nature of the two classes *Authors* and *Papers*, it is more appropriate to make each one of the four blocks stochastic and to use suitable parameters to tune the influence of authorship over the influence of citations.

Proceeding as in Section 2, we scale the rows of the four blocks so as to obtain four stochastic matrices, and use the entries of such matrices to weight the amount of importance that each subject belonging to either the class *Papers* or to the class *Authors* provides to the other subjects. More precisely, define

$$Q = \begin{bmatrix} \operatorname{diag}(A\boldsymbol{e})^{-1}A & \operatorname{diag}(K\boldsymbol{e})^{-1}K \\ \operatorname{diag}(K^T\boldsymbol{e})^{-1}K^T & \operatorname{diag}(H\boldsymbol{e})^{-1}H \end{bmatrix},$$

where $A = KK^T$ and the symbol $\boldsymbol{e}$ denotes the vector of all ones of dimension m, n or $m+n$ depending on the context. Let $\Gamma = (\gamma_{i,j})$ be a 2×2 row-stochastic matrix, and consider the matrix

$$P = Q \odot \Gamma = \begin{bmatrix} \gamma_{1,1} \operatorname{diag}(A\boldsymbol{e})^{-1}A & \gamma_{1,2} \operatorname{diag}(K\boldsymbol{e})^{-1}K \\ \gamma_{2,1} \operatorname{diag}(K^T\boldsymbol{e})^{-1}K^T & \gamma_{2,2} \operatorname{diag}(H\boldsymbol{e})^{-1}H \end{bmatrix}. \tag{3.1}$$

For the sake of notational simplicity, given a $q \times q$ block matrix $A = (A_{i,j})$ and a $q \times q$ matrix $B = (b_{i,j})$ we denote by $A \odot B$ the $q \times q$ block matrix having blocks $b_{i,j}A_{i,j}$. We have the following result.

PROPOSITION 3.2. *For $A = (A_{i,j})_{i,j=1,n}$, where $A_{i,j}$, $i, j = 1, n$ are row stochastic, and for a row-stochastic matrix $B = (b_{i,j})_{i,j=1,n}$, it holds that $A \odot B$ is row-stochastic.*

ETNA
Kent State University
etna@mcs.kent.edu

Proof. One has

$$A \odot Be = \begin{bmatrix} \sum_i b_{1,i} A_{1,i} \boldsymbol{e} \\ \vdots \\ \sum_i b_{n,1} A_{n,i} \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \sum_i b_{1,i} \boldsymbol{e} \\ \vdots \\ \sum_i b_{n,i} \boldsymbol{e} \end{bmatrix} = \boldsymbol{e}. \qquad \square$$

In particular, the matrix P in (3.1) is row-stochastic. In this way we can define our model by means of the eigenvalue equation

$$\boldsymbol{p}^T = \boldsymbol{p}^T P \tag{3.2}$$

with P being the matrix in (3.1), where in the vector $\boldsymbol{p}$ the first m components describe the importance of the authors, while the remaining components describe the importance of the papers.

Equation (3.2) states that the importance of a paper is the sum of the importances given by the authors of the paper, weighted by the factor $\gamma_{1,2}$, plus the importance given by the citations received by the paper, weighted by the factor $\gamma_{2,2}$. Similarly, the importance of an author is the sum of the importances received by the co-authors, weighted with the factor $\gamma_{1,1}$, plus the importances received by the papers that he/she has written, weighted by the factor $\gamma_{2,1}$. More precisely, we have

$$\begin{aligned} p_j &= \gamma_{1,1} \sum_{i=1}^{m} p_i \frac{a_{i,j}}{\sum_t a_{i,t}} + \gamma_{2,1} \sum_{i=1}^{n} p_{m+i} \frac{k_{j,i}}{\sum_t k_{t,i}}, \qquad j = 1, \ldots, m, \qquad \text{authors} \\ p_j &= \gamma_{1,2} \sum_{i=1}^{m} p_i \frac{k_{i,j}}{\sum_t ki,t} + \gamma_{2,2} \sum_{i=1}^{n} p_{m+i} \frac{h_{i,j}}{\sum_t h_{i,t}}, \qquad j = m+1, \ldots, n, \qquad \text{papers.} \end{aligned}$$

Since Γ is full, it is obviously irreducible, and so P is also irreducible and the vector $\boldsymbol{p}$, normalized so that $\sum p_i = 1$, exists and is unique. Observe also that since it is meaningless to compare subjects of different classes, namely authors and papers, the normalization of $\boldsymbol{p}$ is still meaningful if restricted separately to the subvector containing the first m components and the subvector containing the remaining n components.

It is interesting to point out that, denoting by $\mu_A = \sum_{i=1}^{m} p_i$ and $\mu_P = \sum_{i=1}^{n} p_{m+i}$ the overall amount of the importance of authors and of papers, respectively, the vector (μ_A, μ_P) is a left eigenvector of Γ corresponding to the eigenvalue 1. Moreover, if we replace the matrix Γ by $\Gamma' = D\Gamma D^{-1}$, where D is any nonsingular diagonal matrix, then the left Perron vector $\boldsymbol{p}'$ of $P' = Q \odot \Gamma'$ differs from $\boldsymbol{p}$ only by scaling factors depending on the values of μ_A and μ_P. Therefore, in order to evaluate separately the subvectors of $\boldsymbol{p}$ related to authors and papers, respectively, it is enough to consider a matrix Γ of the kind

$$\begin{bmatrix} 1-\alpha & \alpha/\theta \\ \beta\theta & 1-\beta \end{bmatrix}$$

for α and β suitable scalars in $[0, 1]$ and $\theta > 0$ any arbitrary constant. In particular, we may chose $\theta = \alpha/\beta$, which makes Γ column-stochastic, or $\alpha = \sqrt{\alpha/\beta}$ which makes Γ symmetric.

The parameters $\gamma_{i,j}$ determine the amount of influence that each class has on the other classes. In particular, choosing $\Gamma = I$ provides an uncoupled problem where the matrix P is block diagonal. In this case, the ranking of papers is independent of that of authors and coincides with the ranking obtained in the one-class model of Section 2.

In this special case, the authors receive importance only from authorship and not from the importance of their papers.

ETNA
Kent State University
etna@mcs.kent.edu

We observe that this model has an annoying drawback; namely, the importance received by a paper from its co-authors is proportional to the number of co-authors. The more co-authors, the more importance they receive through the authorship. In this way, a paper having many authors might be more important than a paper having a single author even though the former has many fewer citations. This drawback, which is illustrated in the next example, in principle could be removed by normalizing the block $(1, 2)$ of P by columns. This corresponds to evaluating the importance received by the co-authorship as the *mean*, instead of the sum, of the importances brought to the paper by the co-authors. Note that the block $\widehat{K} = K\,\mathrm{diag}(K^T\boldsymbol{e})^{-1}$ that we would obtain by normalizing the matrix K this way is no longer stochastic.

One would think that the column normalization followed by a row normalization is enough to get a row-stochastic matrix in which the importance that a paper receives from its authors is the average of the importances of the authors. Unfortunately, this is false in general. Consider, for instance, a matrix K in which the first column has the first q entries equal to one and the remaining entries zero, and in which the first q rows vanish except for their first and last entries. The column normalization transforms the ones in the first column into $1/q$ and those in the last column into $1/m$. But the subsequent row normalization turns the entries in the first column into $m/(q+m)$ and the ones in the last column into $q/(m+q)$. For instance, if $q = m/2$ then each one of the first q authors would give $2/3$ of its importance to the first paper instead of $1/q$ as we desired.

Therefore, after the normalization by columns has been performed, we have to apply a slightly modified row-normalization. More precisely, row-normalization is performed only if the row sum of the entries is greater than or equal to one. Otherwise, if the row sum is less than one, we leave the entries unchanged except for the entry corresponding to the dummy paper, which is changed in such a way that the row sum is one.

This simple normalization, which generates a row-stochastic block $\widetilde{K}$, is described in the following

ALGORITHM 1. For each $i \in \{1, \ldots, m\}$, compute $s_i = \sum_{j=1}^{n} \widehat{k}_{i,j}$.
If $s_i \leq 1$, set $\widetilde{k}_{i,j} = \widehat{k}_{i,j}$, for $j = 1, \ldots, n-1$, and $\widetilde{k}_{i,n} = 1 - \sum_{j=1}^{n-1} \widehat{k}_{i,j}$.
Else, divide the ith row of $\widetilde{K}$ by the sum of its entries, that is, set

$$\widetilde{k}_{i,j} = \widehat{k}_{i,j}/s_i.$$

Output $\widetilde{K} = (\widetilde{k}_{i,j})$.

We may immediately verify that for $s_i = 1$ the two different normalizations described in the above algorithm provide the same result. Observe also that the normalization for $s_i \leq 1$ leaves unchanged the amount of importance that author i yields to papers j, $j = 1, \ldots, n-1$, after the column scaling and assigns to the dummy paper the remaining amount of importance that is missing.

Row normalization and Algorithm 1 yield the following matrix

$$P = \begin{bmatrix} \gamma_{1,1}\ \mathrm{diag}(A\boldsymbol{e})^{-1}A & \gamma_{1,2}\ \widetilde{K} \\ \gamma_{2,1}\ \mathrm{diag}(K^T\boldsymbol{e})^{-1}K^T & \gamma_{2,2}\ \mathrm{diag}(H\boldsymbol{e})^{-1}H \end{bmatrix}, \tag{3.3}$$

where the matrix $\widetilde{K}$ is obtained by means of Algorithm 1.

EXAMPLE 3.3. In order to understand the different normalization of block (1,2) in the two models, let us consider the case of Example 3.1 with weights $\gamma_{i,j} = 1/2$, $i, j = 1, 2$. Computing the Perron vector in the model described in (3.1), we have that the first four components of the left Perron vector of the matrix P (the ones corresponding to authors) are given by

$$(0.238912,\ 0.238912,\ 0.238912,\ 0.283265);$$

they are normalized to sum to 1. The remaining seven components (the ones corresponding to papers) are

$$(0.0778083,\ 0.0778083,\ 0.0778083,\ 0.176898,\ 0.104652,\ 0.145862,\ 0.339163);$$

they also are normalized to sum to 1. Observe that the first three authors have the same rank, while the fourth author has higher rank. In fact, he/she is the author of two important papers which confer importance to him/her. Moreover, the first three papers still keep the same rank as in the one-class model, but the fourth and the fifth papers have different ranks. In particular, the fourth paper reaches the maximum rank followed by paper 6 and 5. The reason is that paper 4 has many authors and then it accumulates importance from all the authors.

By following the model described in (3.3), in which the average of the importances of the authors is considered instead of their sum, one obtains

$$(0.237763, 0.237763, 0.237763, 0.28671)$$

for authors and

$$(0.11009, 0.11009, 0.11009, 0.137613, 0.126243, 0.150923, 0.25495)$$

for papers. This time, as one would expect, paper 6 is the one with the highest rank, while paper 4 is more important than paper 5. The fourth author still has a higher rank than the remaining authors.

The following example shows that on a basis of equivalent papers, an author with more papers is more important.

EXAMPLE 3.4. Consider the simple case of three papers with a cyclic graph of citations as shown below.

$$\begin{tikzcd} 3 \arrow[rr, bend left] & 2 \arrow[l] & 1 \arrow[l] \end{tikzcd}$$

Each paper has a single citation and the adjacency citation matrix H is given by

$$H = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 \end{array}\right]$$

including the dummy paper. In the one-class model, the three papers have the same importance. In fact, the computed vector p, including the dummy component, is given by

$$\boldsymbol{p}^T = (0.222222,\ 0.22222,\ 0.222222,\ 0.333333).$$

In the two-class model, assuming that there are three authors and that each paper has a single different author, the matrix K is given by

$$K = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

ETNA
Kent State University
etna@mcs.kent.edu

and the matrix $A = KK^T$ is

$$KK^T = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

The computed Perron vector with weights $\gamma_{i,j} = 1/2$ is

$$(0.333333,\ 0.333333,\ 0.333333)$$

for authors and

$$(0.233333,\ 0.233333,\ 0.233333,\ 0.3)$$

for papers. We can see that all the papers, except for the dummy, as well all the authors, have the same rank.

Now assume that there are three authors; author i is author of paper i for $i = 1, 2, 3$. Moreover, author 1 is also co-author of paper 3. In the two-class model, author 1 is expected to receive more importance than the other authors since he/she has written more papers of roughly the same rank. This implies that also his/her two papers should slightly increase their importance. With this data the matrices K and A are given by

$$K = \left[\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right] \quad A = KK^T = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}.$$

In fact, with $\gamma_{i,j} = 1/2$, the computed vector $\boldsymbol{p}$ in the part concerning authors is

$$(0.423170, 0.302289, 0.274541),$$

and in the part concerning papers, dummy paper included, it is

$$(0.226729, 0.222693, 0.234666, 0.315913).$$

Author 1 has increased his/her importance together with the importance of the papers co-authored by him/her. This confirms the consistency of our model.

EXAMPLE 3.5. Consider the situation in Example 2.1 and assume that author i is (co)author of paper i for $i = 1, 2, 3, 4, 5, 6$, while author 6 is co-author of paper 1. From the graph of citations, we expect that paper 6 is the most important (and this is true in the one-class model). Further, we expect that author 6 has higher rank and that he/she raises the rank of paper 1 of which he/she is co-author. In this case, the matrices K and $A = KK^T$ are given by

$$K = \left[\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right], \quad KK^T = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 & 1 \\ 2 & 1 & 1 & 1 & 1 & 3 \end{bmatrix}.$$

Computations performed with $\gamma_{i,j} = 1/2$ shows that the author components of vector $\boldsymbol{p}$ are

$$(0.139891,\ 0.148972,\ 0.147190,\ 0.166273,\ 0.166273,\ 0.231402)$$

and the paper components are

$$(0.120885,\ 0.100520,\ 0.0972106,\ 0.132651,\ 0.132651,\ 0.157310,\ 0.258772).$$

Once again the result of the computation confirms the consistency of the model.

REMARK 3.6. *It is possible to show that Theorems 2.3 and 2.5 still hold for the matrix P defined by (3.3) if the perturbation concerns an entry in the lower right block of P.*

4. Three-class model. Besides the classes of *Papers* and *Authors*, we introduce the class of *Journals* of cardinality q, and we number the elements of this set from 1 to q. Together with the matrices H and K, we consider the matrix $F = (f_{i,j})$, with $f_{i,j} = 1$ if journal i publishes paper j and $f_{i,j} = 0$ otherwise, and the matrix $G = (g_{i,j})$ such that $g_{i,j} = r$ if the author j has published r papers in the journal i. Similarly define the matrix $E = (e_{i,j})$ such that $e_{i,j}$ is the number of citations from papers published in journal i to papers published in journal j. Direct inspection shows that

$$E = FHF^T, \quad G = FK^T.$$

The full adjacency matrix, which collects all the information about citation, authorship and publications, is given by

$$S = \begin{bmatrix} E & G & F \\ G^T & A & K \\ F^T & K^T & H \end{bmatrix}. \tag{4.1}$$

Similarly to the two-case model, S synthesizes the relationship between the different subjects of our model (journals, authors, and papers) in such a way that $s_{i,j} \neq 0$ if there exists a relationship between subject i and subject j. The kind of relationship depends on the pair of classes which the subjects i and j belong to.

Also, in this case we normalize each block of S by scaling its rows so as to obtain stochastic matrices, and we use a 3×3 stochastic matrix of parameters to better tune the influence of one class on the other ones. In order to achieve this, we require that no block has an entire row of zeros. This was avoided in the previous models by introducing a dummy paper. Here, we can proceed similarly. Since we have to avoid creating privileges among the subjects, we may proceed in two different ways. Either we assume that the dummy paper is published by all the journals, or that there exists a dummy journal which publishes only the dummy paper. With these two choices we get two different models represented by two suitable modifications of the matrix S of (4.1). The rank vector is the Perron vector of a suitable modification of S. The analysis of these models is left for future work.

5. Numerical tests. We tested the approaches discussed in previous sections using the CiteSeer dataset, which can be freely downloaded from the CiteSeer web site [6]. CiteSeer is a scientific literature digital library and search engine that focuses primarily on the literature in computer and information science [10]. CiteSeer crawls and gathers academic and scientific documents on the web and uses autonomous citation indexing to permit querying by citation or by document and then ranking them by citation impact.

For our experiments we used the CiteSeer index downloaded on June 2007 consisting of about 800,000 papers. This dataset was first cleaned to remove some incorrect references, such as items without an author or isolated items. We obtained a dataset consisting of approximately 250,000 authors and 350,000 papers in XML format. The data was then parsed to produce the matrices H and K.

Despite the fact that every item in the XML format contains much information, it is not easy to recover the journal where the paper was published, in part because the journals are not

Paper	pos.	cit.
Diffie, Hellman - New Directions in Cryptography	31	553
Rivest, Shamir, Adleman - Public Key Cryptography	3	1218
Bryant - Boolean Functions Manipulation, BDD	1	1636
Kirkpatrick, Gelatt, Vecchi - Simulated Annealing	2	1337
Floyd, Jacobson - TCP/IP Protocol	4	1125
Canny - Computational Approach to Edge Detection	10	834

TABLE 5.1

Experimental results for the one-class model. The top papers in our model are listed in decreasing rank order. In the first column, papers are identified by their authors and title. The second column contains the position of the paper in a list ordered by decreasing number of citations received, and the third column gives the number of citations the paper received.

Author	num. cit	num. pap.	av. num. cit.
Randal Bryant	2615	83	31.5
Sally Floyd	4950	91	54.4
John K. Ousterhout	2214	23	96.3
Luca Cardelli	2112	91	23.2
Van Jacobson	4719	40	118.0
Rakesh Agrawal	4745	83	57.2
Jack J. Dongarra	2799	291	9.6
Raj Jain	1038	116	8.9
Douglas C. Schmidt	2980	329	9.1
Vern Paxson	2735	66	41.4
John McCarthy	911	41	22.2
Thomas A. Henzinger	3694	176	21.0

TABLE 5.2

Experimental results for the two-class model for the subject Author. In the first column the top authors are listed in decreasing order of rank. In the remaining columns we report the number of citation received, the number of papers by the author that are indexed in the dataset, and the average number of citation per paper.

associated with a unique identifier. This means that with these data we were not able to test the effectiveness of our three-class model. However, experimental results on the MR [1] dataset[1], indicates that our journal rankings also capture concepts such as prestige and authority.

We present the results of two different numerical tests. The first test addresses the problem of the ranking of papers by using the one-class model. The results, reported in Table 5.1 shows the top six papers obtained with our model. We can recognize among these papers great pieces of work such as fundamental papers in cryptography; the paper by Bryant introducing the binary decision diagram (BDD), a data structure for describing Boolean functions; or the paper in which the TCP/IP protocol is been proposed.

We see that the position occupied in our ranking by these papers, does not coincide with that occupied by simply sorting the papers by descending number of citations received. This is due to the fact that here, not all the citations are regarded the same, but citation by important papers have a greater weight. For example it is possible to see that the paper by Diffie and Hellman is contained in the reference list of the paper by Rivest, Shamir and Adleman, and hence it gets a higher rank even if it receives fewer citations.

In Table 5.2 we report the top authors obtained by choosing uniform weights. We can

[1]The AMS denied us the authorization to publish results obtained using part of their index.

ETNA
Kent State University
etna@mcs.kent.edu

Paper	pos.	cit.
Kirkpatrick, Gelatt, Vecchi - Simulated Annealing	2	1337
Bryant - Boolean Functions Manipulation, BDD	1	1636
Rivest, Shamir, Adleman - Public Key Cryptography	3	1218
Canny - Computational Approach to Edge Detection	10	834
Floyd, Jacobson - TCP/IP Protocol	4	1125
Diffie, Hellman - New Directions in Cryptography	31	553
John K. Ousterhout - Tcl and the Tk Toolkit	8	913
Harel - Statecharts Formalism	6	1042
Elman - Neural Networks	26	589
Jones - Vienna Development Method	23	609

TABLE 5.3

Experimental results for the two-class model for the subject Paper. In the first column, papers are identified by their authors and title. The papers are listed in decreasing order of rank according to our model. The second column contains the position in the list ordered by decreasing number of citations, and the third column contains the number of citations received by the paper.

recognize very important computer scientists who wrote important papers in many areas of information science. Some of the authors in the list rank higher than one would expect from looking at the number of papers written, mainly because they have important co-authors. However, we can smooth the effect of co-authorship by reducing the corresponding coefficient in the weight matrix.

In Table 5.3 we report the results for the subject *Paper* obtained with uniform weights. The differences with Table 5.1 are essentially in the order of the best papers, that in the two-class model are influenced also by the authority of the authors.

6. Conclusions and open problems. We proposed integrated models for evaluating papers, authors, and journals based on citations, co-authorship and publications. After the one-class model for ranking scientific publications, we introduced the two-class model which ranks papers and authors, and the three-class model for ranking papers, authors, and journals. In all the models, the rank vector is the Perron vector of an irreducible stochastic matrix.

Some theoretical results have been proved concerning the variation of the Perron vector of an irreducible stochastic matrix under limited changes of its entries. These results prove that the models behave as one would expect when a new citation occurs.

Simple examples show that our model is better suited for ranking scientific publications than available models based only on the number of citations.

Some open issues remain to be analyzed. A theoretical issue concerns perturbation theorems. In Section 2 we proved that if a paper receives a new citation, then its rank increases more than the rank of the other papers. It would be natural to guess that if more than one paper receives a citation, then *all* the cited papers increase their importance more than the other papers. At the moment, a proof of this property is missing and no counterexample is known. We plan to address this problem in our future work.

A second issue which deserves attention is related to the “static” nature of our model. That is, the time of publication of the papers or the time a citation is received do not play any role in our model. However, it is commonly accepted that a recent paper and an old paper that receive the same number of citations should not have the same rank, since the old paper has had more time to gather citations than the newer one. We are currently investigating this issue trying to insert the factor “time” in our model. Our idea is to study the evolution of the importance of a paper, an author or a journal over the time.

Acknowledgments. The authors are indebted to the two referees whose comments and suggestions helped improve the quality of the presentation.

REFERENCES

[1] AMS, *MathSciNet, Mathematical Reviews on the Web*. http://www.ams.org/mathscinet/.
[2] C. T. BERGSTROM, *Eigenfactor: Measuring the value of prestige of scholarly journals*, C&RL News, 68 (2007).
[3] V. BLONDEL, A. GAJARDO, M. HEYMANS, P. SENNELART, AND P. V. DOOREN, *A measure of similarity between graph vertices : applications to synonym extraction and web searching*, SIAM Rev., 46 (2004), pp. 647–666.
[4] J. BOLLAN, M. A. RODRIGUEZ, AND H. V. DE SOMPEL, *Journal status*, Scientometrics, 69 (2006), pp. 669–687.
[5] S. BRIN AND L. PAGE, *The anatomy of a large-scale hypertextual Web search engine*, Comput. Networks ISDN Systems, 30 (1998), pp. 107–117.
[6] CITESEER.IST, *Computer and Information Science Papers. CiteSeer Publications ReserchIndex*. http://citeseer.ist.psu.edu/.
[7] E. DIETZENBACHER, *Perturbation of matrices: A theorem on the Perron vector and its applications to input-output models*, J. Econom., 48 (1988), pp. 389–412.
[8] L. ELSNER, C. R. JOHNSON, AND M. NEUMANN, *On the effect of the perturbation of a nonnegative matrix on its Perron eigenvector*, Czechoslovak Math. J., 32 (1982), pp. 99–109.
[9] E. GARFIELD, *Citation analysis as a tool in journal evaluation*, Science, 178 (1972), p. 471.
[10] C. L. GILES, K. BOLLACKER, AND S. LAWRENCE, *CiteSeer: An automatic citation indexing system*, in Digital Libraries 98 - The Third ACM Conference on Digital Libraries, I. Witten, R. Akscyn, and F. M. Shipman III, eds., Pittsburgh, PA, June 23–26 1998, ACM Press, pp. 89–98.
[11] J. E. HIRSH, *An index to quantify an individual's scientific research output*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 16569–16572.
[12] J. M. KLEINBERG, *Authoritative sources in a hyperlinked environment*, J. ACM, 46 (1999), pp. 604–632.
[13] A. N. LANGVILLE AND C. D. MEYER, *Updating Markov chains with an eye on Google's PageRank*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 968–987.
[14] L. I. MEHO AND K. YANG, *Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar*, J. Am. Soc. Inform. Sci. Tech., 58 (2007), pp. 2105–2125.
[15] I. PALACIOS-HUERTA AND O. VOLIJ, *The measurement of intellectual influence*, Econometrica, 72 (2004), pp. 963–977.
[16] G. PINSKI AND F. NARIN, *Citation influence for journal aggregates of scientific publications- theory, with applications to literature of physics*, Inform. Process. Manag., 12 (1976), pp. 297–312.
[17] G. TAUBES, *Measure for measure in science*, Science, 260 (1993), pp. 884–886.
[18] R. J. W. TIJSSEN, M. S. VISSER, AND T. N. VAN LEEUWEN, *Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference?*, Scientometrics, 54 (2002), pp. 381–397.