University of Udine

Department of Mathematics and Computer Science

PREPRINT

A multiparameter model for link analysis of citation graphs

Enrico Bozzo, Dario Fasino

Preprint nr.: 2/2011

Reports available from: http://www.dimi.uniud.it/preprints

A multiparameter model for link analysis of citation graphs

Enrico Bozzo[a,1,*], Dario Fasino[a,1]

[a] *Dipartimento di Matematica e Informatica, Università di Udine, Udine, Italy.*

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 [4] can be obtained by the suitable instantiation 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 simplify 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 random sequence graph we can obtain an explicit estimation of the behavior of the entries of the ranking vector.

Keywords: Link analysis, citation graph, random graphs, ranking.
2000 MSC: 05C80, 15A18, 15B51.

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 [8, 9]. Starting from the two papers [6, 11] 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. In fact, it is usual to model web surfing as a Markov chain, where the states are the web pages or the sites and a transition probability is associated with every hyperlink. To enforce irreducibility (and then guarantee the existence of a unique invariant distribution) the popular PageRank algorithm modifies the chain by allowing random jumps, performed with a prescribed probability usually tuned by means of a parameter $0 \leqslant \alpha < 1$, from every node to every other one. In this model, ranking is related to the mean time spent in every node by a random surfer and is obtained by computing the invariant probability vector of the modified chain and comparing its entries. 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, 14]. For a comprehensive introduction see [13].

Recently link analysis has been proposed also as a tool for ranking scientific authors and products, see e.g. [4, 5, 15] and the references therein. Usually, these models rely on suitable Markov chains obtained from relationships between papers, authors, and journals. For example, a collection of papers can be described as a citation graph, where every citation corresponds to an arc endowed by a positive transition probability. Random walks on that graph give rise to a Markov chain. In [4] the chain is modified by adding a dummy paper which cites and is cited by all the papers in the collection. With this addition, the chain becomes irreducible, and a meaningful ranking can be obtained by computing the invariant probability vector of the modified chain. For shortness, we will refer to this as to the dummy paper model. We note that in [15] link analysis is used in an indirect way in order to obtain improved version of indicators such as impact factor or h-index and the dummy paper is used to solve the problem of citation outside the database.

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 state. This can be used to obtain some form of control the models as simple examples show. In

*Corresponding author
Email addresses: enrico.bozzo@uniud.it (Enrico Bozzo), dario.fasino@uniud.it (Dario Fasino)

[1] Partly supported by PRIN 2008 project no. 20083KLJEZ "Problemi di algebra lineare numerica strutturata: analisi, algoritmi e applicazioni".

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 [8, 9], we will perform an average analysis of the family of models by making the assumptions that the citation graph can be modeled as a fixed degree sequence random graph [1, 8, 9]. In this way we obtain an explicit estimate of the behavior of the entries of the ranking vector for the models of the family.

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 a couple of examples where the adjacency matrix is chosen with triangular structure. In Section 4 the triangular structure is exploited in order to study in a direct way the problem of node update. In Section 5 we will present an average analysis of the family of models. The last section discusses an obsolescence mechanism that dampens the relevance of older papers.

2. A family of models

Given n papers numbered from 1 to n, let $A = (a_{i,j})$ be the $n \times n$ matrix such that $a_{i,j} = 1$ if paper i cites paper j, and $a_{i,j} = 0$ 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 counts the number of papers cited by paper i.

We want to define a meaningful ranking of a set of papers, based on the invariant probability vector of a suitable Markov chain describing random walks on the citation graph. We recall that primitive Markov chains ensure uniqueness, positivity, and ergodicity of the invariant probability vector [13]. For that reason, primitivity is forced in generic Markov chains usually by means of one of two main techniques, see [12, Sect. 6.3], that are described in what follows.

2.1. The random jump model

In the PageRank algorithm, link-following navigation on the graph is interleaved by random jumps to uniformly chosen nodes [6, 13]. This idea is implemented by a suitable modification of the transition matrix, obtained as follows: We define the vector $w = (w_k)$ where, for $k = 1, \ldots, n$

$$w_k = \begin{cases} 1, & a_k = 0; \\ 0, & \text{otherwise.} \end{cases}$$

and we construct the matrix $\widehat{A} = A + we^T$. Let us set $\Delta = \mathrm{Diag}(\delta_k)$, where

$$\delta_k = \begin{cases} 1/n, & \text{if } a_k = 0; \\ 1/a_k, & \text{otherwise.} \end{cases}$$

Then, the matrix $\Delta\widehat{A}$ is row stochastic. Let $0 \leqslant \alpha < 1$ be a real parameter and let us consider the convex combination,

$$G = \alpha\Delta\widehat{A} + \frac{1-\alpha}{n}ee^T. \tag{1}$$

The matrix G is positive and, by virtue of Perron theorem, see [13], there exists a unique positive vector π such that $\pi^T e = 1$ and $\pi^T G = \pi^T$. The vector π is the invariant probability vector of the Markov chain that G represents, and its entries can be used for ranking purposes.

2.2. The dummy paper model

A different technique is the one described in [12, Sect. 6.3] and exploited e.g., in [4, 5]. It involves the addition of an auxiliary node to the starting graph, hence it is particularly suited for citation graphs. We will refer to the added node interchangeably as the *dummy node* or *dummy paper*. Starting with the adjacency matrix A we introduce a dummy paper in such a way that the new adjacency matrix is

$$\begin{pmatrix} A & e \\ e^T & 0 \end{pmatrix}.$$

Hence, the dummy paper cites and is cited by all other papers. Furthermore, let

$$D = \mathrm{Diag}(d_k), \qquad d_k = \frac{1}{1+a_k}.$$

Then

$$P = \begin{pmatrix} DA & De \\ \frac{1}{n}e^T & 0 \end{pmatrix} \tag{2}$$

is row stochastic. Moreover, apart the trivial case where $A = O$, the matrix P is primitive: actually it is easy to show that P^4 is positive, i.e., there is a path of exactly four steps between any two nodes of the augmented citation graph. Since Perron theorem holds for primitive matrices, the normalized (left) Perron vector of P has positive components and can be seen as a ranking vector for the considered paper set.

2.3. A generalized model

In what follows, we define a family of models to construct a primitive Markov chain, starting from an arbitrary citation graph. The random jump model and the dummy paper model occur as special cases within this family.

Recall that the *censored chain* associated to a subset $\mathcal{S}$ of states of a given a Markov chain is the chain that records the location of the parent chain only when the parent chain visits states in $\mathcal{S}$, see e.g., [13, 14]. The transition matrix of the censored chain is the stochastic complement of the matrix of the parent chain relative to the states in $\mathcal{S}$, and its invariant probability vector is obtained by normalizing the subvector relative to the states in $\mathcal{S}$ of invariant probability vector of the parent chain.

In the dummy paper model, the stochastic complement of the matrix P in (2) relative to all the papers except the dummy one is

$$Q = DA + \frac{1}{n}Dee^T. \tag{3}$$

A different representation of Q is presented in the forthcoming theorem:

Theorem 1. *Using the notations introduced in §2.1, if we let* $D_\alpha = \mathrm{Diag}(\frac{a_k}{1+a_k})$*, then the matrix* Q *in* (3) *can be expressed as*

$$Q = D_\alpha \Delta \widehat{A} + \frac{1}{n}(I - D_\alpha)ee^T.$$

Proof. We have $D_\alpha \Delta A = DA$ (altough $D_\alpha \Delta \neq D$) and $I - D_\alpha = D$. Moreover, $D_\alpha w = 0$. Using commutativity of diagonal matrices we have

$$\begin{aligned} D_\alpha \Delta \widehat{A} + \frac{1}{n}(I - D_\alpha)ee^T &= D_\alpha \Delta (A + we^T) + \frac{1}{n}(I - D_\alpha)ee^T \\ &= D_\alpha \Delta A + D_\alpha \Delta we^T + \frac{1}{n}(I - D_\alpha)ee^T \\ &= DA + \frac{1}{n}(e + D_\alpha(w - e))e^T = DA + \frac{1}{n}Dee^T, \end{aligned}$$

and the claim follows from (3). ∎

As a result, the matrix Q in (3) can be seen as a variant of the PageRank-style matrix (1), where the scalar α is replaced by a suitable diagonal matrix; the probability of performing the random jump now depends on the starting node in such a way that it diminishes when the number of references increases.

Hence the random jump model and the dummy paper model belong to the family of Markov chains associated to the parametrized matrix

$$\Gamma = D_\alpha \Delta \widehat{A} + \frac{1}{n}(I - D_\alpha)ee^T, \tag{4}$$

where now the matrix D_α has the more general definition $D_\alpha = \mathrm{Diag}(\alpha_i)$ with $i = 1, \ldots, n$ and $0 \leqslant \alpha_i < 1$. The random walk interpretation of the Markov chain associated to Γ is straightforward: the parameter α_i represent the probability of *not* performing a random jump starting from node i. As stated in Theorem 1, by choosing $\alpha_i = a_i/(1 + a_i)$ for $i = 1, \ldots, n$ we have $\Gamma = Q$ in (3), while by choosing $\alpha_i = \alpha$ for $i = 1, \ldots, n$ then $\Gamma = G$, see (1). By the same arguments that are well known in the PageRank setting, a left Perron vector of Γ can be computed by solving a linear system:

Theorem 2. *Let π be the left Perron vector of Γ, such that $\pi^T\Gamma = \pi^T$ and $\pi^T e = 1$. Moreover, let x be the solution of*

$$x^T(I - D_\alpha \Delta A) = e^T. \tag{5}$$

Then,

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

PROOF. The claim follows by simple adaptations of the arguments found e.g., in [12, Sect. 5.2]. ∎

As a consequence, for ranking purposes x is equivalent to π. Moreover, since by construction $||D_\alpha \Delta A||_\infty < 1$, we have

$$Z = (I - D_\alpha \Delta A)^{-1} = \sum_{k=0}^{\infty} (D_\alpha \Delta A)^k, \tag{6}$$

so that

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

In particular, this explicit expression implies that $x_i \geqslant 1$ for $i = 1, \ldots, n$. Furthermore, if $a_i = 0$ then α_i does not influence the ranking in any way. Hence, we can safely assume that $\alpha_i = 0$ whenever $a_i = 0$.

3. The triangularity assumption

Considering the ranking of scientific publications, it is quite natural to assume that nodes and arcs are added to the citation graph on a chronological basis, and that newer nodes can link only to older nodes. Hence, in what follows we assume that the resulting graph is acyclic. As a consequence, the nodes can be numbered so that the incidence matrix A is strictly upper triangular, and we can use the equation (5) in order to express x as a function of the α_i.

Hereafter, we present a couple of examples in order to compare the orderings obtained by different models of the family.

Example 1. *Consider the case of a linear chain of n papers, where every publication cites the next one. The resulting adjacency matrix is*

$$A = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}. \tag{7}$$

Clearly,

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

so that the solution of (5) *is*

$$x_1 = 1, \qquad x_i = 1 + \alpha_{i-1} x_{i-1} = \sum_{j=0}^{i-1} \prod_{k=i-j}^{i-1} \alpha_k, \qquad i = 2, \ldots, n.$$

If $\alpha_i = \alpha > 0$ for $i = 1, \ldots, n-1$ then

$$x_i = \sum_{k=0}^{i-1} \alpha^k = \frac{1 - \alpha^i}{1 - \alpha}, \tag{8}$$

so that $1 = x_1 < x_2 < \ldots < x_n$. This solution includes both the random jump model and the dummy paper model, where in this specific example we have $\alpha_i = 1/2$ for $i = 1, \ldots, n-1$. On the other hand, let

$2 \leqslant p < n$ and let us assume that $\alpha_i = \alpha > 0$ for $i = 1, \ldots, p-1$ and $\alpha_i = \beta > 0$ for $p \leqslant i \leqslant n$. Then (8) *holds for $i = 1, \ldots, p$ while*

$$x_{p+k} = \frac{1-\beta^k}{1-\beta} + \beta^k \frac{1-\alpha^p}{1-\alpha} = \frac{1}{1-\beta} + \beta^k \left(\frac{1-\alpha^p}{1-\alpha} - \frac{1}{1-\beta} \right), \qquad k = 0, \ldots, n-p.$$

We observe that if

$$\frac{1-\alpha^p}{1-\alpha} - \frac{1}{1-\beta} > 0$$

i.e., if

$$\beta < \alpha \frac{1-\alpha^{p-1}}{1-\alpha^p},$$

then $x_p > x_{p+1} > \ldots > x_n$. This suggests that a suitable tuning of the coefficients α_i can be useful, for example, in order to introduce time dependent features (e.g., obsolescency) in these models.

Our second example illustrates how different parameter choices may modify the overall ranking of the papers:

Example 2. *Let us consider the adjacency matrix*

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$I - D_\alpha \Delta A = \begin{pmatrix} 1 & -\frac{\alpha_1}{3} & -\frac{\alpha_1}{3} & 0 & -\frac{\alpha_1}{3} & 0 \\ 0 & 1 & -\alpha_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\frac{\alpha_3}{2} & -\frac{\alpha_3}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & -\alpha_4 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

so that the linear system (5) *becomes $x_1 = 1$ and*

$$\begin{cases} x_2 = 1 + \frac{\alpha_1}{3} x_1 \\ x_3 = 1 + \frac{\alpha_1}{3} x_1 + \alpha_2 x_2 \\ x_4 = 1 + \frac{\alpha_3}{2} x_3 \\ x_5 = 1 + \frac{\alpha_1}{3} x_1 + \frac{\alpha_3}{2} x_3 \\ x_6 = 1 + \alpha_4 x_4. \end{cases}$$

It is easy to show that if $\alpha_i = \alpha$ for $i = 1, \ldots, 4$ then $x_3 > x_5 > x_4 > x_2 > x_1$ and $x_6 > x_5$. However,

$$x_6 > x_3 \iff x_1 + \alpha x_4 > x_3 \iff x_1 + \alpha x_1 + \frac{\alpha^2}{2} x_3 > x_3 \iff (1+\alpha) x_1 > (1 - \frac{\alpha^2}{2}) x_3$$

and since $x_3 = (1+\alpha) x_2 = (1+\alpha)(1+\alpha/3) x_1$ we obtain

$$x_6 > x_3 \iff \alpha^2 + 3\alpha - 2 > 0.$$

Hence $x_6 > x_3$ for $(-3+\sqrt{17})/2 < \alpha < 1$, while for $0 < \alpha < (-3+\sqrt{17})/2$ then $x_3 > x_6$. In the dummy paper model $\alpha_1 = 3/4$, $\alpha_2 = 1/2$, $\alpha_3 = 2/3$, $\alpha_4 = 1/2$ so that we obtain $x_3 = x_5 > x_6 > x_4 > x_2 > x_1$.

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 faced in [4, 14]; our approach can be much more direct in the view of the triangularity structure of the adjacency matrix. Here, the triangularity of the adjacency matrix corresponds to the assumption that the added paper does not receive citations from the previous papers.

Let us start from a collection of n papers, with scores $x^T = (x_1, \dots, x_n)$ given by (5). 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

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

Introducing the notations

$$\tilde{D}_\alpha = \begin{pmatrix} \tilde{\alpha} & 0 \\ 0 & D_\alpha \end{pmatrix}, \qquad \tilde{\delta} = \frac{1}{m}, \qquad \tilde{\Delta} = \begin{pmatrix} \tilde{\delta} & 0 \\ 0 & \Delta \end{pmatrix}, \tag{9}$$

the updated transition matrix is

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

Owing to Theorem 2, 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

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

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

$$\hat{x}^T = x^T + \tilde{\alpha}\tilde{\delta} y^T. \tag{10}$$

Note that, if $b = e$ 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 couple of preliminary lemmas:

Lemma 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. Partition

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

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

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

and the proof is over. ∎

Lemma 2. *The matrix $Z \equiv (z_{ij})$ given by (6) 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 1. To prove the secon part, let $1 \leqslant i < j \leqslant n$ and consider the partitioning

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

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

$$x_i z_{ij} = x_i e_i^T \begin{pmatrix} I & Z_{11}^{-1}Z_{12} \\ & Z_{22} \end{pmatrix} e_j \leqslant (x_1, \ldots, x_i, 0, \ldots, 0) \begin{pmatrix} I & Z_{11}^{-1}Z_{12} \\ & Z_{22} \end{pmatrix} e_j.$$

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

$$\begin{aligned} x_i z_{ij} \leqslant (e^T Z_{11}, 0^T) \begin{pmatrix} I & Z_{11}^{-1}Z_{12} \\ & Z_{22} \end{pmatrix} e_j &= (e^T, 0^T) \begin{pmatrix} Z_{11} & \\ & I \end{pmatrix} \begin{pmatrix} I & Z_{11}^{-1}Z_{12} \\ & Z_{22} \end{pmatrix} e_j \\ &= (e^T, 0^T) Z e_j < e^T Z e_j = x_j. \end{aligned}$$

The case where $j < i$ is straightforward. ■

Remark that, in the componentwise sense, $\hat{x} - x = \tilde{\alpha}\tilde{\delta}y \geqslant 0$, that is, the updated scores are not smaller than the older ones. The forthcoming theorem establishes a quantitative result comparing the increase in score of non cited papers with respect to that of the cited ones, both in relative and in absolute sense:

Theorem 3. *Let $\mathcal{I} = \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$, let $b = \sum_{i \in \mathcal{I}} e_i$, and let $j \notin \mathcal{I} = \{i_1, \ldots, i_m\}$. In the previously introduced notations 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 the harmonic mean of $x_{i_1}, \ldots, x_{i_m}$,*

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

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

Proof. Firstly, observe that for any $1 \leqslant j \leqslant n$

$$y_j = y^T e_j = \sum_{i \in \mathcal{I}} e_i^T Z e_j = \sum_{i \in \mathcal{I}} z_{ij}.$$

From the updating formula (10) and Lemma 2, 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 2 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 over. ■

We observe that, when $m = 1$ and $\mathcal{I} = \{i\}$, the results in the foregoing theorem 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 $j \neq i$. 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 [4] for the dummy paper model.

5. An average case analysis

In this section we perform an average case analysis of a special family of triangular random graphs, following the strategy suggested in [8, 9]. Our goal is to obtain asymptotic estimates on the behaviour of the solution of (5) on large citation graphs. In the random jump model, analogous results can be found in [2, 10], where asymptotic or average properties of PageRank scores are obtained for families of large, direct acyclic graphs, under additional simplifying assumptions on the node degrees.

In a probabilitic setting, we suppose that for any two papers $1 \leqslant i < j \leqslant n$, the arc $i \to j$ 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 of the adjacency matrix A as a random variable A_{ij} whose distribution is binomial with parameter $\mathbb{P}(i \to j)$; the arc $i \to j$ 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 (5). Although this vector cannot be interpreted as a mean Perron vector of the family, it gives some insight on what can be expected in an average case.

In what follows, we analyze the case where we are given the numbers $0 \leqslant a_i < n - i$ that denote the out-degree of node i, that is, the numbers of papers cited by paper i. The nonzero entries in the i-th row of U are spread uniformly in the positions $i + 1, \ldots, n$, and we have

$$\mathbb{P}(i \to j) = \frac{a_i}{n - i}, \qquad 1 \leqslant i < j \leqslant n.$$

It this way the citations form a *fixed degree sequence* random graph [1, 8, 9].

Theorem 4. *Suppose that the citation graph belongs to a fixed degree sequence family of random graphs defined by the degree sequence $a_1, \ldots, a_n$, with $0 \leqslant a_i < n - i$ and $\alpha_i = 0$ 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 $\min_i \alpha_i \leqslant \rho \leqslant \max_i \alpha_i$ such that*

$$1 = x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n \leqslant (\mathrm{e}n)^\rho.$$

Proof. In the case where $a_i \neq 0$, then for $1 \leqslant i < j \leqslant n$, the entry U_{ji} 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,

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

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

$$x = (x_1, \ldots, x_n)^T, \qquad x_1 = 1, \qquad x_i = \prod_{j=1}^{i-1} (1 - \beta_j) = \prod_{j=1}^{i-1} \Big(1 + \frac{\alpha_j}{n - j}\Big). \tag{11}$$

We have that $1 = x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$, with equality in the k-th place if and only if $\alpha_k = 0$. Let

$$\rho = \left(\sum_{i=1}^{n-1} \frac{\alpha_i}{n - i} \right) \Big/ \left(\sum_{i=1}^{n-1} \frac{1}{n - i} \right).$$

Clearly, we have $\min_i \alpha_i \leqslant \rho \leqslant \max_i \alpha_i$, with strict inequalities for $\min_i \alpha_i \neq \max_i \alpha_i$. From $\log(1+\varepsilon) \leqslant \varepsilon$ and $\sum_{i=1}^{n-1} 1/i \leqslant 1 + \log n$, we get

$$\log x_n \leqslant \sum_{i=1}^{n-1} \frac{\alpha_i}{n - i} = \rho \sum_{i=1}^{n-1} \frac{1}{n - i} \leqslant \rho(1 + \log n).$$

The claim follows by taking exponentials. ∎

Hence, the vector x computed above assigns a paper score which depends essentially on age. We note that the largest component diverges almost linearly, while the number of citations received by the oldest paper grows only logarithmically, whenever the degree sequence $a_1, a_2, \ldots$ is bounded. Indeed,

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

The forthcoming corollary shows an upper bound on the entries of the ranking vector, that does not depend neither on the degree sequence $\{a_i\}$ nor on the parameters α_i:

Corollary 1. *In the notations of the preceding theorem, we have*

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

Proof. Use $\alpha_i < 1$ into the rightmost equation of (11). ∎

6. Including obsolescence

As it is apparent from the previous arguments, in a cycle-free citation graph older papers tend to dominate over younger papers by drawing a considerable part of the overall score. To mitigate this outcome, we can introduce in our model an obsolescency mechanism that, while keeping track of all past citations, assign a larger relevance to papers that get cited by recently added papers.

We explain this idea using the "incremental" setting of Section 3: While adding a new paper to a collection, we replace all the current values of the parameters α_i by $\theta\alpha_i$, where $0 < \theta < 1$ is a decay factor accounting for the lack of interest of older papers due to the arrive of newest one; this modification consist in replacing the definition of $\tilde{D}_\alpha$ in (9) by

$$\tilde{D}_\alpha = \begin{pmatrix} \tilde{\alpha} & \\ & \theta D_\alpha \end{pmatrix}.$$

More sophisticated time awareness schemes, where the obsolescence factor θ depends on the year of publication of each paper, are considered in [7]. The previously discussed, time oblivious model is recovered for $\theta = 1$.

The following example shows the effect of this simple decaying scheme on the linear chain considered in Example 1:

Example 3. *If the adjacency matrix of the citation graph is the same as in* (7) *and* $\alpha_i = \alpha > 0$ *for* $i = 1, \ldots, n-1$*, then the solution of* (5) *obeys the recurrence relation*

$$x_1 = 1, \qquad x_i = 1 + \alpha\theta^{i-1} x_{i-1}, \qquad i = 2, \ldots, n.$$

Simple passages show that $\lim_{i\to\infty} x_i = 1$*. Moreover, for* $\theta < 1/(1+\alpha)$ *one has* $1 + \alpha\theta = x_2 > \ldots > x_n$*.*

Our last result addresses the impact of a nontrivial obsolescence factor on the average case analysis as carried out in the previous section. The conclusion is that, whilst paper score still grows with age, the largest score is now bounded by a constant that does not depend on n:

Corollary 2. *In the same hypotheses and notations of Theorem 4, if the obsolescence factor* $0 < \theta < 1$ *is inserted into the model, then*

$$1 = x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n < \exp(1/(1-\theta)).$$

Proof. It is sufficient to observe that the time aware model corresponds to a time oblivious model with α_j replaced by $\alpha_j\theta^{j-1}$. To get the upper bound, use the inequalities $\alpha_j\theta^{j-1}/(n-j) < \theta^{j-1}$ and $\log(1+\varepsilon) \leqslant \varepsilon$ in the rightmost equation of (11). ∎

References

[1] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for power law graphs. *Experiment. Math.*, 10(1):53–66, 2001.

[2] Konstantin Avrachenkov and Dmitri Lebedev. PageRank of scale-free growing networks. *Internet Math.*, 3(2):207–231, 2006.

[3] Konstantin Avrachenkov and Nelly Litvak. The effect of new links on Google PageRank. *Stoch. Models*, 22(2):319–331, 2006.

[4] Dario A. Bini, Gianna M. Del Corso, and Francesco Romani. Evaluating scientific products by means of citation-based models: a first analysis and validation. *ETNA*, 33:1–16, 2008-2009.

[5] Dario A. Bini, Gianna M. Del Corso, and Francesco Romani. A combined approach for evaluating papers, anthors and scientific journals. *Journal of Comput. and Appl. Math.*, 234:3104–3121, 2010.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.

[7] Gianna 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*, pages 1–6, ICST, Brussels, Belgium, Belgium, 2009. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[8] Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, and Horst D. Simon. Pagerank, HITS and a unified framework for link analysis. Technical Report 49372, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA. November 2001 (updated September 2002).

[9] Chris H. Q. Ding, Hongyuan Zha, Xiaofeng He, Parry Husbands, and Horst D. Simon. Link analysis: hubs and authorities on the World Wide Web. *SIAM Rev.*, 46(2):256–268, 2004.

[10] Santo Fortunato and Alessandro Flammini. Random walks on directed networks: the case of PageRank. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, 17(7):2343–2353, 2007.

[11] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998)*, pages 668–677, New York, 1998. ACM.

[12] Amy N. Langville and Carl D. Meyer. Deeper inside PageRank. *Internet Math.*, 1(3):335–380, 2004.

[13] Amy N. Langville and Carl D. Meyer. *Google's PageRank and beyond: the science of search engine rankings*. Princeton University Press, Princeton, NJ, 2006.

[14] Amy N. Langville and Carl D. Meyer. Updating Markov chains with an eye on Google's PageRank. *SIAM J. Matrix Anal. Appl.*, 27(4):968–987, 2006.

[15] Karol Życzkowsky. Citation graph, weighted impact factors and performance indices. *Scientometrics*, http://dx.doi.org/10.1007/s11192-010-0208-6, 2010.