Università di Pisa
Dipartimento di Informatica

Technical Report: TR-08-10

A combined approach for evaluating papers, authors and scientific journals

Dario A. Bini Gianna M. Del Corso F. Romani

May 30, 2008

Address: Largo B. Pontecorvo 3, 56127 Pisa, Italy. Tel: +39 050 2212700 Fax: +39 050 2212726

A COMBINED APPROACH FOR EVALUATING PAPERS, AUTHORS AND SCIENTIFIC JOURNALS

DARIO A. BINI*, GIANNA M. DEL CORSO†, AND F. ROMANI †

Abstract. An integrated model for ranking scientific publications together with authors and journals recently presented in [Bini, Del Corso, Romani, ETNA 2008] is closely analyzed. The model, which relies on certain adjacency matrices H, K and F obtained from the relations of citation, authorship and publication, provides the ranking by means of the Perron vector of a stochastic matrix obtained by combining H, K and F. Some perturbation theorems concerning the Perron vector previously introduced by the authors are extended to more general cases and a counterexample to a property previously addressed by the authors is presented. The theoretical results confirm the consistency and effectiveness of our model. Some paradigmatic examples are reported together with some results obtained on a real set of data.

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

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 for evaluating the quality of a scientific publication are based on the analysis of the citations received.

Recently, in [3] an integrated model was proposed where publications and authors are ranked based on the relationship of citation and (co-)authorship. The main idea is that the rank of a subject is given by the weighted sum of the ranks of the other subjects which are related to it by means of either citation or authorship. So, for instance, a paper receives importance by the papers that cite it, as well as by the authors who have written it. Similarly, an author receives importance by the papers that he/she has written, as well as by his/her coauthors. A more general model based on three classes, *Authors*, *Papers* and *Journals*, is also outlined and some problems are addressed.

In this paper we further investigate the three-class model outlined in [3], and discuss about the possible normalization techniques which make the behavior of the model closer to the desired behavior of the evaluation system still preserving the overall mathematical consistency.

We extend the perturbation results presented in [3] and provide some examples which give more insights to the properties of the Perron eigenvector which represents the ranking of the subjects involved in the evaluation system.

The results of some numerical experiments performed with the three-class model are reported. We performed tests with real and with synthetic data. The experiments showed that our model is robust to spam, in the sense that it measures quality without favoring journals which publish too many papers or authors who write many low quality papers. The rank of a journal turns out to be independent of the number of papers that it publishes being closer to the average quality of its papers.

The paper is organized as follows. In Section 2 we formalize the model by showing that the desired rank is the Perron vector of a suitable 3×3 block matrix. Then we discuss about different normalizations of each block that, still preserving the row-stochasticity of the matrix, provide a model which is consistent with the real requests of the evaluation system. A brief discussion on the role of time and the comparison

---

*Dipartimento di Matematica, Università di Pisa, Largo B. Pontecorvo 5, 56127 Pisa, Italy.
†Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy.

with the existing models, of which our proposal turns out to be a generalization, conclude the section.

In Section 3 we introduce some perturbation theorems that generalize the ones presented in [3] and provide a sort of theoretical validation of our model. In fact, the theorems confirm the expected behavior where a paper which receives new citations must increase its rank. By means of a counterexample we show that if more papers receive new citations, not necessarily all of them will increase their rank. This issue was addressed in [3] as an open problem.

Section 4 is devoted to numerical experiments; conclusions are drawn in Section 5.

2. The basic model.

Usually, the evaluation of the research performed by a scholar is done by analyzing the papers he/she has published. The quality of his/her papers is mainly measured by considering the notoriety of the journal where the papers appear; sometimes also the citations count is taken into account as criterion. In this section we elaborate the integrated three-class model introduced in [3] where papers, authors and journals mutually contribute to the attribution of a ranking score of each others.

The idea is that in order to evaluate an author we have to consider not only the quality of the journals where his/her papers have been published but also the quality of every single paper of this author. Moreover, also the quality of the co-authors must be taken into account. In fact, an important author who writes a joint paper with a less important one, expresses a sort of trusting vote by conferring to that author more visibility with respect to the international community. Similarly, to evaluate the quality of a paper one has to look at the quality of the journal where the paper is published, at the citations received and at the prestige of its authors. Also when evaluating a journal we take into account not only the cross-citations among journals as done by many methods such as Impact Factor [7], Eigenfactor [2], and many others [4, 11], but also the quality of every single paper published there and the authoritativeness of the scholars writing on that journal.

Let us formalize this idea in mathematical terms.

Assume we are given n_P papers together with the bibliographic data of each paper. In particular, of each paper we know the authors, the journal where the paper is published and the list of citations contained in the paper. With this information we can construct three matrices: the matrix F accounting for the journal publishing each paper, the matrix K which stores information about authorship and the matrix H which records the citation structure among papers. In particular, let n_J be the total number of distinct journals where the n_P papers are published, and let n_A the number of distinct authors who authored the given n_P papers. We find that $F = (f_{i,j})$ is an $n_J \times n_P$ binary matrix where

$$f_{i,j} = \begin{cases} 1 & \text{if paper } j \text{ is published in journal } i \\ 0 & \text{otherwise,} \end{cases}$$

$K = (k_{i,j})$ is an $n_A \times n_P$ binary matrix such that

$$k_{i,j} = \begin{cases} 1 & \text{if author } i \text{ has written paper } j \\ 0 & \text{otherwise,} \end{cases}$$

and $H = (h_{i,j})$ is an $n_P \times n_P$ matrix such that

$$h_{i,j} = \begin{cases} 1 & \text{if paper } i \text{ has paper } j \text{ in its reference list} \\ 0 & \text{otherwise.} \end{cases}$$

We can combine these three matrices to obtain the following 3×3 block matrix

$$A = \begin{bmatrix} FHF^T & FK^T & F \\ KF^T & KK^T & K \\ F^T & K^T & H \end{bmatrix} \tag{2.1}$$

of size $N = n_J + n_A + n_P$.

Each block of this matrix expresses the relationship between the subjects belonging to the three classes *Journals*, *Authors* and *Papers*. More specifically, the entry in position (i, j) of the block FHF^T contains the number of citations that the papers published in journal j receives from the papers published in journal i; the entry in position (i, j) of the block FK^T contains the number of papers that author i has published in journal j; the entry in position (i, j) of the block KK^T contains the number of joint papers that author i has in collaboration with author j.

We can scale the rows of A to obtain a row-stochastic matrix $P = \mathrm{Diag}(\boldsymbol{d})^{-1}A$, where $\boldsymbol{d} = A\boldsymbol{e}$, and $\boldsymbol{e} = (1, \ldots, 1)^T$, provided that $\boldsymbol{d}$ has no null component. In this way, the entries of $P = (p_{i,j})$ can be used as weights to transfer amounts of importance from a subject to another subject. More precisely, numbering the subjects from 1 to N, the importance π_j of subject j is the weighted sum of the importances π_i of all the other subjects i which are in relation with j, where the weights are $p_{i,j}$, that is

$$\pi_j = \sum_{i=1}^{N} \pi_i p_{i,j}.$$

This condition expresses the fact that $\boldsymbol{\pi} = (\pi_i)$ is eigenvector of P corresponding to the eigenvalue 1:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T P.$$

The row stochasticity of P implies that the overall amount of importance that a subject i transfers to the other subjects coincides with the importance of i. In other words, the amount of importance in the system is neither created nor destroyed.

To guarantee the existence of a unique vector $\boldsymbol{\pi}$, such that $\pi_i > 0$ and $\sum_i \pi_i = 1$, we need A to be irreducible. Under this condition, the vector $\boldsymbol{d}$ has nonzero components so that the matrix P can be constructed, and the Perron Frobenius theorem [9] guarantees the existence and the uniqueness of $\boldsymbol{\pi}$. Moreover, in order to have nice convergence properties of iterative algorithms for the computation of $\boldsymbol{\pi}$ we need A to be aperiodic.

There are many ways to enforce irreducibility, for example as done for the Google matrix [5]. A way to obtain an irreducible and aperiodic matrix which fits better in our model is to introduce a dummy paper, a dummy author, and a dummy journal, similarly to what done for the one-class model [3]. The dummy paper is cited by every paper and it cites back all the papers except itself. The dummy paper is written by the dummy author and is published in the dummy journal. Mathematically, this corresponds to consider the matrices $\widehat{H}, \widehat{K}$ and $\widehat{F}$ obtained from H, K and F as follows,

$$\widehat{H} = \left[\begin{array}{c|c} H & \boldsymbol{e} \\ \hline \boldsymbol{e}^T & 0 \end{array}\right], \quad \widehat{K} = \left[\begin{array}{c|c} K & \boldsymbol{0} \\ \hline \boldsymbol{0}^T & 1 \end{array}\right], \quad \widehat{F} = \left[\begin{array}{c|c} F & \boldsymbol{0} \\ \hline \boldsymbol{0}^T & 1 \end{array}\right],$$

and to replace H, K and F in (2.1) with $\widehat{H}, \widehat{K}$ and $\widehat{F}$, respectively.

Theorem 2.1. *The matrix $\widehat{A}$ obtained by replacing the blocks H, K and F in* (2.1) *with the blocks $\widehat{H}$, $\widehat{K}$, and $\widehat{F}$, respectively, is irreducible and aperiodic.*

Proof. In order to prove that $\widehat{A}$ is irreducible we have to show that the underlying graph is strongly connected. This is true since every paper is connected to its authors and to the journal publishing it and every paper is connected to every other paper trough the dummy paper. To prove that $\widehat{A}$ is aperiodic it is sufficient to prove that there are cycles of length 2 and 3, since the period of an irreducible nonnegative square matrix is the greatest common divisor of the lengths of the cycles (see [9], Ch. 9), and $\gcd(2,3) = 1$. This is always the case if $H \neq 0$ with the cycles $i \to \text{dummy} \to i$ of length 2 and $i \to j \to \text{dummy} \to i$ of length 3, where $i \to j$ corresponds to a nonzero (non-diagonal) entry of H, and dummy denotes the dummy paper. □

2.1. Row and column scaling. In the previous section we simply propose to scale the rows of A in order to obtain a row-stochastic matrix. A more flexible way, introduced in [3], consists in performing a separate normalization of the blocks of A. That is, each block of A is normalized to yield nine row-stochastic matrices; then these matrices are compounded with weights $\Gamma = (\gamma_{i,j})_{i,j=1,3}$, where Γ is row stochastic, into a new stochastic matrix. The entries of this new matrix are used to weight the amount of importance that each class (*Journal*, *Authors*, and *Papers*) gives to the other classes. In this section we discuss some issues related to the different kinds of normalization.

Denote by

$$Q = \begin{bmatrix} J_J & J_A & J_P \\ A_J & A_A & A_P \\ P_J & P_A & P_P \end{bmatrix}, \tag{2.2}$$

where each block is row-stochastic and is obtained from the corresponding block in the matrix $\widehat{A}$ of Theorem 3.1, so for example J_J is the stochastic matrix obtained by the row-normalization of $\widehat{F}\widehat{H}\widehat{F}^T$.

Here, the notation used in (2.2) points out the role of each block with respect to the classes *Journals*, *Authors* and *Papers*. For instance, the entries of the block J_A weight the amount of importance that *Journals* transfer to *Authors*.

Let $\Gamma = (\gamma_{i,j})$ be a 3×3 row-stochastic matrix, then the matrix

$$P = \begin{bmatrix} \gamma_{1,1}\, J_J & \gamma_{1,2}\, J_A & \gamma_{1,3}\, J_P \\ \gamma_{2,1}\, A_J & \gamma_{2,2}\, A_A & \gamma_{2,3}\, A_P \\ \gamma_{3,1}\, P_J & \gamma_{3,2}\, P_A & \gamma_{3,3}\, P_P \end{bmatrix}. \tag{2.3}$$

is row-stochastic and its entries $p_{i,j} \geq 0$ express the amount of importance that subject i transfers to subject j. The parameters $\gamma_{i,j}$ can be used to tune the role that each class has with respect to the other classes. For instance, choosing $\gamma_{3,3}$ greater than $\gamma_{2,3}$ and $\gamma_{1,3}$ means to base the importance of papers more on the citations that they receive rather than on the importance of their authors or of the journals where they are published.

Row normalization is not always well suited with the real model. In fact, row normalization of the block in position $(2,3)$, used to compute the influence of the authors on the papers, would imply that the importance received by a paper from its authors is the sum of the importances of the authors. In this way, papers coauthored by many authors would receive much more importance than papers authored by a single author having a comparable prestige. In this case, a column normalization of the block would be more suited since it corresponds to assign to a paper the average of the importances of its authors.

Similarly, in our model the importance of a journal should not depend on the number of papers published but on their quality. This means that we have to apply a column normalization to blocks $(1,1)$ and $(3,1)$ in order to assign to the journal the average instead of the sum of the importances of papers and of the authors who publish in this journal.

Unfortunately, column normalization of a block B does not provide a row-stochastic matrix. Row-stochasticity is required if we wish to respect the condition that each subject evenly distributes its importance to all the other subjects to which is related.

A solution to this problem proposed in [3] is the following. Assume that we are given an $m \times n$ block B where its last column concerns the dummy subject. Assume that B has been normalized in some way so that it is not stochastic anymore. In order to make B stochastic, still keeping almost unchanged the role of its entries as weights in the evaluation model, we apply the following normalization on the matrix B

$$s_i = \sum_{j=1}^{n} b_{i,j}, \qquad b_{i,j} \leftarrow \begin{cases} b_{i,j}/s_i, \ j = 1, \ldots, n & \text{if } s_i \geq 1 \\ b_{i,j}, \ j = 1, \ldots, n-1, \ \ b_{i,n} = 1 - s_i & \text{if } s_i < 1 \end{cases} \tag{2.4}$$

Observe that with this normalization, if the sum s of the entries in a row exceeds 1 then all the entries are divided by their sum. If the sum of the entries is less than 1, their values are left unchanged except for the dummy entry whose value is the amount needed to have a stochastic row. To motivate this strategy, consider the case where a row has only one nonzero entry with a small value. The customary row normalization would turn this entry to 1, our normalization would leave this value unchanged.

Now we separately examine each block of the matrix Q and discuss the kind of normalization that is more suited for the model to make it more realistic.

2.1.1. Importance of papers. The blocks that concur in assigning importance to a paper are: J_P, A_P and P_P.

- The block P_P is obtained from $\widehat{H}$ simply dividing each row by the number of its nonzero entries, that is $P_P = \mathrm{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$. In this way, a paper receives importances by the papers that cite it, moreover that paper distributes its importance uniformly among the papers in its reference list.
- A_P is obtained from $\widehat{K}$ by first normalizing the latter matrix by columns, and then by rows according to the procedure (2.4); in this model a paper receives a sort of weighted average of the importances of its authors instead of the sum of their importances.
- J_P is obtained from $\widehat{F}$ by dividing each entry by the maximum μ among the number of papers published on each journal and then by applying the row-normalization (2.4). This means that

  $$J_P = \begin{bmatrix} \frac{1}{\mu}F & (I - \frac{1}{\mu}F)\boldsymbol{e} \\ 0 & 1 \end{bmatrix}.$$

  In this way, the importance that a journal gives to a paper does not depend on the number of papers published in the journal. That is, each journal gives the same part of its importance to its papers independently of the number of papers published. The importance left is assigned to the dummy paper.

2.1.2. Importance of authors. The blocks that concur in assigning importance to an author are: J_A, A_A and P_A. The rank of an author should depend on the quality of the paper the author has written, the authoritativeness of his/her co-authors and the prestige of the journals where his/her papers are published. This is guaranteed by the following normalization of the blocks J_A, A_A and P_A.

- Matrix P_A is obtained by row-normalization of the matrix $\widehat{K}^T$, that is
$$P_A = \mathrm{diag}(\widehat{K}^T \boldsymbol{e})^{-1}\widehat{K}^T.$$
- Block A_A is obtained by row-normalization of the matrix $\widehat{K}\widehat{K}^T$, that is
$$A_A = \mathrm{diag}(\widehat{K}\widehat{K}^T \boldsymbol{e})^{-1}\widehat{K}\widehat{K}^T.$$
- The matrix J_A is obtained from $\widehat{F}\widehat{K}^T$ which contains the number of papers that each author has published on a given journal. An author will receive from each journal a fraction of the importance of the journal where he/she has published. In order to guarantee the rank of an author to be a right trade-off between quality and quantity, we divide uniformly each entry of the matrix by the maximum number of authors publishing in all the journals. Let ω be such a constant computed as $\omega = \max(\widehat{F}\widehat{K}^T \boldsymbol{e})$. We then apply procedure (2.4), assigning the importance left to the dummy author. Note that dividing each entry by ω guarantees that authors publishing in journals where only a restricted number of authors have published, are not favored.

2.1.3. Importance of journals. The rank of a journal depends on the quality of the papers published in that journal, on the quality of its authors and from the citations received by the other journals. In order to guarantee that quality wins over quantity we propose the following scaling of the rows of J_J, A_J and P_J to make them row-stochastic.

- The block P_J is obtained from $\widehat{F}^T$ first normalizing by column. In fact, the rank of a journal should not depend on the number of papers published by that journal, but rather on their quality. After the column normalization we apply procedure (2.4) to obtain a row stochastic matrix.
- A_J is obtained from matrix $\widehat{K}\widehat{F}^T$ which contains the number of papers written by an author on a given journal. We first divide each entry of $\widehat{K}\widehat{F}^T$ by $\beta = \max(\widehat{K}\widehat{F}^T \boldsymbol{e})$, that is the number of papers written by the author who wrote more papers and then we normalize by row using the procedure (2.4). Scaling the entries by β guarantees that the rank of a journal depends also on the number of different authors publishing in it.
- Block J_J accounts for how citations among journals contributes to the rank of a journal. We start with matrix $\widehat{F}\widehat{H}\widehat{F}^T$, which contains the citation's count among journals. To obtain J_J we first normalize by column dividing by the total number of papers published on a given journal, and then we apply normalization (2.4).

To better understand the way we normalize each block, let us consider the following example.

Example 1. Consider the case where we have 6 papers, 4 authors and 3 journals and the matrices involved (including the dummy subjects) are the following:

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

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

Performing the normalization described in Section 2.1 we obtain the blocks:

$$J_P = \left[\begin{array}{cccccc|c} \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right], A_P = \left[\begin{array}{cccccc|c} \frac{3}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right],$$

and P_P is obtained by row normalization of H. The matrices involved in the computation of the rank of authors are

$$J_A = \left[\begin{array}{cccc|c} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right], \quad A_A = \left[\begin{array}{cccc|c} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right],$$

and

$$P_A = \left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

Finally, for the rank of journals we get the following matrices

$$J_J = \left[\begin{array}{ccc|c} \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & \frac{1}{2} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} \\ \frac{2}{7} & \frac{1}{7} & 0 & \frac{4}{7} \\ \hline \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{array}\right], A_J = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right], P_J = \left[\begin{array}{ccc|c} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \hline 0 & 0 & 0 & 1 \end{array}\right].$$

Using uniform weights $\gamma_{i,j} = 1/3$, it turns out that the dominant left Perron vector is $\boldsymbol{\pi} = (\boldsymbol{\pi}_J, \boldsymbol{\pi}_A, \boldsymbol{\pi}_P)$, where

$$\begin{aligned}
\boldsymbol{\pi}_J &= (0.0658, 0.0566, 0.0541, 0.1567)\\
\boldsymbol{\pi}_A &= (0.0387, 0.0429, 0.0422, 0.0648, 0.1448)\\
\boldsymbol{\pi}_P &= (0.0305, 0.0301, 0.0295, 0.0362, 0.0351, 0.0351, 0.0391, 0.1329).
\end{aligned}$$

Then, disregarding the dummy players, journal 1 is more important than journal 2 and journal 3, author 4 is the one with higher rank, and paper 6 although is receiving just 2 citations, is the one with a higher rank because it is written by the best author on the best journal.

2.2. Comparisons with the existing models. Indeed the value of a scientific product can change over the time. In fact, the Impact Factor (IF) [7] of a journal varies along the years, and similarly, an author can be more important in certain years and less important in others. In our model we can determine the change of importance of a subject over the time, and we may also evaluate the change of the importance of this subject for a given year, depending on the year in which the evaluation is performed. So, for instance, the value of a given paper, or of a given journal issue can change year after year. And it can be interesting to track the importance of a subject along the time, like to evaluate the "average life" of a paper, i.e., how long a paper remains interesting.

Our model can easily perform this if we modify it in a simple way. First, we sort the papers by time in increasing order. Second, we need to disaggregate journals in the following way: a single journal is split into many "copies" depending on the year of publication. This splitting can be performed in pairs (like it is done with the computation of IF) or in single years. In this way, each subject is a journal plus the year, and they are ordered in increasing time. Similarly we should do with authors if we want to evaluate an author in his scientific life. The adjacency matrix that we obtain this way is much larger than the original one.

If we wish to take a picture of the importance of the different scientific subjects from year Y, we have just to consider the principal submatrix of the full adjacency matrix formed by the rows and columns related to the years preceding Y and apply the algorithm.

Observe that the column sum of the entries in the block J_J provides the number of citations received by each journal. In the model where journals are split according to the publication year in groups formed by two years, this sum provides the impact factor (IF) [7]. In fact, for computing the IF of the journals for a given year it is enough to consider the submatrix of the block J_J of the adjacency matrix formed by the citations up to that year and take the column sums related to the journals of the last pair of years, normalizing by the total number of papers published in the two previous years.

Observe also that the left eigenvector of the stochastic matrix obtained by scaling the rows of the block J_J by means of the row-sum of its entry is related to the *Eigenfactor* of [2]. The difference is that in [2] the adjacency matrix of journal citations is constructed without introducing a dummy journal, but is introduced a rank-one correction like in the Google approach in order to have irreducibility and acyclicity.

The novelty of our approach is that each subject contributes to the rank of the other subjects, while the previous approaches [2, 4, 11, 12] propose to rank journals only on the basis of the citations cumulatively received by all the papers published on a given journal. On the contrary we are ranking a journal on the basis of the importance of every single paper therein published. In [1] three different methods for ranking departments on the basis of the PhD placements of their faculty is considered. The matrices involved in those models are three as in our, but just one (that counting the number of faculties who received a degree from university i employed in university j) is used to get the rank.

It is important to point out that by choosing the parameter $\gamma_{i,j}$ such that $\gamma_{i,i} = 1$, $i = 1, 2, 3$ and limiting the analysis to a single class among *Journal*, *Authors*, *Papers*, we get the one-class model for the single classes. Choosing $\gamma_{1,1} = 1$, $\gamma_{1,2} = \gamma_{1,3} = 0$, and limiting the analysis to the classes *Authors* and *Papers* we obtain the two-class model of [3].

Concerning the one-class (*Papers*) and the two-class (*Authors, Papers*) models we can prove some perturbation results in the line followed in [3]. This is done in the next section.

3. Perturbation results. In a citation-based model, one expects that a paper which receives a new citation should increase its rank more than any other paper does. The following theorem, proved in [3], formalizes this property in terms of perturbation of the Perron vector and can be applied to the one-class and the two-class models where adding a new citation changes only one entry of the row-stochastic matrix.

THEOREM 3.1. *Let H be an irreducible adjacency matrix, let (r, s) be a pair of integers such that $h_{r,s} = h_{r,r} = 0$ and q be the number of nonzero entries in the r-th row. Define $\widehat{H} = (\widehat{h}_{i,j})$ such that $\widehat{h}_{r,s} = 1$, $i = 1, \dots, k$, $\widehat{h}_{i,j} = h_{i,j}$ otherwise. Let $P = \mathrm{diag}(H\boldsymbol{e})^{-1}H$, $\widehat{P} = \mathrm{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$ and denote by $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ their corresponding left Perron vectors. Then*

$$\sigma\frac{\widehat{\pi}_r}{\pi_r} \le \frac{\widehat{\pi}_j}{\pi_j} \le \frac{\widehat{\pi}_s}{\pi_s} \quad j = 1, \dots, n, \tag{3.1}$$

for $\sigma = q/(q + 1)$. Moreover,

$$\frac{\widehat{\pi}_j}{\pi_j} < \frac{\widehat{\pi}_s}{\pi_s}, \quad \textit{if} \;\; h_{r,j} \neq 0 \tag{3.2}$$

and

$$1 < \frac{\widehat{\pi}_s}{\pi_s}. \tag{3.3}$$

We may ask what happens if several papers receive a citation from the same paper. The following result formalizes this case.

THEOREM 3.2. *Let H be an irreducible adjacency matrix, let $(r, s_1), \dots, (r, s_k)$ be pairs of integers such that $h_{r,s_i} = h_{r,r} = 0$ for $i = 1, \dots, k$ and q be the number of nonzero entries in the r-th row. Define $\widehat{H} = (\widehat{h}_{i,j})$ such that $\widehat{h}_{r,s_i} = 1$, $\widehat{h}_{i,j} = h_{i,j}$ otherwise. Let $P = \mathrm{diag}(H\boldsymbol{e})^{-1}H$, $\widehat{P} = \mathrm{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$ and denote by $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ their corresponding left Perron vectors. Then*

$$\sigma\frac{\widehat{\pi}_r}{\pi_r} \le \frac{\widehat{\pi}_j}{\pi_j} \le \left(\prod_{i=1}^{k}\frac{\widehat{\pi}_{s_i}}{\pi_{s_i}}\right)^{1/k} \le \max_i \frac{\widehat{\pi}_{s_i}}{\pi_{s_i}} \quad j = 1, \dots, n, \tag{3.4}$$

for $\sigma = q/(q+k)$. *Moreover,*

$$\frac{\widehat{\pi}_j}{\pi_j} < \left(\prod_{i=1}^{k} \frac{\widehat{\pi}_{s_i}}{\pi_{s_i}} \right)^{1/k}, \quad \text{if } h_{r,j} \neq 0 \tag{3.5}$$

and

$$1 < \left(\prod_{i=1}^{k} \frac{\widehat{\pi}_{s_i}}{\pi_{s_i}} \right)^{1/k}. \tag{3.6}$$

Proof. Let us consider the case $k = 2$. The matrices $\widehat{H}$ and $\widehat{P}$ can be viewed as obtained in two elementary steps where we first compute $\widehat{H}_1$ from H by adding the link from r to s_1 and then we compute $\widehat{H}_2 = \widehat{H}$ from H_1 by adding the link from r to s_2. Denoting $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ the left Perron vectors of P, $\widehat{P}_1$ and $\widehat{P}_2 = \widehat{P}$, respectively, from Theorem 3.1 applied to H and $\widehat{H}_1$ we find that

$$y_{s_1}/x_{s_1} \geq y_i/x_i, \quad i \neq s_1$$

where the inequality is strict if $h_{r,i} \neq 0$. From Theorem 3.1 applied to $\widehat{H}_1$ and $\widehat{H}_2$ we find that

$$z_{s_2}/y_{s_2} \geq z_i/y_i, \quad i \neq s_2$$

where the inequality is strict if $h_{r,i} \neq 0$. From the above two expressions we deduce that

$$\frac{z_i}{x_i} = \frac{z_i}{y_i} \cdot \frac{y_i}{x_i} \leq \frac{z_{s_2}}{y_{s_2}} \cdot \frac{y_{s_1}}{x_{s_1}},$$

whence we get

$$\frac{z_i}{x_i} \leq \frac{z_{s_2}}{x_{s_2}} \cdot \frac{x_{s_2}}{y_{s_2}} \cdot \frac{y_{s_1}}{x_{s_1}}.$$

Since the role of s_1 and s_2 can be interchanged, we obtain the dual inequality

$$\frac{z_i}{x_i} \leq \frac{z_{s_1}}{x_{s_1}} \cdot \frac{x_{s_1}}{y_{s_1}} \cdot \frac{y_{s_2}}{x_{s_2}}.$$

Multiplying both sides of the latter two inequalities yields

$$\left(\frac{z_i}{x_i}\right)^2 \leq \frac{z_{s_1}}{x_{s_1}} \cdot \frac{z_{s_2}}{x_{s_2}}, \quad i \neq s_1, s_2,$$

which proves the main inequality in (3.4) for $k = 2$. Moreover, the inequality is strict if $h_{r,i} \neq 0$. The lower bound on $\widehat{\pi}_j/\pi_j$ given in (3.4) follows similarly. In the general case where $k > 2$ the proof can be carried out with the same technique by looking at $\widehat{H}$ as the matrix obtained after performing k transformation steps where at each step we add only one citation from paper r to paper s_i, for $i = 1, 2, \ldots, k$. Equation (3.6) follows directly from Theorem 3.1 applied at the last step of the transformations chain. □

The above theorem states that in the one-class and in the two-class models, if we add k new citations from paper r to k different papers, then for at least one paper the increase of rank is larger than the one obtained by the other papers. One would expect that *all* the papers that receive a citation should increase their rank more than the other papers. This is false in general, as we will show in the next section.

A similar result can be proved if we assume that only paper s receives a citation from papers $r_1, r_2, \dots, r_k$.

THEOREM 3.3. *Let H be an irreducible adjacency matrix, let $(r_1, s), \dots, (r_k, s)$ be pairs of integers such that $h_{r_i,s} = h_{r_i,r_i} = 0$ for $i = 1, \dots, k$ and q_i be the number of nonzero entries in the r_i-th row. Define $\widehat{H} = (\widehat{h}_{i,j})$ such that $\widehat{h}_{r_i,s} = 1$, $\widehat{h}_{i,j} = h_{i,j}$ otherwise. Let $P = \operatorname{diag}(H\boldsymbol{e})^{-1}H$, $\widehat{P} = \operatorname{diag}(\widehat{H}\boldsymbol{e})^{-1}\widehat{H}$ and denote by $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ their corresponding left Perron vectors. Then*

$$\min_{i=1,k} \sigma_i \frac{\widehat{\pi}_{r_i}}{\pi_{r_i}} \le \left(\prod_{i=1}^{k} \sigma_i \frac{\widehat{\pi}_{r_i}}{\pi_{r_i}} \right)^{1/k} \le \frac{\widehat{\pi}_j}{\pi_j} \le \frac{\widehat{\pi}_s}{\pi_s} \quad j = 1, \dots, n,$$

for $\sigma_i = q_i/(q_i + 1)$. Moreover,

$$\frac{\widehat{\pi}_j}{\pi_j} < \frac{\widehat{\pi}_s}{\pi_s}, \quad \text{if } \; h_{r,j} \ne 0,$$

and $1 < \frac{\widehat{\pi}_s}{\pi_s}$.

Proof. The proof is carried out with the same technique used for Theorem 3.2. ☐

3.1. A counterexample.

If a set of papers receive new citations, one would expect that the ranks of all these papers increase more than the ranks of the remaining papers do. Proving this property was addressed in [3] as an open problem. Here we provide an example which shows that this apparently intuitive property is false.

Indeed, there may exist situations where a paper which is not cited has an increase of rank larger than the one of some of the cited papers. This happens if this paper is cited by the papers which receive a citation and if the latter papers are sufficiently many. There are cases where some cited paper has even a *decrease* of rank since the number of cited papers is so large that not all of them can have an increase of rank due to the conservation of the overall rank. These counter-intuitive situations are shown by means of a very simple example.

Consider the one-class model formed by papers numbered from 1 to n plus the dummy paper, which is numbered as $n + 1$. The dummy paper cites and is cited by all the other papers. Moreover paper i cites paper $i + 1$, for $i = 1, \dots, n$. For $n = 5$, the adjacency matrix of this system is given by

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

while the row-stochastic matrix obtained by H is

$$P = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1/n & 1/n & 1/n & 1/n & 1/n & 0 \end{bmatrix}$$

Now assume that paper 1 cites papers $2, 3, \ldots, n-1$. In this way, the adjacency matrix is

$$\widehat{H} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

while the row-stochastic matrix obtained by H is

$$\widehat{P} = \begin{bmatrix} 0 & 1/(n-1) & 1/(n-1) & 1/(n-1) & 0 & 1/(n-1) \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1/n & 1/n & 1/n & 1/n & 1/n & 0 \end{bmatrix}$$

The graphs associated with the two systems are shown in Figure 3.1

For $n = 7$, the left Perron vectors $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ of P and $\widehat{P}$ are

$$\boldsymbol{\pi} = (0.05259, 0.07888, 0.09203, 0.09860, 0.10189, 0.10353, 0.10436, 0.36812)$$

$$\widehat{\boldsymbol{\pi}} = (0.05119, 0.05972, 0.08958, 0.10451, 0.11197, 0.11570, 0.10904, 0.35830)$$

while their ratio is given by

$$\widehat{\boldsymbol{\pi}}/\boldsymbol{\pi} = (0.97334, 0.75704, 0.97334, 1.05986, , 1.09893, 1.11754, 1.04487, 0.97334)$$

Observe that, among the papers that receive a new citation, i.e., papers 3, 4, 5 and 6, only paper 6 has the highest increase of rank (1.11754); papers 4 and 5 have a lower increase while paper 3 has a decrease of rank. Observe also that paper 7 which does not receive new citations has still an increase of rank. The latter change is due to the fact that paper 7 is cited by paper 6 which is cited by paper 4 which is cited by paper 3. A new citation received by papers 3, 4 , 5 and 6 provides an induced increase of rank of paper 7. The fact that paper 3 has a decrease of rank, despite it receives a new citation, seems at first glance quite odd, but it can be easily explained. The number of new citations, i.e., 4, is rather large with respect to the overall number of papers, moreover, paper 7 and 8 which do not receive citations receive importance indirectly. However there cannot be a general increase of importance because the rank of all the papers sum up to 1. This way some papers must decrease their rank.

4. Numerical experiments. We performed a number of experiments both on real and synthetic data. The goal of our experimentation is twofold. On one hand we use these experiments to validate the model by showing that the normalization proposed in Section 2.1 are effective. On the other hand we want to show that the ranking provided by our model is not equivalent to a simple counting of citations received, but it captures also concepts such as prestige or reputation. We report some results obtained applying our method to the matrix P in (2.3) with uniform weights, that is, choosing $\gamma_{i,j} = 1/3$.

For our experiments we use real data taken from the CiteSeer dataset, which can be freely downloaded from the CiteSeer web site [6]. CiteSeer is a scientific literature

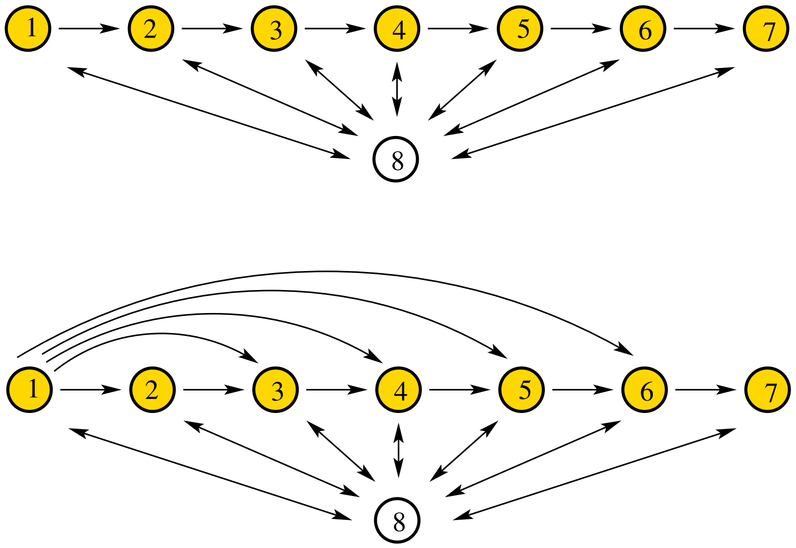

FIGURE 3.1. *Two systems, made of* n *papers plus the dummy paper, differ for* $n-3$ *new citations from paper* 1 *to papers* $3, 4, \ldots, n-1$*. Paper* $n-1$ *receives the largest increase of rank. For* $n \geq 10$*, paper* n *has an increase of rank larger than that of paper* $n-3$*. For* $n \geq 7$*, paper 3 has a decrease of rank*

digital library and search engine that focuses primarily on the literature in computer and information sciences [8]. 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.

The CiteSeer index used is a download of 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. However, in the XML format it is not contained the information about the journals where the papers were published. For a small part of papers it is possible however to recover the journal information crossing the XML file with the Bib-TeX file still available from the CiteSeer site. We obtained in this way a set of about 37,000 papers, written by approximately 41,000 authors and published in 2829 journals or conference proceedings.

The dataset obtained is not well suited for our algorithm for many reasons. Firstly, among the 2829 journals, more than a half (exactly 1636), appear with just a paper in the database. This means that we have to rank a journal on the basis of the quality of only one paper therein published. For the subject *Authors* we have a similar problem, since many well-known scholars appear as authors of only a small number of papers, this is mainly due to the fact that the CiteSeer index is built starting from papers available on-line (mostly technical reports) or spontaneously submitted. This means that the index doesn't contain all the papers that appear on a certain journal issue, but just those appearing in the reference list of some other indexed paper. Another problem with this database is that proceedings of the same conference appear several times as different journals, one for each year the conference has taken place. It should be more appropriate to group them under the same journal, since the reputation of a conference builds up over the years.

These observations tell us that we have to evaluate the results in Tables 4.1, 4.2

`paper`	`pos.`	`cit.`
Kirkpatrick, Gelatt, Vecchi - Simulated Annealing - SCIENCE	1	174
Diffie, Hellman - Cryptography- IEEE Trans. Inform. Theory,	3	122
Elman- Finding Structure in Time - Cognitive Science	10	84
Van Gelder, Ross, Schlipf - General Logic Programs - J. ACM	2	127
Sunderam - PVM - Concurrency, Practice and Experience	8	92
Harel - Statecharts - Science of Computer Programming	4	107

TABLE 4.1

Experimental results for the subject Papers. In the first column papers in order of decreasing rank are listed with the name of the authors and a short identification of the title of the paper, and the abbreviation of the Journal. The second column contains the position in the list ordered by decreasing number of citations, the third column reports the number of citation received by the paper.

and 4.3 in the light of these specific features of the database. For instance, concerning the subject *Journals*, we adopted a column-normalization for the block J_J in order to make the rank of a journal independent of the number of papers published by enhancing quality over quantity. However, for the characteristics of our data, it happens (as for the journal *SCIENCE* or *J R STATIST SOC B* in Table 4.3) that journals with just a paper, pops up in the ranking because the only paper published there [1] is very good. Also not so prestigious journals with just a paper score a rank higher than expected because the importance of the only paper therein published is higher than the average importance of papers published on other journals. This particular example show that this database is too incomplete, since we should consider in the game also those papers published on a given journal that did not get any citations. This will decrease the importance of the journal because of the normalization of block J_J.

Note that the results obtained with this model for the category *Papers* and *Authors* differs from that obtained in [3] using the two-class model. Indeed, we are using different data where important and highly cited papers which score a high value of rank are no more present in our database. In fact, we could keep in the database only the items with an associated journal, and for most of the papers this information was missing.

The behavior of our method on a more realistic dataset is described in Figures 4.1. These plots represent the validation of our method on a set of 260,000 papers from a single discipline. These papers were published in 1131 journals by more than 100,000 different authors. Here we present just the statistical behavior of the method.

From 4.1 we can see that the rank of a journal doesn't depend on the number of papers that were published in that journal. However, it turns out that the journals publishing many papers receive a good ranking score. The second plot in Figure 4.1 shows the dependence on the total number of citations received.

As underlined before, the CiteSeer database but also the papers used to produce Figure 4.1 are not complete, in the sense that they do not index complete issues of a journal, but the index is constructed bottom-up, starting from a bunch of papers and adding to the index the papers in their bibliography. A more suitable index should be constructed considering instead a more complete database, where all the papers appearing on a given journals are considered. Unfortunately, we do not have

---

[1] In the first example S. Kirkpatrick, C. D. Gelatt, Jr., M.P. Vecchi, *Optimization by Simulated Annealing*, Science, Number 4598, 13 May 1983.

Author	num. cit	num. pap.	av. num. cit.
Oded Goldreich	196	75	2.6
Moni Naor	209	53	3.9
Douglas C. Schmidt	131	49	2.7
Sally Floyd	372	55	13.3
Henry G. Baker	57	24	2.4
Moshe Vardi	88	47	1.8
Jun Zhang	54	29	1.9
G. W. Stewart	49	31	1.6
Marek Karpinski	83	54	1.5
Jack J. Dongarra	166	69	2.4

TABLE 4.2

Experimental results for the subject Author. In the first column the top authors are ranked in decreasing order of rank. In the remaining columns there are reported: the number of citation received, the number of papers by the author and indexed in the dataset, and the average number of citation per paper.

Journal	pos.	cit.	#cit/#pap.
LECTURE NOTES IN COMPUTER SCIENCE	1	4839	0.72
SCIENCE	69	174	174
THEORETICAL COMPUTER SCIENCE	2	1915	1.68
15TH INT. CONF. DISTR. COMP. SYS.	220	35	35
J R STATIST SOC B	220	35	35
ELECTR. COLLOQ. COMP. COMP. (ECCC)	42	319	1.09
INFORMATION PROCESSING LETTERSG.	20	1194	0.94

TABLE 4.3

Experimental results for the subject Journal. In the first column the top Journals or conference proceedings are ranked in decreasing order of rank. The second column contains the position in the list ordered by decreasing number of citations, the third column reports the number of citations received by the journal. In the fourth column the average number of citations obtained by papers published in the corresponding journal is reported. Note that the fourth and fifth journal are ex-aequo at the 220-th position in the list ordered by number of citations, in fact they collect 35 citations. Actually, both these journals are represented by only one paper in our database, and this paper is of a good quality so that the rank of the two journals pops up in the list.

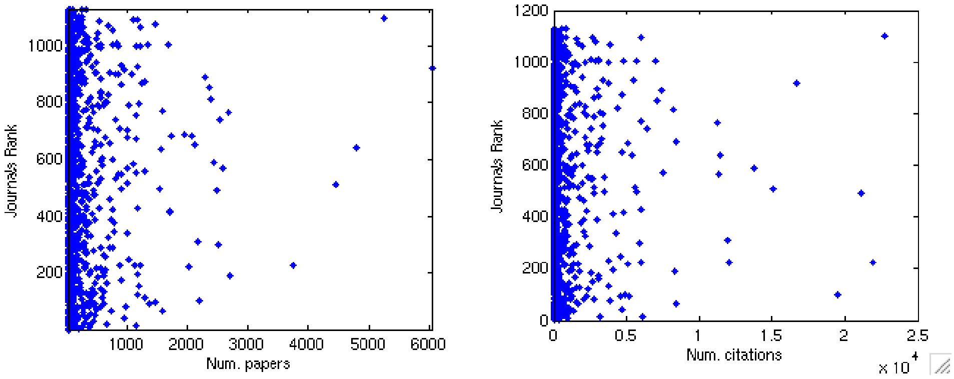

FIGURE 4.1. *Experimental results for the subject Journal. In the first plot it is shown the dependence of the rank of journals on the number of papers it published. The second plot shows the dependence on the number of citation received.*

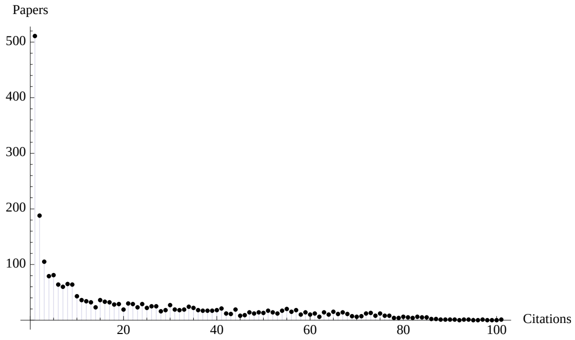

FIGURE 4.2. *Histogram of number of papers receiving a given number of citations. More than 500 papers receive just a citation, very few of them receive more than 80 citations.*

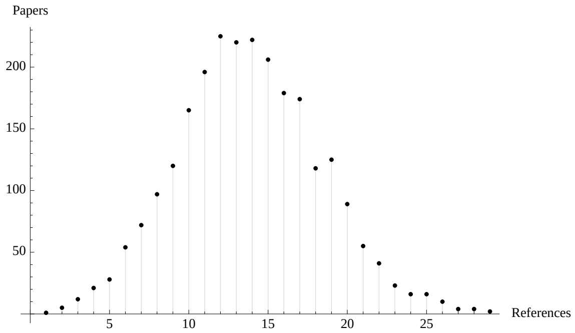

FIGURE 4.3. *Histogram of number of bibliographic items contained in the papers. It has a Gaussian-like shape where most of the papers have a list of around 15 items in the reference list and very few papers have less than 5 or more than 25 items.*

access to a database of this kind so, we simulated this behavior by building up a synthetic database, with statistical properties similar to those of real data. The index we constructed is small but the result produced show a good behavior of our method. We simulated an index with 2,500 papers, 150 authors and 20 journals. Figure 4.2 shows the histogram of the citations received by papers. The idea was to build up a database where very few papers receive many citations, and most of them just one or two citations. A power-law distribution [10] can model this kind of behavior. The length of the reference list of papers, has the behavior described by histogram 4.3.

Concerning authorship, we generated data where the majority of papers have a single author, and where papers written by 5 or more authors are less numerous. Indeed, this feature is closer to scientific areas like pure mathematics rather than other areas like experimental physics where it is more likely to encounter papers with a huge list of coauthors. However, the normalizations performed on blocks A_J, A_A and A_P guarantee that the rank does not depend on the number of authors of a paper.

The papers are more or less uniformly distributed over the 20 journals, so we have that each journal publish on the average 125 papers with a minimum of 99 and a maximum of 157.

The experiments performed on these data go into three different directions. We want to analyze the behavior of the rank with respect to the quantity, the sum of quality and the mean quality. In particular, the goal is to show that journal rank is

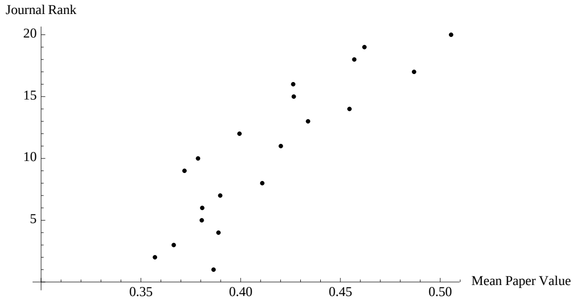

Figure 4.4. *Dependence of the rank of a journal on the mean rank value of the papers it publishes*

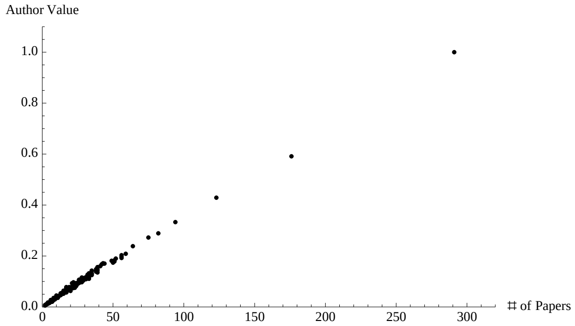

Figure 4.5. *Author's value versus number of papers written.*

independent of the number of papers it publishes, but it depends on the quality of the papers therein published and on the quality of the authors publishing there. In Figure 4.4 it is reported the dependence of the rank of journals on the mean "quality" of the papers published there. We see an almost linear dependence, that means that journals ranking higher publish papers with a higher rank. The rank of Journals is almost uncorrelated with the number of papers or the number of authors publishing there, while the dependence of the rank of a journal on the average of the values of the authors with a paper on that journal is very similar to that reported in Figure 4.4.

Concerning the rank of authors, we expect this rank to be somehow dependent on the number of authored papers and on their value. Figures 4.5 and 4.6 report the result obtained. We see that the value of an author grows almost linearly with the number of papers written while the rank of authors increases with the sum of the values of its papers but the behavior is not linear.

In Figure 4.7 we plot the value of papers versus the number of citations received. We have a strong dependence on the number of incoming references but the rank of a paper is not obtained by simply counting the number of citation received.

The dependence of the rank of a paper on the value of its "mean" author is plotted in Figure 4.8. Even if there is not a strong dependence, we see that there are no low rank papers written by important authors. On the contrary also important papers are written by authors with a low mean value.

5. Conclusions. We presented a model for evaluating the quality of scientific publications, the productiveness of authors and the prestige of journals. The main

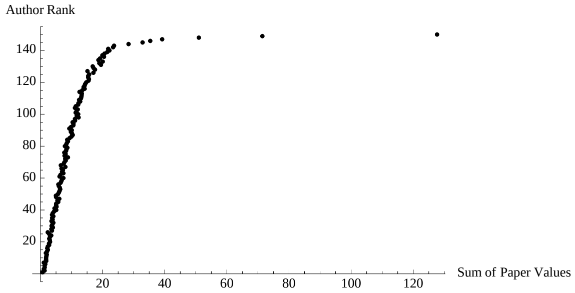

FIGURE 4.6. *Author's rank versus sum of the values of the papers written.*

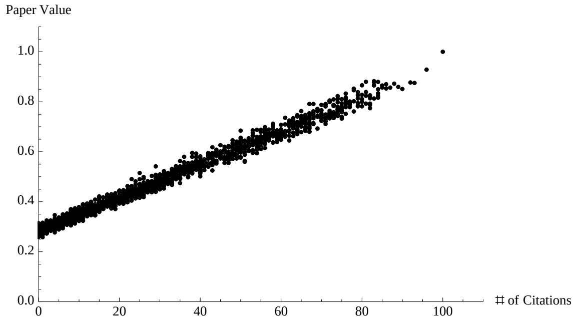

FIGURE 4.7. *Rank of papers versus the number of citations received.*

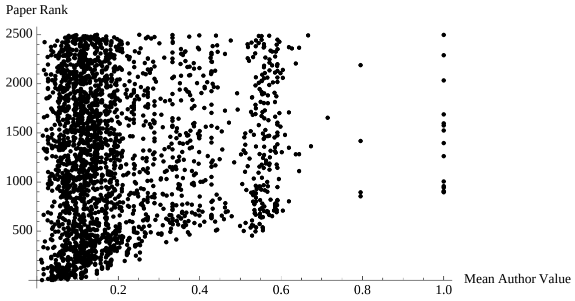

FIGURE 4.8. *The rank of papers is plotted versus the value of the mean author of the paper. For each paper, the value of the mean author is given by the arithmetic mean of the coauthors.*

idea is that the rank of a subject (such as a journal, an author or a paper) is obtained as the weighted sum of the ranks of the other subjects. In this way the rank of a paper does not depend only on the citations received by the paper, but also on the prestige of the journal where the paper is published and on the quality of the author writing it. We discussed how the rank of every subject can be obtained as the Perron vector of a stochastic, irreducible matrix that can be constructed from the citation matrix H, the matrix K accounting for authorship and the matrix F which keeps track of the journal where a paper is published. We obtained a 3×3 block matrix, and in Section 2.1 we discussed the kind of normalization more suited for each block to make the model more realistic. We also gave some perturbation results for a simpler model proposed in [3] accounting only for the subjects Authors and Papers, together with a counterexample concerning an issue pointed out in [3].

The proposed model was tested both on real and synthetic data. The experiments showed that the normalization of the block of the matrix make the model robust to spam, in the sense that it measures quality without favoring journals which publish too many papers or authors who write many low quality papers. This is reflected on the fact that the rank of a journal is independent of the number of papers that it publishes but it is closer to the average quality of its papers, and that the rank of a paper cannot be obtained simply counting the number of references received. The model can be however further improved introducing the factor time which allows to weight differently new and old citations. It is possible to introduce another improvement to avoid self citations. This can be done by removing from the list of incoming reference all citations coming from papers written by one of the authors of the paper.

The normalization of the various blocks proposed in Section 2.1 are uniform except for the dummy subject. It is however possible to normalize accordingly to a different distribution. For example, this can be done in the case of the authors of a paper are listed in a non alphabetic order or when it is clear which of the authors should get more credit for a paper.

We proposed this ranking method in the framework of research evaluation, it is however possible to use the same approach for other ranking problems where the influence of a subject depends on the importance of other subjects, such as the problem of ranking news stories and news agencies or the ranking of web sites on the basis of web pages.

REFERENCES

[1] R. Amir and M. Knauff. Ranking economics departments worldwide on the basis of PhD placements. *The Review of Economics and Statistics*, 90(1):185–190, 2008.

[2] C. T. Bergstrom. Eigenfactor: Measuring the value of prestige of scholarly journals. *C&RL News*, 68(5), 2007.

[3] D. A. Bini, G. M. Del Corso, and F. Romani. Evaluating scientific products by means of citation-based models: a first analysis and validation. *Electron. Trans. Numer. Anal.*, 2008. to appear.

[4] J. Bollan, M. A. Rodriguez, and H. Van de Sompel. Journal status. *Scientometrics*, 69(3):669–687, 2006.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN Systems*, 30:107–117, 1998.

[6] Citeseer.IST. Computer and Information Science Papers. CiteSeer Publications ResearchIndex. http://citeseer.ist.psu.edu/.

[7] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471, 1972.

[8] C. Lee Giles, Kurt Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In Ian Witten, Rob Akscyn, and Frank M. Shipman III, editors, *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA, June 23–26 1998. ACM Press.

[9] L. Hogben, editor. *Handbook of Linear Algebra*. Chapman Hall/CRC, 2007.

[10] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1:226–251, 2003.

[11] I. Palacios-Huerta and O. Volij. The measurement of intellectual influence. *Econometrica*, 72(3):963–977, 2004.

[12] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications-theory, with applications to literature of physics. *Inform. Process. Manag.*, 12:297–312, 1976.