Versatile weighting strategies for a citation-based research evaluation model

Gianna M. Del Corso F. Romani

Abstract

After a quick review of the most used numerical indicators for evaluating research, we present an integrated model for ranking scientific publications together with authors and journals. Our model relies on certain adjacentcy matrices obtained from the relationship between papers, authors, and journals. These matrices are first normalized to obtain stochastic matrices and then are combined together using appropriate weights to form a suitable irreducible stochastic matrix whose dominant eigenvector provides the desired ranking. Our main contribution is a in-depth analysis of various strategies for choosing the weights, showing their probabilistic interpretation and showing how they affect the outcome of the ranking process. We also prove that, by solving an inverse eigenvector problem, we can determine a weighting strategy in which the relative importance of papers, authors, and journals is chosen by the final user of the ranking algorithm. The impact of the different weighting strategies is analyzed also by means of extensive experiments on large synthetic datasets.

1 Introduction

Evaluation of scientific research has always been a very important problem. Recently, the number of scientific journals and papers has increased at an almost exponential rate [23] making the task of using and evaluating scientific literature much harder than in the past. For example, researchers now rely on search engines such as `Google Scholar` to choose what to read or what to cite. This

Received by the editors July 2009.
Communicated by A. Bultheel.
2000 *Mathematics Subject Classification* : 65C20, 65F15.
Key words and phrases : PageRank, Perron vector, perturbation results, impact factor.

problem doesn't affect only researchers but also funding agencies, university administrators, and reviewers called to evaluate productivity of researchers and institutions. Most of the times it is impossible to give an in-depth evaluation of the research performed by a scholar or institution, and it is becoming common to use indirect indicators of quality.

Among such indicators, the most popular are currently those based on citation analysis. In addition to the obvious over-simplification introduced by using just a few numbers to quantify the scientific merit, citation analysis has other weaknesses. For example, it is based on the assumption that a citation is a sort of trusting vote; this is not always the case since an author can cite a paper to criticize its content. In addition, since authors know the mechanism of citation analysis papers sometimes contain unnecessary self-citations. Another criticism to the use of citations as the "corner stone" to asses quality of research is that many items contained in the reference list of a paper are papers written by people in the entourage of the authors.

On the other hand, as soon as a paper is discovered to contain errors usually it does not receive further citations. In addition, many studies [2, 15, 14] show that self-citations do not inflate citation rates as one could expect since they rapidly lose their weight as time elapses, aging much faster than citation coming from other sources. Finally, one can argue that a thorough peer review process of the published papers should guarantee the appropriateness of the reference list. Summing up, citation analysis is usually recognized as a credible and convincing approach, especially for a quick, simple and objective evaluation of a large amount of data when peer review in not practicable.

In the literature one can find many different metrics for evaluating papers, journals, or researchers. The reason is that there are many possible different purposes for ranking. For instance, the ranking of journals is interesting for librarians to decide subscriptions and for authors to decide where to publish. The ranking of papers is becoming useful for untangling the maze of papers published everyday, and decide what to read or what to cite. Likewise, it is becoming common to evaluate of scholars on the basis of their scientific productivity for distributing funds, or even for hiring people.

Among the different methods proposed in the literature for ranking scientific research we can distinguish between methods based on citation statistics — such as Impact Factor (IF) (see [12] and references therein for an historical review), simple Citation Count, the MCQ by the American Mathematical Society [3] — and methods based on approaches similar to Google PageRank, such as Eigenfactor [4], SCImago [22] and others [21, 20].

Metrics based on citation statistics are easy to compute but not all the scientific community agrees on the effectiveness of these metrics to capture concepts such as reputation or influence. In particular, two major objections against the use of such measures are that the same journal may publish papers with very different citation rates, and that the "culture" of citations depends on the scientific areas [1]. For examples, in fields such as mathematics or economics, the process of citation gathering is much slower that in fields such as cell biology, and it can take decades before the process stabilizes [23]. This reflects on the fact that the average length of the reference list significantly varies among different disciplines.

For the ranking of individual papers the use of citation-based statistics is even more questionable. Assessing the quality of a single paper on the basis of the prestige of the journal where it is published is certainly a crude approximation since it amounts to ascribing the properties of a journal to all the individual articles it contains. Similarly, the approach of evaluating the quality of a paper counting the citations received is not fair with respect to relatively recent papers. Recently, citation-based statistics have been used also for the ranking of individual researchers. Indexes such as the h-index [16], g-index [9], m-index [16] are based on the citations received by most cited papers of a given author. These indexes are relatively easy to compute, but again they don't measure the impact of an author on the scientific community. In fact, scientists with a short career are at an inherent disadvantage, regardless of the importance of their discoveries.

Metrics based on PageRank-like techniques seem more appealing since the effectiveness of PageRank for ranking web pages is proved by everyday use, and citations in paper have a similar role than links in web pages. The main idea is that not all the citations are equal and that, rather than the number, one should consider the "quality" of citations. These metrics have nice mathematical properties: for example in [20] it is shown that a centrality measure similar to PageRank is the only ranking satisfying a number of axiomatic requirements.

In [5, 6] we propose an integrated three-class model for the ranking of papers, authors, and journals loosely inspired by the PageRank algorithm. In our model papers, authors, and journals represent three distinct classes that mutually contribute to the attribution of a ranking score to each element of each class. The idea is that to evaluate an author we consider not only the quality of the journals where his/her papers have been published, but also the quality of every single paper he/she authored. In addition, we take into account also the "quality" of the co-authors. 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 in the scientific community. Similarly, to evaluate the quality of a paper we consider the quality of the journal where the paper is published, the citations received, and at the reputation 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 [11], Eigenfactor [4], and others [7, 20] — but also the quality of every single paper published there, and the authoritativeness of the authors who published on that journal.

In this paper we describe a probabilistic interpretation of our model and consider the problem of choosing the weighting parameters of the model. The weighting parameters allow one to tune the relative importance of the three classes. Hence, different weighting strategies can be used according to the intended use of the ranking algorithm. The role of the weighting parameters is also investigated by means of experiments on syntetic data. Finally, we compare the ranking provided by our method with those returned by Impact Factor.

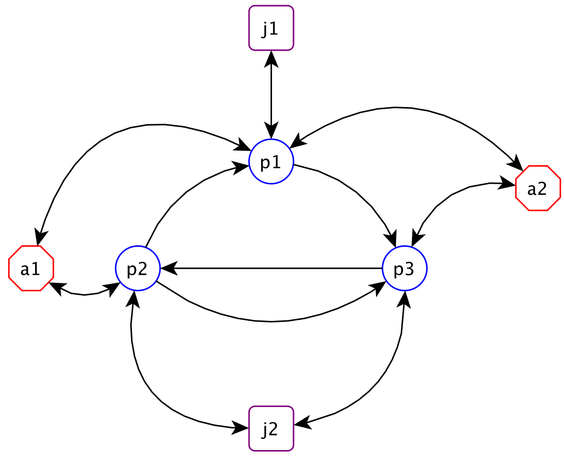


Figure 2.1: A graph where we have different nodes for each category. We have three papers, two authors and two journals.

2 The basic model

Assume we are given n_P papers together with their bibliographic data. More precisely, 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 construct a graph with three different kind of nodes (see Figure 2.1). We associate with this graph three matrices, one for every kind of nodes: the matrix F which records which journal has published 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 denote the number of distinct authors who authored the n_P papers. We define $F = (f_{i,j})$ as the $n_J \times n_P$ binary matrix such that

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

$K = (k_{i,j})$ as the $n_A \times n_P$ 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 $H = (h_{i,j})$ as the $n_P \times n_P$ matrix such that

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

In the example of Figure 2.1 we have

$$F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad K = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad H = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

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

$$A = \begin{bmatrix} FHF^T & FK^T & F \\ KF^T & KK^T & K \\ F^T & K^T & H \end{bmatrix} \tag{1}$$

of size $N = n_J + n_A + n_P$. For the example in Figure 2.1 it is

$$A = \left[\begin{array}{cc|cc|ccc} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 & 0 & 1 & 1 \\ \hline 1 & 1 & 2 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 2 & 1 & 0 & 1 \\ \hline 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}\right].$$

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

We can scale the rows of A to obtain a row-stochastic matrix P, that is $P\boldsymbol{e} = \boldsymbol{e}$, where $\boldsymbol{e} = (1, \dots, 1)^T$. Then, we compute the ranking score of the subjects as the left eigenvector corresponding to the eigenvalue 1,

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

More precisely, numbering the subjects from 1 to N, the rank value (or 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

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

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 and uniqueness of a solution we need A to be irreducible. Under this condition, it is always possible to find a scaling technique such that the matrix P can be constructed. The Perron Frobenius theorem [17] guarantees the existence of a unique vector $\boldsymbol{\pi}$, such that $\pi_i > 0$ and $\sum_i \pi_i = 1$.

We refer to $\boldsymbol{\pi}$ as the *Perron vector* of P. Moreover, in order to have nice convergence properties of iterative algorithms for the computation of $\boldsymbol{\pi}$ we need A to be aperiodic.

Note that working with the stochastic matrix P rather than computing the dominant eigenvector of A has advantages also from a numerical point of view. In fact, the approximation of the dominant eigenvector is done using an iterative procedure, and working with a stochastic matrix guarantees that we don't need to perform a normalization at each step to limit the growth of the entries of the intermediate vectors.

There are many ways to enforce the irreducibility of A; we use a technique inspired by the Google PageRank model [8] but which is different from a mathematical point of view. Similarly to what we have done for the one-class model [5], we obtain an irreducible and aperiodic matrix introducing a dummy paper, a dummy author, and a dummy journal. We assume that the dummy paper is cited by every paper and cites back all the papers except itself. We also assume that 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,

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

and to replace H, K and F in (1) with $\widehat{H}, \widehat{K}$ and $\widehat{F}$, respectively. It is easy to prove the following theorem [6].

Theorem 1. *The matrix $\widehat{A}$ obtained by replacing the blocks H, K and F in* (1) *with the blocks $\widehat{H}$, $\widehat{K}$, and $\widehat{F}$, respectively, is irreducible and aperiodic.*

2.1 Row and column scaling

In the previous section we pointed out the importance of scaling the rows of A to obtain a row-stochastic matrix. The simplest strategy is dividing each row of A by the sum of the entries in the row. A more flexible strategy, introduced in [5, 6], consists in performing a separate normalization of each block of A. That is, each block of A is normalized to yield nine row-stochastic matrices; then these matrices are compounded with weights $\Gamma = (\gamma_{i,j})_{i,j=1,2,3}$, where Γ is row stochastic, into a new stochastic matrix. The entries of Γ can be used to weight the amount of importance that each class (*Journal*, *Authors*, and *Papers*) transfers to the other classes. An in-depth discussion about the different possible normalization strategies of the single blocks is presented in [6] where a proposal for the normalization of each block is discussed. In this paper we use the same normalizing techniques presented and motivated in [6].

Denote by

$$Q = \begin{bmatrix} J_J & J_A & J_P \\ A_J & A_A & A_P \\ P_J & P_A & P_P \end{bmatrix}, \tag{2}$$

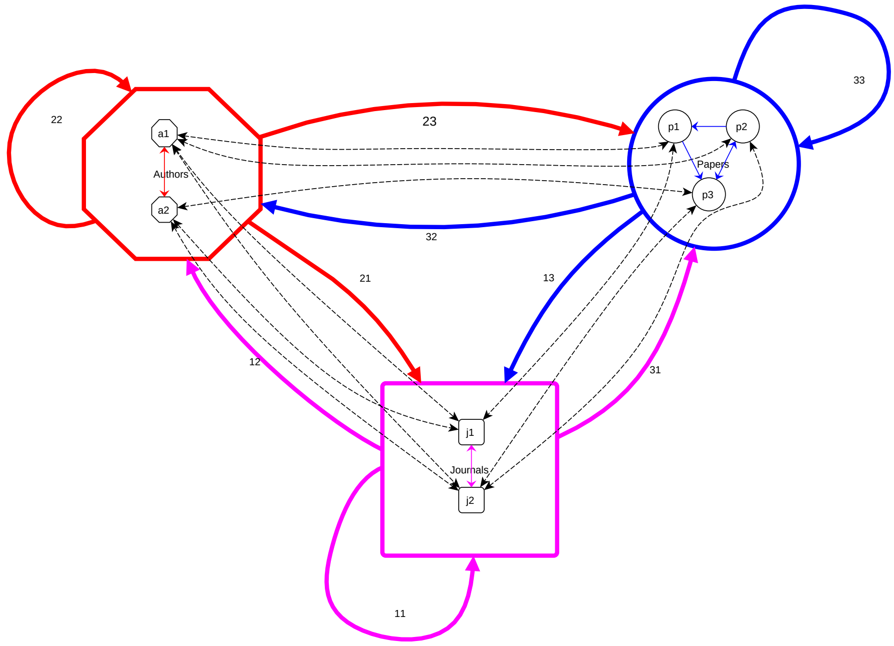


Figure 2.2: A two-level graph, representing the matrix P. Tick solid lines represent direct arcs between the three different classes and are labeled with the weights expressed by γ. Following the dashed arcs we can recover informations about the authors of a given paper and the journal where the paper has been published. Thin solid directed arcs between subjects in the same class represent the link described by the diagonal blocks of Q.

the matrix obtained from the corresponding block in the matrix $\widehat{A}$ of Theorem 1. Each block is row-stochastic, and for example J_J is the stochastic matrix obtained by the row-normalization of $\widehat{F}\widehat{H}\widehat{F}^T$.

The notation used in (2) shows 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

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

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 that the importance of papers comes more from the citations they receive rather than from the importance of their authors or of the journals where they are published. Figure 2.2 shows a representation of the graph associated to the matrix P.

3 Probabilistic interpretation and choice of weights

Similarly to what was done for the PageRank model, we can give a probabilistic interpretation of our model in terms of a "random reader" or "random evaluator". According to this interpretation, the dummy journal represents the library, the dummy author is the librarian, and the dummy paper is the catalog of the library. Note that we should expect the rank of the dummy subjects to be higher than that of the subjects belonging to the same class, since the random reader consults more frequently the library or the catalog than a single paper or journal.

The random reader after entering the library and asking for the catalog, picks a paper $\mathcal{P}$ and then she performs one of the following three actions. She keeps reading papers choosing among the ones in the reference list of $\mathcal{P}$, or she jumps to one of the coauthors of $\mathcal{P}$, or she looks at the journal where $\mathcal{P}$ is published. Each of these actions happens with probability $\gamma_{3,i}, i = 1,2,3$. While examining an author $\mathcal{A}$ the random reader with probability $\gamma_{2,2}$ chooses one of the coauthors, with probability $\gamma_{2,1}$ she browses the journals where $\mathcal{A}$ has published or with probability $\gamma_{2,3}$, she starts reading one of the papers written by $\mathcal{A}$. While examining a journal $\mathcal{J}$, the reader can move to another journal cited by papers in $\mathcal{J}$, can pick a paper published in $\mathcal{J}$ or can start examining an author who has published papers in journal $\mathcal{J}$. The random reader jumps from a class to another with a probability described by the 3×3 Markov chain Γ. The probability of picking in a particular class is, on the other hand, ruled by the underlining Markov chain described by the nine stochastic matrices $J_J, J_A, J_P, A_j, A_A, A_P, P_J, P_A$ and P_P.

The choice of modeling the problem with stochastic matrices combining them with the weights γ_{ij} allows one to tune how much of the importance of a class we want to transfer to another class. Define

$$\mu_J = \sum_{i=1}^{n_J} \pi_i, \quad \mu_A = \sum_{i=n_J+1}^{n_J+n_A} \pi_i, \quad \mu_P = \sum_{i=n_J+n_A+1}^{n_J+n_A+n_P} \pi_i, \tag{4}$$

It turns out that the vector (μ_J, μ_A, μ_P) is the left Perron eigenvector of Γ corresponding to the eigenvalue 1. In fact, the following Theorem holds [19].

Theorem 2 (Coupling Theorem)**.** *Let P be an $n \times n$ irreducible stochastic matrix partitioned as*

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & P_{1,3} \\ P_{2,1} & P_{2,2} & P_{2,3} \\ P_{3,1} & P_{3,2} & P_{3,3} \end{bmatrix},$$

with square diagonal blocks. Then the stationary distribution vector for P is given by

$$\pi^T = (\xi_1 \boldsymbol{s}_1^T, \xi_2 \boldsymbol{s}_2^T, \xi_3 \boldsymbol{s}_3^T)$$

where $\boldsymbol{s}_i$ is the unique stationary distribution vector for the stochastic Schur complement S_{ii}. The vector

$$\xi^T = (\xi_1, \xi_2, \xi_3)$$

is the unique stationary distribution vector for the 3×3 *irreducible stochastic matrix C whose entries are defined by*

$$c_{ij} = \boldsymbol{s}_i^T P_{ij} \boldsymbol{e}. \tag{5}$$

The matrix C is called the coupling matrix and the scalars ξ_i are called the coupling factors. ■

The Coupling Theorem applied to matrix P of equation (3), can be used to show that Γ is the *coupling matrix*. In fact, since the nine matrices $J_J, J_A, J_P, \dots$ are stochastic, from (5) we get

$$\begin{aligned} c_{ij} &= \boldsymbol{s}_i^T P_{ij} \boldsymbol{e} \\ &= \gamma_{ij} \boldsymbol{s}_i^T \boldsymbol{e} = \gamma_{ij} \end{aligned}$$

where the last equality holds because $\boldsymbol{s}_i$ are distribution vectors and $\boldsymbol{s}_i^T \boldsymbol{e} = 1$. Moreover, the scalars μ_J, μ_A, μ_P, introduced in (4), are proportional to the coupling factors since $\xi_1 = \|\xi_1 \boldsymbol{s}_1\|_1 = \sum_{i=1}^{n_J} \pi_i = \mu_J$ and similarly for μ_A and μ_P. This means that the vector $(\mu_J, \mu_A, \mu_P)^T$ is the Perron eigenvector of the 3×3 matrix Γ, or equivalently, it corresponds the unique stationary distribution vector of the coupling matrix Γ.

In [6] it was suggested to use uniform weights, which corresponds to having $\Gamma = 1/3 \boldsymbol{e}\boldsymbol{e}^T$. The dominant (left) eigenvector of Γ, or equivalently the stationary distribution of the coupling matrix, is the vector $1/3\, \boldsymbol{e}$. With this choice each class has the same role in determining the importance of each subject since $\mu_J = \mu_A = \mu_P = 1/3$. This also implies that the mean value of a journal is $1/(3\, n_J)$, and the mean value of an author and a paper are respectively $1/(3\, n_A)$ and $1/(3\, n_P)$. However, in practical situations the number of journals, authors, and papers differ in order of magnitude. Typical values [3] are $n_J \approx 10^3$, $n_A \approx 10^5$, $n_P = \approx 10^6$, making the mean value of a journal two or three orders of magnitude larger that the mean values of authors and papers. This means that journals play a bigger role in the determination of the ranking of the other subjects, while papers and authors have a smaller role. Of course, citations are still important because they influence the rank of journals in block J_J.

To correct this situation, we can think to a different weighting criteria. We ask the following question: Which is the best weighting strategy if we want the average paper to hold as much as the average journal or author? Since the average value of each class is μ_i / n_i, with $i \in \{J, A, P\}$, the solution to this problem relies on solving an inverse problem, where the Perron eigenvector is $(n_J/N, n_A/N, n_P/N)^T$ with $N = n_J + n_A + n_P$, and we are seeking a stochastic 3 matrix Γ. By direct substitution we see that a solution is given by

$$\Gamma = \frac{1}{N} \begin{bmatrix} n_J & n_A & n_P \\ n_J & n_A & n_P \\ n_J & n_A & n_P \end{bmatrix}. \tag{6}$$

In Section 4 we present experimental results to evaluate the differences between these weighting strategies.

In view of the above considerations, we see that working with a symmetric stochastic Γ will always produce an unbalanced average importance for each class. In fact, if

$$\Gamma = \begin{bmatrix} 1-a-b & a & b \\ a & 1-a-c & c \\ b & c & 1-b-c \end{bmatrix}, \quad a, b, c \in [0, 1], \tag{7}$$

then we obtain a Perron vector equal to $1/3(1,1,1)^T$ and the average value of each class will be the same as in the uniform model. Of course, the actual value of each subject will change even if the average value remains the same for each choice of the parameters in (7). Note moreover, that for small values of the parameters, we obtain a diagonally dominant matrix. Hence, the rank of each class will depend mostly on the values within the class but the problem will become close to reducibility and therefore numerically unstable.

Summing up, in our model we can influence the average outcome values for each class with an appropriate choice of the weights. This requires solving an inverse eigenvector problem and looking for a 3×3 stochastic matrix Γ, with the prescribed eigenvector corresponding to the eigenvalue 1. As an example, suppose we want to force our method to rely more on citations rather than on authorship or importance of journals. More in general, we ask which is a possible choice of Γ such that the average importance of a paper is k times that of a journal, and that of an author is h times that of a journal. To answer these questions we need to find a stochastic matrix Γ with Perron vector equal to $(n_J/C, hn_A/C, kn_P/C)$ where $C = n_j + hn_A + kn_P$. One possible Γ is

$$\Gamma = \frac{1}{C}\begin{bmatrix} n_J & h\,n_A & k\,n_P \\ n_J & h\,n_A & k\,n_P \\ n_J & h\,n_A & k\,n_P \end{bmatrix}. \tag{8}$$

Note that in the probabilistic interpretation of the model presented at the beginning of this section, the value γ_{ij} represents the probability of jumping from class i to class j, for $i, j \in \{J, A, P\}$. Hence, choosing Γ as in (8), with $k > h > 1$, means that the random evaluator will spend more time reading papers than examining authors or journals.

We point out that, although it is possible to know in advance the average value of a particular class by looking at the matrix Γ, we cannot predict or influence the outcome of the algorithm. In fact, the rank value of each subject is influenced by too many factors, and in particular by the citation structure, authorship and the importance of journals.

4 Numerical experiments

In [5, 6] the results of several tests on real and synthetic data using an uniform weight matrix are reported. The experiments reported in this section address two related questions. The first one is associated with the validation of the model on reliable data. In fact, as stressed in [6], real dataset are either not publicly available and usable, or so incomplete that the characteristics of the bibliographic items do not correspond to those recognized in real cases. In this respect, a generative model for building up synthetic matrices describing the subjects journals, authors and papers is proposed. The synthetic data produced agree with the properties observed on real datasets, allowing us to test the algorithm on a larger set of data, where we can evaluate the robustness of our ideas on special critical situations. For example, we plan to use synthetic data to discover malicious situation where a set of papers cites each other to increase their citations count.

The second question addressed in this section is the dependence of the rank on the choice of the weight matrix Γ as discussed in Section 3.

Since most of the tests have been performed on synthetic data, let us first discuss the generative model. The problem consists of generating the three matrices H, K and F when the parameters of the problems, that is the number of papers n_P, the number of authors n_A and the number of journals n_J variates. These parameters are strictly related one to the other by a proportionality dependence. For example, when the number of papers increases, one should expect an increase also of n_A and n_J. The factors of proportionality between n_P, n_A and n_J respect those observed in large bibliographic collections such as Mathematical Review [3] or ISI Web of Science [10].

The matrix H is a $n_P \times n_P$ boolean matrix where $H_{i,j} = 1$ if paper i contains j in its reference list. To make the model more realistic we assume that, for each paper i, we know the publication year $y(i)$ as well as the incoming $\mathcal{I}(i)$ and outgoing citations $\mathcal{O}(i)$. The matrix H has to satisfy some basic requirements as well as some statistical properties observed on real data. In particular, the matrix H needs to satisfy the following requirements.

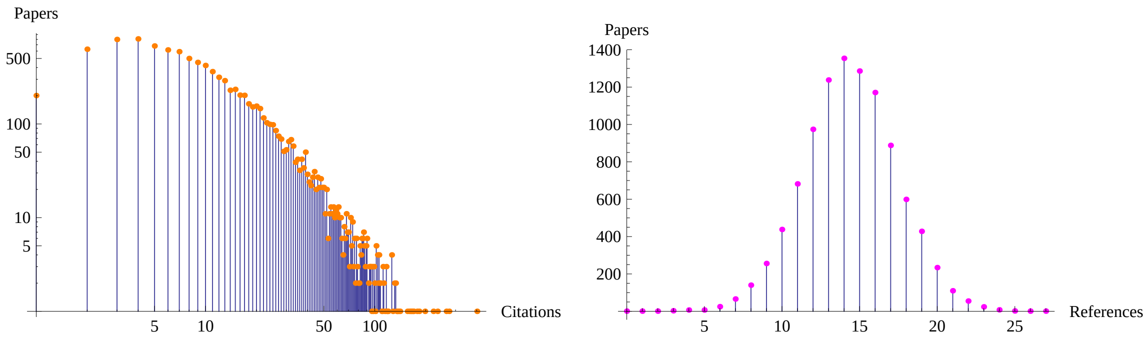
- A paper i can cite a paper j only if $y(i) \geq y(j)$, that means that i can cite only already published papers at the time i was issued.

- The distribution of the outgoing citations $\mathcal{O}(i)$ follows a normal distribution with mean 15 and variance 3. The choice of these parameters was based on the study of the distribution of the references of a large portion of the Mathematical Review database.

- As it has been observed in the literature (see for example [23] and the references therein) incoming citations $\mathcal{I}(i)$ follow a Zipf law. In fact, a few papers receive many citations while the majority are cited seldom.

The publication year is chosen randomly in a preset interval, so that the papers are distributed uniformly over the years. To implement the Zipf distribution we used a method proposed in [23]. We assign to each paper a "quality index" $Q(i)$ which follows a normal distribution, $Q(i) \in N(\mu, \sigma)$. The index of quality drives the citation process, so that the number of citation $c(i)$ that a paper i is going to receive is such that $c(i) \leq \lfloor 10^{Q(i)} - \gamma \rfloor$, where γ is a minimum standard of quality a paper needs for assuming it will receive at least a citation. For our experiments the values of $\mu = 1$ and $\sigma = 0.4$ have been chosen on the ground of the experimental results presented in [23]. It is possible to show that, in this way, the incoming citations follow a Zipf law.

The informations about the incoming and outgoing citations extracted from a randomly generated H with $n_P = 10^6$ are depicted in Figures 4.1. In particular, Figure 4.1 (a) represents an histogram in a log-log scale of papers versus citations. We see that there are many papers receiving few citations, while less than 10 papers receive hundreds of citations. In Figure 4.1(b) we see the distribution of the outgoing citations $\mathcal{O}(i)$ with a shape resembling a normal distribution (gaussian) with mean 15.



(a) (b)

Figure 4.1: Figure (a) represents s log-log scale histogram showing the distribution of papers versus citations. There are few papers with many citations, while the big majority receives less than 5 citations each. In Figure (b) the histogram showing the distribution of references on papers.

Matrix K stores information about authorship. K is a boolean matrix $n_A \times n_P$ and $K(a, p) = 1$ iff a is an author of paper p. From the analysis of real datasets [3, 13] we can see that an author has in general peaks where (s)he is more productive and periods in his (her) career where (s)he writes a minor number of papers. This leads to a model where the distribution of the publications of each author follows a normal distribution over the time with a normal standard deviation which is a proper characteristic of an author. The distribution of papers for each author follows a Zipf law, in fact we can hypothesize that a few authors publish a larger amount of papers, while the majority publish a restricted number of papers [18]. To enforce into the model the presence of coauthors, we have to assume that the sum of papers written by all the authors is greater than the number of distinct articles published. This guarantees that at the time of the matching between authors and papers, we can assign the same paper to more than one author.

Matrix F keeps track of the journal where a paper is published. Of course, this boolean matrix has only an entry equal to one for each column, since each paper cannot be published twice. The construction of F is done assigning uniformly papers to journals and forcing each journal to publish at least a paper. The histogram showing the resulting distribution is visualized in Figure 4.2.

We used the above generative algorithms to produce a dataset with one million of papers, half a million of authors and 5,000 journals, which respects the proportion of the cardinality of the classes in real databases [3]. We tested our methods with different choices of the weighting matrix Γ. In particular, in the results appearing below we denoted by $G1$ the uniform choice of the parameters γ as in equation (6) and by $G2$, $G3$ and $G4$ the choice of the weights in accordance with equation (8) for different choices of h and k. More precisely, for $h = 1, k = 1$ we have $G2$, for $h = 5, k = 1$, $G3$, and finally when $h = 1, k = 10$ we get $G4$. From the discussion carried on in Section 3, choosing as weighting technique matrix $G1$,

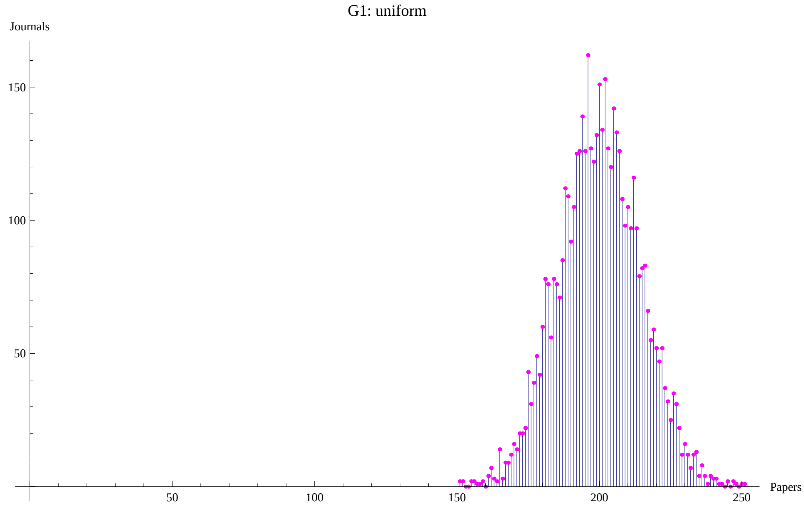


Figure 4.2: Histogram of the distribution of papers on journals. It emerges the gaussian shape.

which corresponds to using uniform weights, the rank will depend essentially on journals. With $G2$, the mean value of a generic subject is the same, independently of the class the subject belongs to. Hence, for determining the rank of a subject, on the average, we are gathering the same amount of importance from the citing papers, from authors and from the journals. The choice of weighting techniques in accordance with $G3$ or $G4$ depicts more extreme solutions, where we give more importance to authors, and more importance to citations, respectively.

Since we are interested in the relative rank rather than on the absolute value of the rank of subjects, the plot in this section are obtained by normalizing the value to span from 0 to 1.

In Figure 4.3, for the four different choices of the weights Γ, the behavior of the rank respect to the number of citations received is shown. As expected and desired, we notice a linear dependence on the number of citations, however, this dependence is less evident in the last three plots, where for the same value of rank, there are papers receiving a number of citations belonging to a quite large interval. For example, the plot corresponding to $G3$, where authors contribute more to the value of papers, the linear dependence is less strong. In particular we can identify two clusters of papers which show a different linear dependence on the number of incoming citations. This behavior might depend on the fact that the coauthor-ship matrix has different connected components, since there are independent sets of authors working together which don't have strong connections with another group of authors. With this weighting strategy, where authors have more importance than journals or papers, the reducibility of the diagonal block A_A in (2) starts to emerge. As observed experimentally, this poses also problems of convergence for higher values of the parameter h.

On the other hand, using $G1$, we see that the dependence on the citations is much stronger, since the rank of papers depends on the rank of the journals,

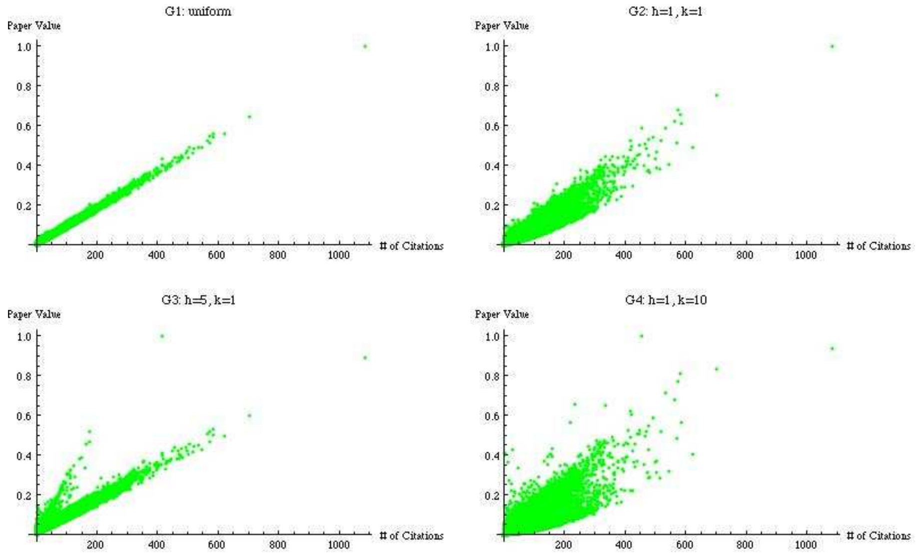


Figure 4.3: Dependence of the rank of papers from the number of citation received for the four models.

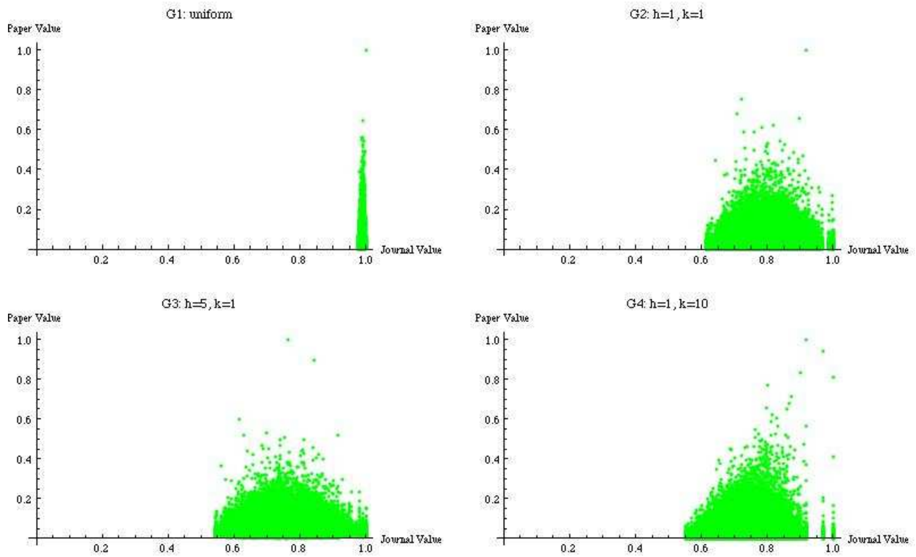


Figure 4.4: Dependence of the rank of papers from the Journal value for the four models.

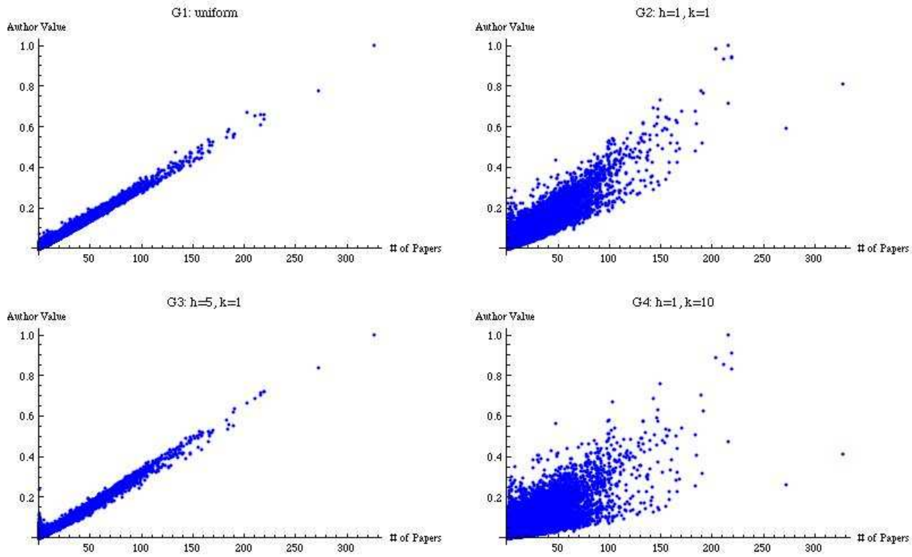


Figure 4.5: For the four models it is shown the dependence of the rank of authors on the number of papers written.

and the rank of the journals is ruled by cross citations between journals. Plots in Figure 4.4 show the dependence of the rank of papers on the rank of the journals where they are published. Again, we see that with weights $G2$ and $G4$, it is more evident that good papers are published in good journals while the contrary is not true in general. In fact, in good journals also low ranking papers can appear. When we give more importance to citations using $G4$, we see that it completely disappears the situation of good ranking papers that are published in low ranking journals. When using weight matrix $G3$ the rank depends more on the quality of the authors than on citations, and then we have high ranking papers appearing also on low ranking journals . In Figure 4.5 it is shown the dependence of the rank of authors from the number of papers written. The relationship between authors and number of papers is similar to the one between rank of papers and number of citations depicted in Figure 4.3. Again, using as weighting strategy $G2$ or $G4$, we get more interesting results. We still observe that authors publishing more papers have more chances to become important. However, the importance of an author cannot be determined by a mere counting the number of papers published, but also the quality of the papers and of the journal where the papers are published contribute to the final ranking score. Note that using $G3$, we have a clear linear dependence. It is reasonable, in fact that giving more importance to the class "Authors" we have that single authors can be compared on the basis of the number of papers written.

The rank of a journal does not depend on the number of papers it publishes, since this was especially requested (see [6]) when designing the normalization techniques of the block (1,1) in matrix A (1). The dependency on the number of citations received is linear but the plot is a sort of cloud. In this case we don't

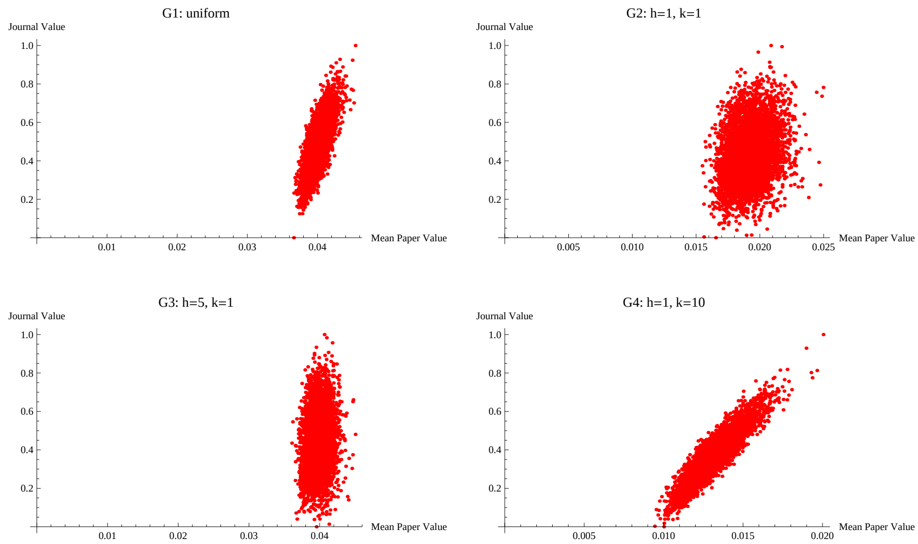


Figure 4.6: For the four models it is shown the dependence of the rank of journals on the mean value of the papers published therein.

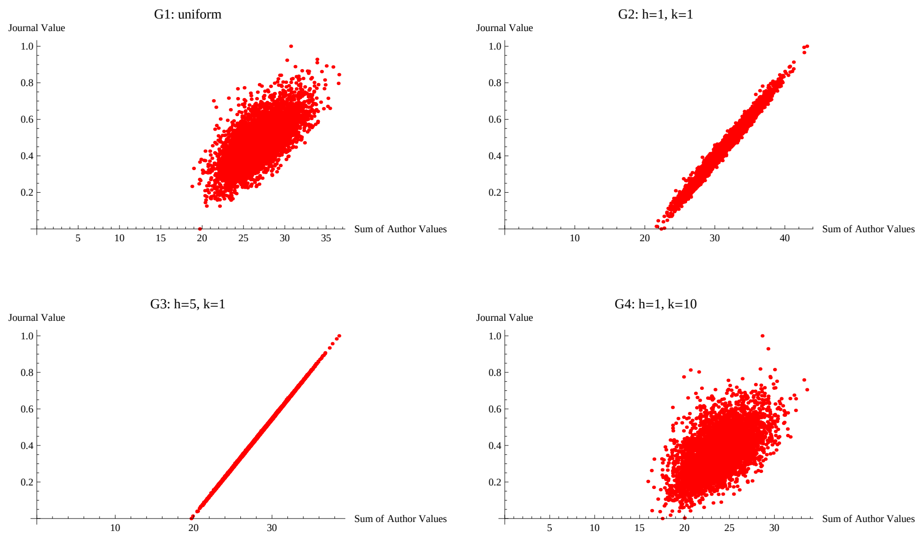


Figure 4.7: For the four models it is shown the dependence of the rank of journals on the sum of the values of the authors publishing therein.

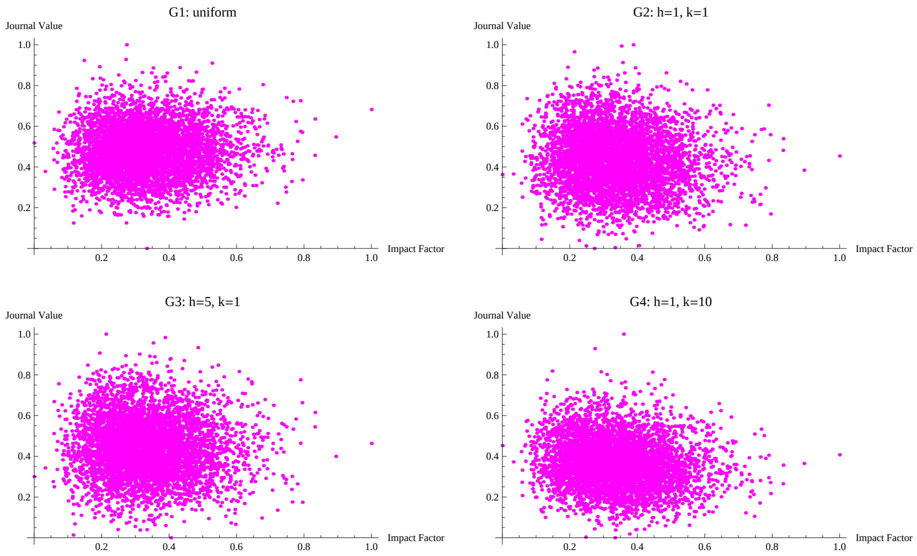


Figure 4.8: For the four models it is shown that there is not a relation between The rank of journals obtained using our method and that provided by applying the two-year Impact Factor algorithm.

have much differences among the four models.

It is interesting to comment on Figure 4.6 where the dependence on the mean value of the papers published on the journals is shown. We see that when using $G1$ and $G4$, we have a linear dependence, that is, the rank of a journal is related to the value of the average paper published therein. This is expected in the model using $G4$, since with this weighting method we are giving 10 times more importance to papers. For models based on $G2$ and $G3$, the results show that the influence of the quality of the average paper on the quality of papers is minimal. In models $G2$ and $G3$, the ranking of journals is however dominated by the sum of the importances of the authors publishing in those journals as can be observed in Figure 4.7. In the author-centedred model $G3$ we have a very neat linear dependence.

Finally, in Figure 4.8 the journal rank is plotted against the Impact Factor over a two-year period showing that independently from the weighting strategy used, the results returned by our method are profoundly different.

A more complete set of plots of the models is available at the companion web site `http://www.di.unipi.it/~romani/JAP4/JAP.html`.

For completeness we tested also the bibliographic items in the CiteSeer dataset, which can be freely downloaded from the CiteSeer web site [13] and is composed mainly of technical reports collected from the net. Since this collection doesn't contain full issues of the journals, but especially unpublished reports, most of the time the information about the journal where the paper is published is missing. As reported in [6], starting from a collection of 800,000 papers, after cleaning the database, and crossing it with a BibTeXfile to recover the information about the journal, we obtain a set of about 37,000 papers, 41,000 authors and 2829 journals. But, among the 2829 journals, more than a half (exactly 1636), appear with just one paper in the database. Because of the extreme sparsity of the journal-paper matrix F, this database doesn't represent a good dataset. However, for completeness, at the address `http://www.di.unipi.it/~romani/JAP4/JAP.html` the statistical behavior of our algorithm for the different choices for the weight matrix is shown. The results obtained for the categories *Authors* and *Papers* show similarity to those obtained with the synthetic data.

From the analysis performed, it is clear that the choice of a weight matrix rather than another, should be ruled by the particular problem one is addressing. For example, if one is interested in ranking papers on the basis of citations, model $G4$ is the more adequate. On the other hand, it can be interesting to value more coauthor-ship, because for example, we want to form a research team and we are looking for researchers with the ability of working in a team. In this case, we can use as weighting matrix something similar to $G3$, eventually lowering the value of h to give more importance to to journals and papers. Model $G2$ keeps balanced the average influence of the various classes, while with the uniform weight distribution $G1$ we provide a ranking dominated by the importance of journals.

To better understand the different choices of parameters also from a numerical viewpoint, we perform experimentally a sensitivity analysis. In particular, denote by $\bar{\pi}_\Gamma = (\bar{\pi}_j; \bar{\pi}_A, \bar{\pi}_P)^T$ the Perron vector of matrix P using as weight matrix Γ. Let $\mathcal{S}$ be the sorting operator, which applied to a vector, returns the vector sorted in a non-increasing order. To compute an approximation of the stationary distribution of P, we use the power method combined with a stopping criterion on the infinity norm of the difference between two successive iterations. Let $\pi_\Gamma^{(*)}$ be the vector obtained at convergence of our method with a stopping criterion of 10^{-15}, and let $\boldsymbol{r}_\Gamma = (\boldsymbol{r}_J; \boldsymbol{r}_A; \boldsymbol{r}_P)^T = (\mathcal{S}(\pi_J^{(*)}); \mathcal{S}(\pi_A^{(*)}); \mathcal{S}(\pi_P^{(*)}))^T$ the rank vector sorted. Denote by $\mathcal{P}_J, \mathcal{P}_A$ and $\mathcal{P}_P$ the permutation induced by the reordering, that is $\pi_J^{(*)}(\mathcal{P}_J) = \boldsymbol{r}_J$ and similarly for the class of authors and papers. Since we are interested in the rank position rather than in the numerical value of the subjects, we analyzed experimentally, for the four models, the sensitivity to the stopping criterion. Our analysis shares the same flavor of a rigorous and theoretical analysis for the Google PageRank model where a proposal of an alternative stopping condition is carried on [24]. Let $\pi_\Gamma^{(i)}$ be the approximation of $\bar{\pi}_\Gamma$ obtained after the $i-$th step of the power method, and let $\boldsymbol{r}_\Gamma^{(i)}$ the vector reordered in accordance with the permutations $\mathcal{P}_J, \mathcal{P}_\mathcal{A}$ and $\mathcal{P}_P$. In Figure 4.9 the behavior of the method separately on the three classes is depicted for the different choices of the weight matrix Γ. In particular, at each iteration is computed the number of mismatches for each class, that is, the percentage of entries of $\boldsymbol{r}_\Gamma^{(i)}$ which are not sorted in a non-increasing order. On the x-axis is reported the number of iterations and on the y-axis the percentages of mismatches, for example, a value of $S = 0.3$ at iteration i denotes that 30% of the entries are placed in the wrong position in vector $\pi_\Gamma^{(i)}$. From the four plots, we see that using $G1$ we get a faster convergence than with the other models and that the class "journal" converges always faster than

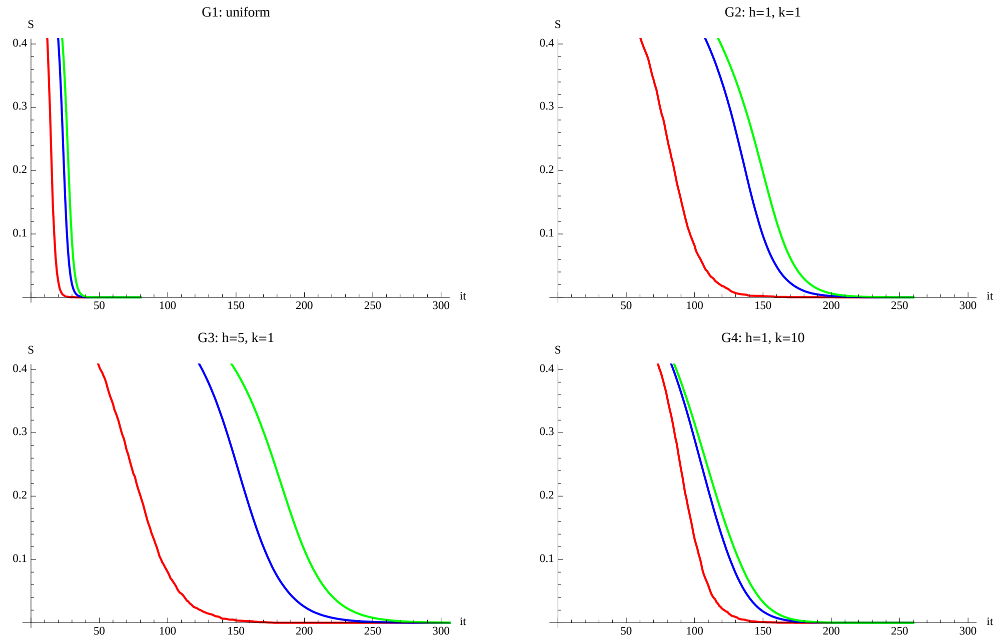


Figure 4.9: For the four models, and for each class the number of mismatches for iteration is plotted. From left to right it is represented the convergence behavior for the class journals, authors and papers.

the other classes. Moreover, using $G3$, we see that only when we have achieved convergence on journals we start converging also on the other entries of the iteration vector, while using $G4$ the convergence is almost simultaneous on the three classes. In Figure 4.10 the negative logarithm of the error at each step is plotted, where the error represents the distance of iterate i-th from $\pi_\Gamma^{(*)}$ in the infinity norm. We see that all the methods have a linear convergence, and the method obtained using $G1$ as weight matrix, achieves a better convergence. The difference in the slope of the curves depends on the closeness to 1 of the second dominant eigenvalue of the iteration matrix. As expected, the method based on $G3$ has a slower convergence, in fact with that choice of Γ, more importance is given to authors, and the co-authorship matrix is highly reducible. We notice that, although we used a stopping criterion with a tolerance of 10^{-15}, we get a precision of around 12 significant digits in the computed approximation.

5 Conclusions

In this paper we have analyzed the performance of a method for evaluating scientific literature [5, 6] on a large synthetic dataset. In particular, we performed an experimental comparison of the dependence of the ranking provided with different choices of the nine weighting parameters presented in the model. We showed that the model analyzed in [5, 6] is tunable and that versatile weighting strategies can be applied to meet the different needs of different users. In the final part of

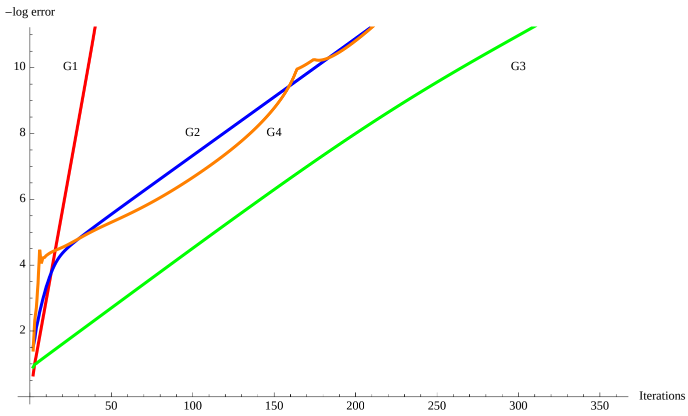


Figure 4.10: Convergence behavior of the four models. On the y-axis the negative logarithm of the error expressed as the distance from the computed solution in the infinity norm.

the paper, a sensitivity analysis is performed.

References

[1] R. Adler, J. Ewing, and P. Taylor. Citation statistics. Technical report, IMU-ICIAM-IMS, June 2008. http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf.

[2] D. W. Aksnes. A macro-study of self citations. *Scientometrics*, 52(2):235–246, 2003.

[3] AMS. MathSciNet, Mathematical Reviews on the Web. http://www.ams.org/mathscinet/.

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

[5] 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.*, 22:1–16, 2008.

[6] D. A. Bini, G. M. Del Corso, and F. Romani. A combined approach for evaluating papers, authors and scientific journals. *Journal of Computational and Applied Mathematics*, 2009. to appear.

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

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

[9] L. Egghe. Theory and practice of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[10] Institute for Scientific Information. Isi web of science. www.isiknowledge.com/.

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

[12] E. Garfield. The history and meaning of journal impact factor. *Journal of the American Medical Association*, 293:90–93, 2006.

[13] 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.

[14] W. Glänzel and B. Thijs. The influence of author self-citations on bibliometric macro indicators. *Scientometrics*, 59(3):281–310, 2004.

[15] W. Glänzel, B. Thijs, and B. Schlemmer. A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, 59(1):63–77, 2004.

[16] J. E. Hirsh. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Science*, November 15 2005.

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

[18] John C. Huber. A new method for analyzing scientific productivity. *J. Am. Soc. Inf. Sci. Technol.*, 52(13):1089–1099, 2001.

[19] C. D. Meyer. Stochastic complementation, uncoupling Markov chains and the theory of nearly reducible systems. *SIAM Rev.*, 31(2):240–272, 1989.

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

[21] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications- theory, with applications to literature of physics. *Information Processing & Management*, 12:297–312, 1976.

[22] SCImago. Sjr– scimago journal & country rank, 2007. http://www.scimagojr.com.

[23] M.J. Stringer, M. Sales-Pardo, and L. A. Nunes Amaral. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, 3(2):e1683, 2008. http://dx.doi.org/10.1371/journal.pone.0001683.

[24] R. S. Wills and I. C. F. Ipsen. Ordinal ranking for Google's PageRank. *SIAM J. Matrix Anal. Appl*, 30(4):1677–1696, 2009.

Dipartimento di Informatica, Università di Pisa,
Largo B. Pontecorvo 3,
56127 Pisa, Italy.