Electronic Transactions on Numerical Analysis.
Volume 39, pp. 464-475, 2012.

ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

A MULTIPARAMETER MODEL FOR LINK ANALYSIS OF CITATION GRAPHS*

ENRICO BOZZO† AND DARIO FASINO‡

Abstract. We propose a family of Markov chain-based models for the link analysis of scientific publications. The PageRank-style model and the dummy paper model discussed in [Electron. Trans. Numer. Anal., 33 (2008), pp. 1–16] can be obtained by a particular choice of its parameters. Since scientific publications can be ordered by the date of publication it is natural to assume a triangular structure for the adjacency matrix of the citation graph. This greatly simplifies the updating of the ranking vector if new papers are added to the database. In addition by assuming that the citation graph can be modeled as a fixed degree sequence random graph we can obtain an explicit estimation of the behavior of the entries of the ranking vector.

Key words. link analysis, citation graph, random graph, Markov chain, ranking

AMS subject classifications. 15B51, 60J20, 65C40

1. Introduction. Link analysis aims at exploring the information cached in large datasets organized as graphs or networks, to infer certain relationships between linked data. Starting from the popularization of the PageRank [8] and HITS [14] algorithms, link analysis changed the information retrieval scene in many respects, in particular by improving the effectiveness of search engines, that became able to rank by importance the retrieved information in an efficient and query independent way. The definition of the importance indicators is usually tied to a mathematical model of user navigation within the network. As a notable example, it is customary to model web surfing as a Markov chain, where the states are web pages or sites and a transition probability is associated with every hyperlink [13, 16]. In this approach, ranking is related to the mean time spent at every web site by a random surfer and is obtained by computing the invariant probability vector of the Markov chain and comparing its entries. To enforce irreducibility in the associated transition matrix, and thus guarantee the existence of a unique invariant probability vector, the PageRank algorithm modifies the original chain by allowing random jumps to arbitrary nodes that can be performed with a prescribed probability usually tuned by means of a parameter $0 \leqslant \alpha < 1$. The resulting navigation model corresponds to the Markov process which at every transition with probability α follows a random walk on the network, and with probability $1 - \alpha$ "teleports" to a random node. Altough the classical treatment of the PageRank algorithm considers uniform random jumps, the transition probabilities associated with random jumps may depend on the arrival node, in which case these probabilities are collected in the so-called personalization vector. For a comprehensive introduction to the subject see, e.g., [12, 16]. It is also worth mentioning that, since the web is not static, great attention has been paid to the problem of the influence on PageRank of link and node updating [3, 17].

Recently link analysis has been proposed also as a tool for ranking scientific authors and products; see, e.g., [6, 7, 9, 19, 20, 21] and the references therein. The Eigenfactor metric [5] is one of the most prominent examples. The underlying mathematical models rely on suitable Markov chains obtained from relationships between papers and/or authors and/or journals. For example, a collection of papers can be described as a citation graph, where every publication corresponds to a node and every citation corresponds to a directed edge. Unlike the PageRank algorithm, in [6, 7, 10] the graph is modified by adding a dummy paper

*Received August 6, 2011. Accepted for publication July 18, 2012. Published online on November 26, 2012. Recommended by C. Brezinski. This work was supported by PRIN 2008 project no. 20083KLJEZ "Problemi di algebra lineare numerica strutturata: analisi, algoritmi e applicazioni".

†Dipartimento di Matematica e Informatica, Università di Udine, Udine, Italy (enrico.bozzo@uniud.it).

‡Dipartimento di Chimica Fisica e Ambiente, Università di Udine, Udine, Italy (dario.fasino@uniud.it).

endowed by suitably weighted citations to and from all the papers in the collection. With this addition too, random walks on the citation graph give rise to a regular Markov chain, and the invariant probability vector of the chain yields a meaningful ranking of the papers in the collection. For shortness, we will refer to this as to the dummy paper model. We mention that in [19] link analysis is used in an indirect way in order to obtain improved versions of bibliometric indicators such as impact factor or h-index, and the dummy paper model is employed to solve the problem of citations outside the database. In [9, 18, 20] the authors experiment PageRank-type algorithms in order to rank by importance a set of about 350000 papers published in the American Physical Society journals. They compare the results with the ranking obtained by simply ordering the papers by the number of citations received. The two rankings are related but PageRank sometimes gives a high score even to papers with a not too high number of citations. The authors argue that this is due to the different weight that references have on the basis of the rank and the number of citations received by citing papers, and support the opinion that PageRank represents a computationally simple and effective way to evaluate the relative importance of publications beyond simply counting citations. Analogous conclusions are also drawn in [21], after a deep analysis of a citation and coauthorship network comprising about 5000 papers and 6000 authors in informetrics.

In this paper we show that the random jump and the dummy paper models belong to a wider family of models depending on n parameters $0 \leqslant \alpha_i < 1$, where $i = 1, \ldots, n$, and n is the number of papers. The parameters tune the probability of the random jump in such a way that it can be different for every node. This has to be compared with [22], where the authors actually came across the same idea while trying to introduce a temporal dimension in Web search. In addition, working with papers, it is quite natural to assume a triangular structure in the adjacency matrix that reflects the chronological order of their publication. By making this assumption, and following the approach suggested in [11, 12], we will perform an average case analysis of the family of models under the hypothesis that the citation graph is an acyclic *fixed degree sequence random graph* [1]. In this way we obtain an explicit estimate of the behavior of the entries of the ranking vector for the models of the family. Actually, the class of random graphs exploited in our average case analysis do not provide accurate models of citation networks (for example, it cannot cater to intrinsic qualities of the papers); nevertheless, fixed degree sequence random graphs are one of the most widespread and useful models of random networks; and the study of that simplified case can already furnish some valuable insight on the ranking method.

The paper is organized as follows. In Section 2 we recall the now classical model based on random jumps, introduce the dummy paper model and compare them, showing how they can be obtained by properly instantiating certain parameters spanning a family of models. In Section 3 we discuss the assumption of acyclicity in citation graphs and derive some of its consequences. In Section 4 the triangular structure of adjacency matrices is exploited in order to study in a direct way the problem of node update. In Section 5 we present an average analysis of the family of models.

2. A family of models. Given n papers numbered from 1 to n, let $A = (a_{ij})$ be the $n \times n$ matrix such that $a_{ij} = 1$ if paper i cites paper j, and $a_{ij} = 0$ elsewhere. This matrix is the adjacency matrix of the citation graph of the paper collection. Moreover, let e be the vector of appropriate order whose entries are all ones, and let $a = Ae$, $a = (a_1, \ldots, a_n)^T$. The entry a_i is the out-degree of node i and counts the number of papers in the collection that are cited by paper i. We want to define a meaningful ranking of the given set of papers, based on the invariant probability vector of a suitable Markov chain describing random walks on the citation graph. To set up the notation, we define the vector $w = (w_1, \ldots, w_n)^T$, where

ETNA
Kent State University
http://etna.math.kent.edu

for $i = 1, \dots, n$

$$w_i = \begin{cases} 1 & \text{if } a_i = 0 \\ 0 & \text{otherwise,} \end{cases}$$

and we construct the matrix $\widehat{A} = A + wp^T$, where $p = (p_1, \dots, p_n)^T$ is a positive stochastic vector sometimes called *personalization vector*. Let us set

$$\Delta = \mathrm{Diag}(\delta_1, \dots, \delta_n), \qquad \delta_i = \begin{cases} 1 & \text{if } a_i = 0 \\ 1/a_i & \text{otherwise.} \end{cases}$$

Then, the matrix $\Delta\widehat{A}$ is row stochastic, i.e., $\Delta\widehat{A}e = e$. In this paper we consider the family of Markov chains associated to the parametrized matrix

$$\Gamma = D_\alpha \Delta \widehat{A} + (I - D_\alpha)ep^T, \tag{2.1}$$

where

$$D_\alpha = \mathrm{Diag}(\alpha_1, \dots, \alpha_n), \qquad 0 \leqslant \alpha_i < 1, \quad i = 1, \dots, n,$$

and, of course I denotes the identity matrix. The random walk interpretation of a Markov chain associated to Γ is straightforward: the parameter α_i represents the probability of *not* performing a random jump starting from node i; and the number p_i is the probability of arriving in node i after a random jump. The matrix Γ is positive and, by virtue of Perron theorem (see, e.g., [16]), there exists a unique positive vector π such that $\pi^T e = 1$ and $\pi^T\Gamma = \pi^T$. The vector π is the invariant probability vector of the Markov chain associated to Γ; its entries can be used for ranking purposes, since they quantify the probability of visiting each node during random walks.

2.1. Special cases. Suitable choices for the matrix D_α give rise to special cases, already discussed in the literature. By choosing $D_\alpha = \alpha I$,

$$\Gamma = \alpha\Delta\widehat{A} + (1 - \alpha)ep^T. \tag{2.2}$$

This matrix is often referred to as the *Google matrix* of the network, and plays a fundamental role in the PageRank algorithm.

Various authors already studied the use of different values of α for every node. For example, by choosing $\alpha_i = a_i/(1+a_i)$ for $i = 1, \dots, n$ after simple passages, the expression (2.1) simplifies to

$$\Gamma = DA + Dep^T, \qquad D = \mathrm{Diag}\big(\tfrac{1}{1+a_1}, \dots, \tfrac{1}{1+a_n}\big), \tag{2.3}$$

and this can be shown to be equivalent to the addition to the network of a *dummy node* as described in [15, Section 6.3] and exploited, e.g., in [6, 7, 10]. Actually, let us ideally introduce a dummy node in the network and add a link from every node to the dummy node and in addition a link from the dummy node to node j weighted with transition probabilities p_j, for $j = 1, \dots, n$, such that the vector p is positive and stochastic. The probability of the transition from node i to node j in the modified network is the sum of the probability of a direct transition from i to j and of a transition passing through the dummy node,

$$\frac{a_{ij}}{1 + a_i} + \frac{1}{1 + a_i}p_j,$$

and this is exactly the (i, j)-entry of the matrix Γ in (2.3).

ETNA
Kent State University
http://etna.math.kent.edu

As another example, in [4] random walks are used as a tool for sampling complex networks. In order to mitigate the sensitivity of the invariant probability vector from the parameter α in (2.2) with $p = \frac{1}{n}e$ and, at the same time, accelerate the convergence of the power iteration, the authors choose to set $\alpha_i = a_i/(\alpha + a_i)$ in (2.1). The resulting transition matrix is

$$\Gamma = DA + \frac{\alpha}{n} Dee^T,$$

where

$$D = \mathrm{Diag}(d_1, \ldots, d_n), \qquad d_i = \frac{1}{\alpha + a_i}.$$

This is easily seen to be equivalent to introducing a dummy node and setting the probability of jumping from the node i to the dummy node equal to $\alpha/(\alpha + a_i)$. Note that the original dummy node approach is obtained by setting $\alpha = 1$.

2.2. The invariant probability vector. By arguments that are well known in the PageRank setting (see, e.g., [15, Section 5.2]), the invariant probability vector of Γ can be computed by solving a linear system, as shown in the following theorem.

THEOREM 2.1. *Let π be the invariant probability vector of Γ in (2.2), which is the positive vector such that $\pi^T\Gamma = \pi^T$ and $\pi^T e = 1$. Moreover, let x be the solution of*

$$x^T(I - D_\alpha \Delta A) = p^T. \tag{2.4}$$

Then, x and π differ by a multiplicative constant as

$$x^T = \frac{1}{1 - \sum_{i:a_i>0} \alpha_i \pi_i} \pi^T.$$

Proof. The eigenvalue problem $\pi^T(D_\alpha \Delta \widehat{A} + (I - D_\alpha)ep^T) = \pi^T$ can be easily recast as the linear system $\pi^T(I - D_\alpha \Delta \widehat{A}) = [\pi^T(I - D_\alpha)e]p^T$. Note that the quantity within square brackets is a scalar. From the normalization condition $\pi^T e = 1$ and the definition $\widehat{A} = A + wp^T$, with some algebra we derive

$$\pi^T(I - D_\alpha \Delta A) = [1 - \pi^T D_\alpha(e - \Delta w)]p^T.$$

Observing that $\Delta w = w$, we obtain

$$p^T(I - D_\alpha \Delta A)^{-1} = \frac{1}{1 - \pi^T D_\alpha(e - w)} \pi^T.$$

To complete the proof it suffices to expand the expression of $\pi^T D_\alpha(e - w)$. □

As a consequence of the preceding theorem, π can be obtained by normalization of x, $\pi = (1/e^T x)x$. Hence, x is equivalent to π for ranking purposes. Moreover, since by construction $||D_\alpha \Delta A||_\infty < 1$, the power series expansion

$$Z = (I - D_\alpha \Delta A)^{-1} = \sum_{k=0}^{\infty} (D_\alpha \Delta A)^k, \tag{2.5}$$

is convergent, whence we can express x also as

$$x^T = p^T Z = p^T \sum_{k=0}^{\infty} (D_\alpha \Delta A)^k.$$

ETNA
Kent State University
http://etna.math.kent.edu

Observe that, if $a_i = 0$ (that is, the ith node is dangling) then α_i does not influence the ranking in any way, since the ith row of $D_\alpha \Delta A$ is zero. Hence, we can safely assume that $\alpha_i = 0$ whenever $a_i = 0$.

Finally, note that for $D_\alpha \to O$ one has

$$x^T = p^T(I + D_\alpha \Delta A) + O(\|D_\alpha\|^2).$$

This equation can be seen as a generalization of formula (2) in [9], which in our notation becomes

$$x_i = \frac{1}{n}\left(1 + \alpha \sum_{j \to i} \frac{1}{a_j}\right) + O(\alpha^2)$$

and describes the behaviour for small α of the invariant probability vector in the PageRank model (2.2) with $p = \frac{1}{n}e$.

3. The triangularity assumption. Considering scientific publications, it is quite natural to assume that nodes and edges are added to the citation graph on a chronological basis, and that newer nodes can link only to older nodes. A consequence of this assumption is that citation graphs do not contain cycles. This remark can be found also in [18], where it is quoted as a feature that differentiates citation networks from other network topologies. By the way, to the best of our knowledge, acyclicity has not been considered as a fundamental assumption in the analysis of citation graphs and, with the exception of some experimental arguments on aging effects [9, 10, 20, 22], its consequences have not been duly analyzed in the scientific literature. Hence, in what follows we assume that the resulting graph is acyclic. In particular, we assume that the nodes are numbered in reverse order with respect to their inclusion in the collection, so that the incidence matrix A is strictly upper triangular; that is, we assign index 1 to the most recent paper in the collection, and increasing numbers are assigned according to paper age. Hence, the linear system (2.4) is upper triangular and that structure can be exploited in order to express x as a function of the α_i.

Certainly, this numbering style may seem counterintuitive as it does not reflect the way the citation network grows with time. Nevertheless, we adopt this convention since it greatly simplifies various expressions occurring in subsequent results, to be shown in the next two sections, which are derived from the formula (2.4).

As the first consequence of the triangularity assumption, the largest entries of matrix Z in (2.5) are the diagonal ones, as shown in the following simple lemma.

LEMMA 3.1. *Let $U \geqslant 0$ be a strictly upper triangular matrix of order n such that $\|U\|_\infty < 1$. Let $V = (I - U)^{-1}$. Then $V_{ii} = 1$ for $i = 1, \dots, n$ and $0 \leqslant V_{ij} < 1$ for $1 \leqslant i < j \leqslant n$.*

Proof. The claim is obvious if $n = 1$. Otherwise partition

$$U = \begin{bmatrix} 0 & u^T \\ & \hat{U} \end{bmatrix}, \qquad V = \begin{bmatrix} 1 & v^T \\ & \hat{V} \end{bmatrix} = \begin{bmatrix} 1 & -u^T \\ & I - \hat{U} \end{bmatrix}^{-1}.$$

Then $\hat{V} = (I - \hat{U})^{-1}$. Hence for $2 \leqslant i \leqslant j \leqslant n$ we obtain the claim by an inductive argument. Moreover, from $(I - U)V = I$ we get $v^T = u^T \hat{V}$, whence

$$v_i = \sum_{j=1}^{n-1} u_j \hat{V}_{ji} \leqslant \max_j \hat{V}_{ji} \sum_{j=1}^{n-1} u_j < 1.$$

The last inequality following from $\sum_{j=1}^{n-1} u_j \le \|U\|_\infty < 1$. $\square$

By virtue of Lemma 3.1, if $p \to e_i$ then the ith paper gets the highest ranking as expected.

ETNA
Kent State University
http://etna.math.kent.edu

3.1. A test case: the linear chain. As an interesting test problem, we consider the case of a linear chain of n papers, where every publication cites the next one. The resulting adjacency matrix is

$$A = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}. \tag{3.1}$$

Clearly,

$$I - D_\alpha \Delta A = \begin{bmatrix} 1 & -\alpha_1 & & \\ & 1 & \ddots & \\ & & \ddots & -\alpha_{n-1} \\ & & & 1 \end{bmatrix},$$

so that the solution of (2.4) is

$$x_1 = p_1, \qquad x_i = p_i + \alpha_{i-1} x_{i-1}.$$

This recurrence is solved by the explicit formula

$$x_i = \sum_{j=1}^{i} p_j \prod_{k=j}^{i-1} \alpha_k. \tag{3.2}$$

Note that in this specific example, the dummy paper model prescribes the values $\alpha_1 = \ldots = \alpha_{n-1} = 1/2$. More generally, if $p = \frac{1}{n}e$ and $\alpha_i = \alpha > 0$ for $i = 1, \ldots, n-1$, then

$$x_i = \frac{1}{n} \sum_{k=0}^{i-1} \alpha^k = \frac{1 - \alpha^i}{n(1-\alpha)}, \tag{3.3}$$

so that $1/n = x_1 < x_2 < \ldots < x_n < 1/(n(1-\alpha))$. Altough bounded, this increasing behavior is considered undesirable, since it assigns the highest rank to the oldest paper, while that paper does not receive citations from newer papers.

Actually, aging is a very relevant issue in citation networks since older papers tend to dominate over younger papers by drawing a considerable part of the overall score by way of long citation paths. To mitigate this outcome, one can introduce various obsolescence mechanisms which, while keeping track of all past citations, assign a larger relevance to papers that get cited by recently added papers; see, e.g., [10, 18, 20, 22]. In [22] the authors propose to tackle aging effects by setting $\alpha_i = \theta^i$, where $0 < \theta < 1$ is a parameter called *DecayRate*. (A similar device is exploited also in [10].) Under that hypothesis, in the case where $p = \frac{1}{n}e$, from (3.2) we obtain

$$x_1 = \frac{1}{n}, \qquad x_2 = \frac{1}{n}(1+\theta),$$

so that $x_1 < x_2$. Moreover, for $i \geq 2$,

$$x_i = \frac{1}{n} \sum_{j=1}^{i} \prod_{k=j}^{i-1} \theta^k \geq \frac{1}{n} \sum_{j=i-1}^{i} \prod_{k=j}^{i-1} \theta^k = \frac{1}{n}(1 + \theta^{i-1}).$$

ETNA
Kent State University
http://etna.math.kent.edu

Since

$$x_{i+1} = \frac{1}{n} + \theta^i x_i \le x_i \quad \Longleftrightarrow \quad x_i \ge \frac{1}{n(1-\theta^i)},$$

we notice that

$$\frac{1+\theta^{i-1}}{n} \ge \frac{1}{n(1-\theta^i)},$$

or equivalently,

$$\theta(1+\theta^{i-1}) \le 1,$$

is a sufficient condition in order to have $x_i \ge x_{i+1} > x_{i+2} > \dots$. The latter inequality for $i = 2$ gives the condition $\theta \le (\sqrt{5}-1)/2 \approx 0.62$ which is sufficient (actually, also necessary) in order to have $x_2 \ge x_3 > x_4 > \dots$. This fact provides a theoretical justification for the empirical choice $\theta = 1/2$ made by the authors of [22] in their experiments: it is one of the largest and simplest values that assigns the highest rank to the second most recent paper in a linear chain; and indeed, one can argue that the paper acquiring the newest citation is the one that motivated the most recent publication.

In [18, 20] the authors propose to avoid aging effects by an approach which is equivalent to setting p_i proportional to $\mathrm{e}^{-T_i/\tau}$, where T_i is the age of the ith paper and τ is a time unit. Following that approach and considering a sequence of papers equispaced in time, we set $p_i = \gamma\theta^i$, where $\gamma = (1-\theta)/(\theta-\theta^{n+1})$ is the normalization factor that makes p a stochastic vector. From equation (3.2), assuming that $\alpha_k = \alpha$ for $k = 1, \dots, n-1$, we find

$$x_i = \gamma \sum_{j=1}^{i} \theta^j \alpha^{i-j} = \begin{cases} \gamma i \alpha^i & \text{if } \theta = \alpha \\ \gamma\theta \frac{\alpha^i - \theta^i}{\alpha - \theta} & \text{otherwise.} \end{cases}$$

We see that, for any values $0 \le \alpha, \theta < 1$ and for any n, one has $x_i \to 0$ as $i \to \infty$, so that this strategy is useful in order to introduce an obsolescence effect on older papers.

4. Node update. In this section we study how the rankings vary if a new paper is added to the database. The same problem is considered in [6, 17]; our approach can be much more direct in the view of the triangularity structure of the adjacency matrix. In this section, to better compare our result with the cited references, we assume that the personalization vector has uniform entries.

Let us start from a collection of n papers, with scores $x^T = (x_1, \dots, x_n)$ given by (2.4). Let A be the adjacency matrix of the associated citation graph. To this collection, we add a new paper citing $m \geqslant 1$ papers in the collection (the case $m = 0$ is trivial). If we give the index 1 to the new paper and shift the others accordingly, the new adjacency matrix takes the form

$$\tilde{A} = \begin{bmatrix} 0 & b^T \\ 0 & A \end{bmatrix}.$$

The vector b^T describes the newly added citations. Denoting

$$\tilde{D}_\alpha = \begin{bmatrix} \tilde{\alpha} & 0 \\ 0 & D_\alpha \end{bmatrix}, \qquad \tilde{\delta} = \frac{1}{m}, \qquad \tilde{\Delta} = \begin{bmatrix} \tilde{\delta} & 0 \\ 0 & \Delta \end{bmatrix}, \tag{4.1}$$

the updated transition matrix is

$$\tilde{\Gamma} = \tilde{D}_\alpha \tilde{\Delta} \tilde{A} + \frac{1}{n+1}(I - \tilde{D}_\alpha) e e^T.$$

ETNA
Kent State University
http://etna.math.kent.edu

Due to Theorem 2.1, the normalized Perron vector of $\tilde{\Gamma}$ is a multiple of the score vector $\tilde{x}^T = (\tilde{x}_1, \dots, \tilde{x}_{n+1})$ given by $\tilde{x}^T = e^T \tilde{Z}$, where

$$\tilde{Z} = (I - \tilde{D}_\alpha \tilde{\Delta} \tilde{A})^{-1} = \begin{bmatrix} 1 & -\tilde{\alpha}\tilde{\delta} b^T \\ 0 & I - D_\alpha \Delta \tilde{A} \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \tilde{\alpha}\tilde{\delta} y^T \\ 0 & Z \end{bmatrix}, \qquad y^T = b^T Z.$$

Obviously $1 = \tilde{x}_1 \leq \tilde{x}_i$ for $i = 2, \dots, n+1$, as the added paper receives no citations. If we let $\tilde{x}^T = (1, \hat{x}^T)$ then the vector $\hat{x}^T = (\hat{x}_1, \dots, \hat{x}_n)$ contains the updated scores of the preexisting papers, and we have the updating formula

$$\hat{x}^T = x^T + \tilde{\alpha}\tilde{\delta} y^T. \tag{4.2}$$

In the componentwise sense we have $\hat{x} - x = \tilde{\alpha}\tilde{\delta} y \geqslant 0$, that is, the updated scores are not smaller than the older ones. If $b = e$, that is, the new paper cites all the preceding papers, then $y^T = x^T$ so that $\hat{x} = (1 + \tilde{\alpha}\tilde{\delta})x$ and the earlier ordering of the papers in the collection (before the addition of the new paper) is not altered. In what follows, we analyze the effect of the new citations in the general case. Before our main results, we need a preliminary lemma.

LEMMA 4.1. *The matrix $Z = (z_{ij})$ given by (2.5) is unit upper triangular, with $0 \leqslant z_{ij} < 1$ for $1 \leqslant i < j \leqslant n$. Moreover, if $x^T = e^T Z$ and $i \neq j$, then we have $z_{ij} < x_j / x_i$.*

Proof. The first part of the claim follows from Lemma 3.1. To prove the second part, let $1 \leqslant i < j \leqslant n$ and consider the partitioning

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ & Z_{22} \end{bmatrix}, \qquad I - D_\alpha \Delta A = \begin{bmatrix} I - P_{11} & -P_{12} \\ & I - P_{22} \end{bmatrix},$$

where the upper leftmost blocks have order $i \times i$. From $(I - P_{11}) Z_{12} - P_{12} Z_{22} = O$ we have $Z_{11}^{-1} Z_{12} = P_{12} Z_{22} \geqslant O$. Now, since Z_{11} is unit upper triangular, we have that

$$x_i z_{ij} = x_i e_i^T \begin{bmatrix} I & Z_{11}^{-1} Z_{12} \\ & Z_{22} \end{bmatrix} e_j \leqslant [x_1, \dots, x_i, 0, \dots, 0] \begin{bmatrix} I & Z_{11}^{-1} Z_{12} \\ & Z_{22} \end{bmatrix} e_j.$$

Moreover, $(x_1, \dots, x_i) = e^T Z_{11}$. Hence we have

$$\begin{aligned} x_i z_{ij} \leqslant \left[e^T Z_{11}, 0^T \right] \begin{bmatrix} I & Z_{11}^{-1} Z_{12} \\ & Z_{22} \end{bmatrix} e_j &= \left[e^T, 0^T \right] \begin{bmatrix} Z_{11} & \\ & I \end{bmatrix} \begin{bmatrix} I & Z_{11}^{-1} Z_{12} \\ & Z_{22} \end{bmatrix} e_j \\ &= \left[e^T, 0^T \right] Z e_j < e^T Z e_j = x_j. \end{aligned}$$

The case where $j < i$ is straightforward. $\square$

The next theorem establishes a quantitative result comparing the increase in score of papers that receive a citation from the most recent paper, with respect to that of the not cited ones, both in relative and in absolute sense; the set $\mathcal{I}$ contains the indices of the newly cited papers.

THEOREM 4.2. *Let $\mathcal{I} = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$, $b = \sum_{k \in \mathcal{I}} e_k$, and $j \notin \mathcal{I}$. In the previously introduced notation we have:*

1. *$\hat{x}_j - x_j < \tilde{\alpha} \leqslant \sum_{i \in \mathcal{I}} (\hat{x}_i - x_i)$, with equality when $m = 1$;*
2. *Let ξ be the harmonic mean of $x_{i_1}, \dots, x_{i_m}$,*

$$\xi = \frac{m}{\sum_{i \in \mathcal{I}} 1/x_i}.$$

Then, $(\hat{x}_j - x_j)/x_j < \sum_{i \in \mathcal{I}} (\hat{x}_i - x_i)/\xi$.

Proof. Let $y^T = b^T Z$. Firstly, observe that for any $1 \leqslant i \leqslant n$

$$y_i = y^T e_i = \sum_{k\in\mathcal{I}} e_k^T Z e_i = \sum_{k\in\mathcal{I}} z_{ki}.$$

From the updating formula (4.2) and Lemma 4.1, we obtain

$$\sum_{i\in\mathcal{I}} \hat{x}_i - x_i = \frac{\tilde{\alpha}}{m} \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{I}} z_{ki} \geqslant \frac{\tilde{\alpha}}{m} \sum_{i\in\mathcal{I}} z_{ii} = \tilde{\alpha}.$$

For $j \notin \mathcal{I}$, using again Lemma 4.1, we have $\sum_{i\in\mathcal{I}} z_{ij} < m$. Hence

$$\hat{x}_j - x_j = \frac{\tilde{\alpha}}{m} \sum_{i\in\mathcal{I}} z_{ij} < \tilde{\alpha},$$

and the proof of the first part of the claim is complete. Furthermore, from $z_{ij} < x_j/x_i$,

$$\frac{\hat{x}_j - x_j}{x_j} = \frac{\tilde{\alpha}}{m} \sum_{i\in\mathcal{I}} \frac{z_{ij}}{x_j} < \frac{\tilde{\alpha}}{m} \sum_{i\in\mathcal{I}} \frac{1}{x_i} = \frac{\tilde{\alpha}}{\xi} \leqslant \frac{1}{\xi} \sum_{i\in\mathcal{I}} \hat{x}_i - x_i,$$

and the proof is complete. ☐

Thus, the overall increase in score of the newly cited papers is greater than the increase in score of every single paper that does not receive a new citation. We observe that, when $m = 1$ and $\mathcal{I} = \{i\}$, the results in Theorem 4.2 take the simple form

$$\hat{x}_j - x_j < \tilde{\alpha} = \hat{x}_i - x_i, \qquad \frac{\hat{x}_j - x_j}{x_j} < \frac{\hat{x}_i - x_i}{x_i},$$

for all $j \neq i$, that is, the ith paper gets the largest score increment, both absolute and relative. In particular, in the overall ranking of the collection, the position of the ith paper cannot decrease. The rightmost inequality, in the equivalent form $\hat{x}_j/x_j < \hat{x}_i/x_i$, can also be traced in [6] for the dummy paper model.

5. An average case analysis. In this section we perform an average case analysis of a special family of acyclic random graphs. Our goal is to obtain asymptotic estimates on the behaviour of the solution of (2.4) on large citation graphs. In the random jump model, analogous results can be found in [2, 13], where asymptotic or average properties of PageRank scores are obtained for families of large, direct graphs, under additional simplifying assumptions on the network topology and with constant node out-degrees.

In a probabilistic setting, we suppose that for any two papers $1 \leqslant i < j \leqslant n$, the edge $i \to j$ may exist or not, according to a certain probability that we denote by $\mathbb{P}(i \to j)$, to be better specified later. More precisely, we consider each entry in the upper triangular part of the adjacency matrix A as a random variable A_{ij} whose distribution is binomial with parameter $\mathbb{P}(i \to j)$; the edge $i \to j$ exists if and only if $A_{ij} = 1$. We compute the mean value $\langle U \rangle$ of $U = D_\alpha \Delta A$ and we consider the properties of the solution of the linear system $x^T(I - \langle U \rangle) = e^T$, corresponding to (2.4). Although this vector cannot be interpreted as a mean Perron vector of the family, it gives some insight on what can be expected on average. In fact, a similar approach was followed in [11, 12] for the average case analysis of the HITS algorithm and some of its variants, in order to prove that various ranking methods for nodes of oriented networks lead to scores which are highy correlated with node in- and out-degrees.

In what follows, we analyze the case where we are given the integer numbers $0 \leqslant a_i \leqslant n - i$ denoting the out-degree of node i for $i = 1, \ldots, n$, that is, the numbers of papers cited

ETNA
Kent State University
http://etna.math.kent.edu

by paper i. The a_i citations of paper i are distributed uniformly among the papers $i+1,\dots,n$. Hence, we have

$$\mathbb{P}(i \to j) = \begin{cases} \frac{a_i}{n-i} & 1 \leqslant i < j \leqslant n \\ 0 & \text{else.} \end{cases}$$

This construction can be considered as an acyclic version of the *fixed degree sequence random graph* class by Aiello, Chung, and Lu [1], in which node degrees are first given and edges are randomly distributed between nodes subject to constraints of node degrees. For this reason, we call *fixed degree sequence random citation graph* the family of random graphs defined by the aforementioned construction.

The next theorem relies on the following simple facts from real analysis: (1) Let $0 \le x \le 1$. Then, $x \log 2 \le \log(1+x) \le x$. (2) Let $v_1,\dots,v_n$ and $w_1,\dots,w_n$ be arbitrary, with $v_i \ge 0$. Then there exists a number $\bar{w}$ such that $\min_i w_i \le \bar{w} \le \max_i w_i$ and $\sum_{i=1}^n v_i w_i = \bar{w}\sum_{i=1}^n v_i$. This result is a discrete version of the mean value theorem.

THEOREM 5.1. *Consider the family of fixed degree sequence random citation graphs defined by the degree sequence $a_1,\dots,a_n$. For any choiche of the coefficients $\alpha_1,\dots,\alpha_n$ such that $\alpha_i = 0$ if $a_i = 0$, let x be the solution of the linear system $x^T(I - \langle U\rangle) = e^T$. Then, there exists a number $2 \le \eta \le \mathrm{e}$, where e is Euler's number, such that*

$$1 = x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n = \eta^{\sigma_n}, \qquad \sigma_n = \sum_{i=1}^{n-1} \frac{\alpha_i}{n-i}.$$

Proof. In the case where $a_i \neq 0$, for $1 \leqslant i < j \leqslant n$, the entry U_{ij} is a random variable that assumes the value α_i/a_i with probability $\mathbb{P}(i \to j)$ and 0 otherwise. Hence, the mean value of the (i,j)-entry in the strictly upper triangular part of U is

$$\langle U\rangle_{ij} = \frac{\alpha_i}{a_i}\mathbb{P}(i \to j) = \frac{\alpha_i}{n-i}.$$

Due to the assumption that $\alpha_i = 0$ if $a_i = 0$ this formula holds also in the case where $a_i = 0$. Therefore,

$$I - \langle U\rangle = \begin{bmatrix} 1 & \beta_1 & \beta_1 & \cdots & \beta_1 \\ & 1 & \beta_2 & \cdots & \beta_2 \\ & & 1 & \ddots & \vdots \\ & & & \ddots & \beta_{n-1} \\ & & & & 1 \end{bmatrix}, \qquad \beta_i = -\frac{\alpha_i}{n-i}.$$

By direct substitution one can show that the solution of the linear system $x^T(I - \langle U\rangle) = e^T$ is

$$x = (x_1,\dots,x_n)^T, \qquad x_1 = 1, \qquad x_i = \prod_{j=1}^{i-1}(1-\beta_j) = \prod_{j=1}^{i-1}\left(1 + \frac{\alpha_j}{n-j}\right). \tag{5.1}$$

We have that $1 = x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$, with equality in the ith place if and only if $\alpha_i = 0$. By using the aforementioned facts, we obtain that there exist numbers $\varepsilon_1,\dots,\varepsilon_n,\bar{\varepsilon}$, all belonging to the interval $[\log 2, 1]$, such that

$$\log x_n = \sum_{j=1}^{n-1}\log\left(1 + \frac{\alpha_j}{n-j}\right) = \sum_{j=1}^{n-1}\varepsilon_j\frac{\alpha_j}{n-j} = \bar{\varepsilon}\sum_{j=1}^{n-1}\frac{\alpha_j}{n-j}.$$

The claim follows by taking exponentials and setting $\eta = \mathrm{e}^{\bar\varepsilon}$. $\square$

As shown in the previous theorem, the vector x assigns a paper score which depends essentially on age. In particular, using $\alpha_i < 1$ in the rightmost equation of (5.1), we obtain the following bounds for the entries of the ranking vector, that depend neither on the degree sequence $\{a_i\}$ nor on the parameters α_i:

$$1 \leqslant x_i < \prod_{j=1}^{i-1} \left(1 + \frac{1}{n-j}\right) = \frac{n}{n-i+1}, \qquad i = 2, \ldots, n.$$

Note that the upper bound on the last component diverges linearly. On the other hand, the expected number of citations received by paper n can be bounded as follows:

$$\sum_{i=1}^{n-1} \mathbb{P}(i \to n) = \sum_{i=1}^{n-1} \frac{a_i}{n-i} \leqslant (1 + \log n) \max_{1 \leqslant k \leqslant n-1} a_k,$$

so that it grows only logarithmically, whenever the degree sequence $a_1, a_2, \ldots$ is bounded. This discrepancy illustrates that the score of older papers may be overwhelmed by the propagation of the importance score from newer papers by way of longer paths in the citation network. This observation prompts the introduction of some obsolescence technique to dampen the relevance of older papers.

Actually, the magnitude of x_n indicates the departure from uniform ranking in the average scenario. In particular, it is the number $\sigma_n = \sum_{i=1}^{n-1} \alpha_i/(n-i)$ that establishes its boundedness, or the lack thereof, when new papers are added to the collection. For example, if there exists a number $\alpha^* > 0$ such that $\alpha_i \geq \alpha^*$, then we have $\sigma_n \geq \alpha^* \log(n-1)$, and we can conclude that x_n diverges at least as $n^{\bar\varepsilon\alpha^*}$, with the same number $\log 2 \leq \bar\varepsilon \leq 1$ occurring in the preceding proof. On the other hand, if the coefficients α_i form a sequence that decays to zero sufficiently fast, as in the models introduced in [10, 22] which assign to α_i a value which decreases exponentially with the age of paper i, then the numbers σ_n, and thus also the numbers x_n, are bounded independently on n.

6. Conclusions. We discussed, from a theoretical perspective, a generalized model for the ranking of scientific publications that comprehends both the PageRank-type model (2.2) and the dummy paper model (2.3). By assuming the triangularity of the adjacency matrix that represents the mutual citations we proved certain common features of this kind of hyperlinked environment, for example, that the insertion of a new citation can only increase the rank of the newly cited paper. In our generalized model each paper has its own parameter ruling the possibility to follow one of its references or to perform a random jump. We suggested a possible strategy for choosing these parameters together with the entries of the personalization vector in order to contrast aging effects. We also proposed an average case analysis considering a special class of acyclic random graphs which, in a simplified framework, quantifies the growth of the importance score of older papers due to indirect citations and provides theoretical foundation to the effectiveness of certain heuristic techniques to introduce obsolescence. We plan to extend this analysis to other families of random graphs.

REFERENCES

[1] W. AIELLO, F. CHUNG, AND L. LU, *A random graph model for massive graphs*, in Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, F. Yao and E. Luks, eds., New York, 2000, ACM Press, pp. 171–180.

[2] K. AVRACHENKOV AND D. LEBEDEV, *PageRank of scale-free growing networks*, Internet Math., 3 (2006), pp. 207–231.

[3] K. AVRACHENKOV AND N. LITVAK, *The effect of new links on Google PageRank*, Stoch. Models, 22 (2006), pp. 319–331.

[4] K. AVRACHENKOV, B. RIBEIRO, AND D. TOWSLEY, *Improving random walk estimation accuracy with uniform restarts*, in Algorithms and Models for the Web Graph, R. Kumar and D. Sivakumar, eds., Lecture Notes in Computer Science, vol. 6515, Springer, Berlin, 2010, pp. 98–109.

[5] C. BERGSTROM, *Eigenfactor: Measuring the value and prestige of scholarly journals*, College & Research Libraries News, 68 (2007), pp. 314–316.

[6] 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., 33 (2008), pp. 1–16.

[7] ———, *A combined approach for evaluating papers, authors and scientific journals*, J. Comput. Appl. Math., 234 (2010), pp. 3104–3121.

[8] S. BRIN AND L. PAGE, *The anatomy of a large-scale hypertextual web search engine*, Computer Networks, 30 (1998), pp. 107–117.

[9] P. CHEN, H. XIE, S. MASLOV, AND S. REDNER, *Finding scientific gems with Google's PageRank algorithm*, J. Informetrics, 1 (2007), pp. 8–15.

[10] G. M. DEL CORSO AND F. ROMANI, *A time-aware citation-based model for evaluating scientific products: extended abstract*, in VALUETOOLS '09: Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, G. Stea, J. Mairesse, and J. Mendes, eds., ICST, Brussels, Belgium, 2009, pp. 1–6.

[11] C. DING, X. HE, P. HUSBANDS, H. ZHA, AND H. D. SIMON, *Pagerank, HITS and a unified framework for link analysis*, in Proceedings of the 2003 SIAM International Conference on Data Mining, D. Barbara and C. Kamath, eds., SIAM, 2003, pp. 249–253.

[12] C. DING, H. ZHA, X. HE, P. HUSBANDS, AND H. D. SIMON, *Link analysis: hubs and authorities on the World Wide Web*, SIAM Rev., 46 (2004), pp. 256–268.

[13] S. FORTUNATO AND A. FLAMMINI, *Random walks on directed networks: the case of PageRank*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 2343–2353.

[14] J. M. KLEINBERG, *Authoritative sources in a hyperlinked environment*, in Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, H. J. Karloff, ed., San Francisco, CA, 1998, ACM Press, pp. 668–677.

[15] A. N. LANGVILLE AND C. D. MEYER, *Deeper inside PageRank*, Internet Math., 1 (2004), pp. 335–380.

[16] ———, *Google's PageRank and Beyond: the Science of Search Engine Rankings*, Princeton University Press, Princeton, NJ, 2006.

[17] ———, *Updating Markov chains with an eye on Google's PageRank*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 968–987.

[18] S. MASLOV AND S. REDNER, *Promise and pitfalls of extending Google's PageRank algorithm to citation networks*, J. Neuroscience, 28 (2008), pp. 11103–11105.

[19] K. ŽYCZKOWSKY, *Citation graph, weighted impact factors and performance indices*, Scientometrics, 85 (2010), pp. 301–315.

[20] D. WALKER, H. XIE, K.-K. YAN, AND S. MASLOV, *Ranking scientific publications using a simple model of network traffic*, J. Stat. Mech. Theory Exp., P06010 (2007).

[21] E. YAN AND Y. DING, *Discovering author impact: A PageRank perspective*, Information Processing & Management, 47 (2011), pp. 125–134.

[22] P. S. YU, X. LI, AND B. LIU, *On the temporal dimension of search*, in WWW Alt. '04: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, eds., ACM Press, 2004, pp. 448–449.