RAIRO-Inf. Theor. Appl. **40** (2006) 227-253
DOI: 10.1051/ita:2006004

GRAPH FIBRATIONS, GRAPH ISOMORPHISM, AND PAGERANK *

Paolo Boldi[1], Violetta Lonati[1], Massimo Santini[1] and Sebastiano Vigna[1]

Abstract. PageRank is a ranking method that assigns scores to web pages using the limit distribution of a random walk on the web graph. A *fibration* of graphs is a morphism that is a local isomorphism of in-neighbourhoods, much in the same way a covering projection is a local isomorphism of neighbourhoods. We show that a deep connection relates fibrations and *Markov chains with restart*, a particular kind of Markov chains that include the PageRank one as a special case. This fact provides constraints on the values that PageRank can assume. Using our results, we show that a recently defined class of graphs that admit a polynomial-time isomorphism algorithm based on the computation of PageRank is really a subclass of *fibration-prime* graphs, which possess simple, entirely discrete polynomial-time isomorphism algorithms based on classical techniques for graph isomorphism. We discuss efficiency issues in the implementation of such algorithms for the particular case of web graphs, in which $O(n)$ space occupancy (where n is the number of nodes) may be acceptable, but $O(m)$ is not (where m is the number of arcs).

Mathematics Subject Classification. 05C50, 05C85, 05C60, 94C15, 60J10, 15A51.

1. Introduction

PageRank [27] is one of the most well-known measures of importance of a web page: inspired by previous works on the mutual citations for determining the relevance of scientific papers, it is based on the intuition that a web page is more

Keywords and phrases. Graph fibrations, PageRank, Markov chain with restart.

* *This work is partially supported by MIUR PRIN Project "Automi e linguaggi formali: aspetti matematici e applicativi" and by EC Project DELIS.*

[1] Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Italy; santini@dsi.unimi.it

important if it is linked to by many important pages. PageRank is one of the factors used by search engines to determine the order of answers to a query, a problem of uttermost importance that is often referred to as *web ranking*, whence the name “PageRank”.

One suggestive metaphor to describe the idea behind PageRank is the following: consider an iterative process where every web page has a certain amount of money that will at the end be proportional to its importance. Initially, all pages are given the same amount of money. Then, at each step, every page gives away all of its money to the pages it points to, distributing it equally among them: this corresponds to the interpretation of links as a way to confer importance. This idea has a limit, however, because there might exist groups of pages that “suck away” money from the system without ever returning it back. Since we want to disallow the creation of such oligopolies, we force every page to give a fixed fraction $1-\alpha$ of its money to the State; the money collected this way is then redistributed among all the pages either equally or according to some criterion, represented as a vector $\boldsymbol{v}$ whose i-th component is the fraction of money that will be given back to page i.

Such a system can be represented as a Markov chain and, as we will show, it reaches a stationary state for every $\alpha < 1$ and for every preference vector $\boldsymbol{v}$. The distribution of such stationary state is the PageRank vector.

This formulation of PageRank can be generalised in many ways, for example allowing parallel links (a choice that will result extremely useful from a technical viewpoint) and considering weighted versions, so that every page can choose how the money given to its successors should be distributed among them. These generalisations naturally lead to recast PageRank as a special case of a perturbed Markov chain [19, 29] that we call *Markov chain with restart*.

The second player in this paper is a particular kind of graph morphism, called *graph fibration* [2]. The elementary definition we shall give has appeared in many places in the scientific literature, most notably in symbolic dynamics (left/right covers [22], regular homomorphisms [25]) and spectral graph theory (divisors [28] and semicovers [14]). However, a graph morphism has an immediate interpretation as a functor between free categories, and in that case Grothendieck’s oldest notion of fibration [10] reduces exactly to the elementary definition we shall use.

The main result of the paper shows that the existence of a fibration $f : G \to B$ preserving the colour on the arcs (*i.e.*, the transition probabilities) implies certain constraints on the value of PageRank (actually: of the limit distribution of any Markov chain with restart); more precisely, the limit distribution associated to G must be fibrewise constant. This result provides a surprising link between a purely combinatorial, discrete construction and the values of a limit process.

In the last part of the paper we study the implications of our results, showing that the class of *Markovian spectrally distinguishable graphs*, introduced by Gori, Sarti and Maggini [8] as a class of graphs possessing polynomial-time isomorphism algorithms, is actually a subclass of *fibration-prime graphs*. The latter has very quick isomorphism algorithms, easily derived from partitioning algorithms developed in the eighties [4], and in fact used by Brendan McKay’s program `nauty` [23] for computing canonical labellings, automorphism groups and isomorphisms between graphs. Finally, we discuss the particular case of web graphs, in which $O(n)$ space occupancy may be acceptable, but $O(m)$ is not (here, n and m are the number of nodes and arcs, respectively).

We discuss results that are at the intersection of several areas: graph theory, Markov chains, graph-isomorphism algorithms, and ranking of web pages. Thus, we spend a significant part of the paper to introduce the definitions that are necessary to state our main results. In passing, we make a number of observations, mainly obtained from mathematical literature, that are apparently not widely known in the computer science community, and that provide more immediate proofs of some known results.

2. Graph-theoretical preliminaries

A *(directed multi)graph* G is defined by a set N_G of nodes, a set A_G of arcs, and by two functions $s_G, t_G : A_G \to N_G$ that specify the source and the target of each arc (we shall drop the subscripts whenever no confusion is possible). Given a set of colours C, we say that a graph G is *C-coloured* if it is endowed with a *colouring* function $c_G : A_G \to C$. We use the notation $G(i,j)$ for denoting the set of arcs from node i to node j, that is, the set of arcs $a \in A_G$ such that $s(a) = i$ and $t(a) = j$; the arcs in $G(i,j)$ are said to be *parallel* to one another. A graph is *separated* iff it has no parallel arcs[1]. A graph is *symmetric* iff it is endowed with an involution $(\bar{\ }) : A_G \to A_G$ such that $s(a) = t(\bar{a})$ (and consequently $t(a) = s(\bar{a})$) for all arcs $a \in A_G$. A *loop* is an arc with the same source and target. Following common usage, we denote with $G(-,i)$ the set of arcs coming into i, that is, the set of arcs $a \in A_G$ such that $t(a) = i$, and analogously with $G(i,-)$ the set of arcs going out of i. We write $d^+_G(i) = |G(i,-)|$ for the *outdegree* of i in G and $d^-_G(i) = |G(-,i)|$ for the *indegree* of i in G. The maximum outdegree (indegree) is denoted by Δ^+_G (Δ^-_G).

A *path* (of length $n \geq 0$) is a sequence $\pi = \langle i_0 a_1 i_1 \cdots i_{n-1} a_n i_n \rangle$, where $i_k \in N_G$, $a_k \in A_G$, $s(a_k) = i_{k-1}$ and $t(a_k) = i_k$. We define $s(\pi) = i_0$, $t(\pi) = i_n$, $|\pi| = n$ and let $G^*(i,j) = \{\, \pi \mid s(\pi) = i, t(\pi) = j \,\}$ (the set of paths from i to j). We shall usually omit the nodes from the sequence when at least one arc is present. We say that i *leads to* j and write $i \leadsto j$ when there is $\pi \in G^*(i,j)$ such that $|\pi| > 0$. We say that i and j *communicate* and write $i \leftrightsquigarrow j$ whenever $i \leadsto j$ and $j \leadsto i$. It is easy to observe that the reflexive closure of $\leftrightsquigarrow$ is an equivalence relation among nodes, whose classes are called the *strongly connected components* of the graph. Moreover, the relation $\leadsto$ naturally induces a partial order among such components.

An *in-tree* is a graph with a selected node r, the root, and such that every other node has exactly one directed path to the root; if t is a node of an in-tree, we sometimes use $t \to r$ for denoting the unique path from t to the root. If T is

[1] The name originates from the fact that such graphs are separated for the double negation topology in the topos of graphs – see [32].

an in-tree, we write $h(T)$ for its *height* (the length of the longest path). Finally, we write $T \restriction k$ for the tree T truncated at height k, obtained by deleting all nodes at distance greater than k from the root.

A *graph morphism* $f : G \to H$ is given by a pair of functions $f_N : N_G \to N_H$ and $f_A : A_G \to A_H$ commuting with the source and target maps, that is, $s_H \circ f_A = f_N \circ s_G$ and $t_H \circ f_A = f_N \circ t_G$ (again, we shall drop the subscripts whenever no confusion is possible). In other words, a morphism maps nodes to nodes and arcs to arcs in such a way to preserve the incidence relation. In the case of C-coloured graphs, f is a *colour-preserving* morphism if $c_G = c_H \circ f$. A morphism is *epimorphic* (or an *epimorphism*) iff f_N and f_A are both surjective. Unless otherwise stated, morphisms between trees are required to preserve the root.

2.1. Fibrations

The central concept we are going to deal with is that of *graph fibration* [2], a particular kind of graph morphism induced by the notion of fibration between categories.

Definition 2.1. A *fibration* between the graphs G and B is a morphism $f : G \to B$ such that for each arc $a \in A_B$ and for each node $i \in N_G$ satisfying $f(i) = t(a)$ there is a unique arc $\widetilde{a}^i \in A_G$ (called the *lifting of a at i*) such that $f(\widetilde{a}^i) = a$ and $t(\widetilde{a}^i) = i$.

We inherit some topological terminology. If $f : G \to B$ is a fibration, G is called the *total graph* and B the *base* of f. We shall also say that G is *fibred (over B)*. The *fibre over a node* $h \in N_B$ is the set of nodes of G that are mapped to h, and shall be denoted by $f^{-1}(h)$.

There is a very intuitive characterisation of fibrations based on the concept of local in-isomorphism: a fibration is a graph morphisms satisfying the

> **Local In-Isomorphism Property:** If $f(i) = f(j)$ there exists a (colour-preserving, if G is coloured) bijection $\psi : G(-,i) \to G(-,j)$ such that $f(s(a)) = f(s(\psi(a)))$, for all $a \in G(-,i)$.

Another possible, more geometric way of interpreting the definition of fibration is that given a node h of B and a path π terminating at h, for each node i of G in the fibre of h there is a unique path terminating at i that is mapped to π by the fibration; this path is called the *lifting of π at i*, and it is denoted by $\widetilde{\pi}^i$.

In Figure 1, we show two graph morphisms; the morphisms are implicitly described by the colours on the nodes. The morphism displayed on the left is not a fibration, as the loop on the base has no counterimage ending at the lower grey node, and moreover the other arc has two counterimages with the same target. The morphism displayed on the right, on the contrary, is a fibration. Observe that loops are not necessarily lifted to loops.

Given a graph G and a node $i \in N_G$, define the in-tree $\widetilde{G}^i$ as follows:

- the nodes of $\widetilde{G}^i$ are the finite paths of G ending in i;

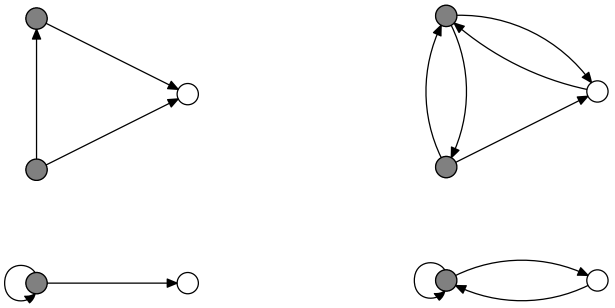

FIGURE 1. On the left, an example of graph morphism that is not a fibration; on the right, a fibration. Colours on the nodes are used to implicitly specify the morphisms.

- there is an arc from the node π to the node π' iff π starts with arc a and continues with path π' for some arc a (if G is coloured, then the arc gets the same colour as a).

We then define the graph morphism $\upsilon_G^i : \widetilde{G}^i \to G$ by mapping each node π of $\widetilde{G}^i$ (*i.e.*, each path of G ending in i) to its starting node, and each arc of $\widetilde{G}^i$ to the corresponding arc of G. It is immediate to check that υ_G^i is a fibration. We call υ_G^i the *universal fibration of G at i*, and $\widetilde{G}^i$ the *universal total graph of G at i*. Such names are motivated by the following properties. If T is an in-tree and $f : T \to G$ is a fibration that maps the root to node i, than T and $\widetilde{G}^i$ are isomorphic. Moreover, every other fibration with base G factors the universal fibration, that is, for every fibration $f : H \to G$ and for every node $j \in f^{-1}(i)$ there is a unique isomorphism $\iota : \widetilde{G}^i \to \widetilde{H}^j$ such that $\upsilon_G^i = f \circ \upsilon_H^j \circ \iota$:

$$\begin{tikzcd}
\widetilde{G}^i \arrow[r, dashed, "\iota"] \arrow[dd, "\upsilon_G^i"'] & \widetilde{H}^j \arrow[d, "\upsilon_H^j"] \\
 & H \arrow[dl, "f"] \\
G &
\end{tikzcd}$$

Now, by the universal property of universal fibrations it is immediate to see that nodes in the same fibres have the same universal total graph (we shall not distinguish isomorphic total graphs). The process can be actually reversed, as to any graph we can associate its *minimum base* $\widehat{G}$, a graph over which G is fibred, and that is *fibration prime* in the sense that it cannot be fibred nontrivially and

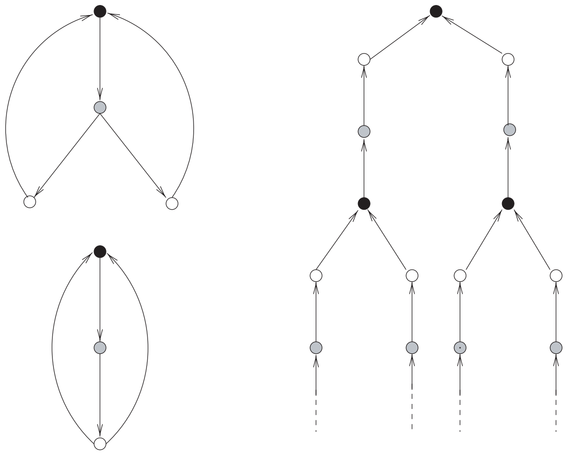

FIGURE 2. A graph (upper left), its minimum base (lower left) and the universal total graph of the black node (right). Colours on the nodes are used to specify implicitly the minimal fibration and the universal fibration.

epimorphically (*i.e.*, every epimorphic fibration $G \to B$ is an isomorphism).[2] As a matter of fact, the nodes of $\widehat{G}$ are actually the nodes of G quotiented with respect to the relation of having the same universal total graph. Hence, all fibrations from G to $\widehat{G}$ (called *minimal fibrations*) have the same node component. Figure 2 shows a graph, its minimum base and the universal total graph of a node.

The construction of $\widehat{G}$ can be made effective by observing that isomorphism of universal total graphs between nodes of the same graph is easily computable, by a result of Nancy Norris [26] that we restate in our terminology:

Theorem 2.2. *If G has n nodes, for all nodes i, j, $\widetilde{G}^i \cong \widetilde{G}^j$ iff $\widetilde{G}^i \restriction (n-1) \cong \widetilde{G}^j \restriction (n-1)$, that is, iff there is an isomorphism between the first $n-1$ levels of the two trees.*

Fibration-prime graphs are *node-rigid* – all their automorphisms are the identity on the nodes; moreover the following property holds:

Proposition 2.3. *A graph is fibration prime iff distinct nodes have non-isomorphic universal total graphs. Moreover, if two fibration-prime graphs have the same set*

[2] In fact, the partition induced by the fibres of the minimum base is the *coarsest equitable partition*, introduced in the late sixties by the community working on graph spectra. Independently, a tradition was developing in computer science about *graph partitioning*, a technique to label graph nodes in a way that is automorphism invariant [5,31]. Finally, in symbolic dynamics the minimum base of a deterministically coloured graph is the *Fischer cover* of the graph seen as a *sofic system* [22]; equivalently, if the graph is seen as a deterministic automaton all whose states are initial and final, the minimum base is the minimum automaton.

of universal total graphs, then the graphs are isomorphic, and the node component of all such isomorphisms is unique.

We also recall that every action of a group on a graph induces a fibration, since the orbits satisfy the local in-isomorphism property; in particular, a graph with an automorphism that is nontrivial on the nodes cannot be fibration prime.

3. Preliminaries about non-negative matrices and Markov chains

Let S be a finite set of *states*. A sequence of S-valued random variables $(X_k)_{k\in\mathbf{N}}$ is said to be a *(homogeneous finite) Markov chain* [16] iff for all $k > 0$ and $i_0, i_1, \dots, i_k \in S$

$$\Pr\{X_k = i_0 \mid X_{k-1} = i_1, X_{k-2} = i_2, \dots, X_0 = i_k\} = \Pr\{X_k = i_0 \mid X_{k-1} = i_1\},$$

and the right-hand side does not depend on k (whenever the left-hand side is defined). The vector[3] $\boldsymbol{p}$ defined by $p_i = \Pr\{X_0 = i\}$ is the *initial distribution*, and the matrix P defined by $P_{ij} = \Pr\{X_k = j \mid X_{k-1} = i\}$ is the *transition matrix* of the Markov chain. For any $k \geq 0$, let $\boldsymbol{p}_{(k)}$ be the vector of the (marginal) probability distribution of X_k, *i.e.*,

$$\left(\boldsymbol{p}_{(k)}\right)_i = \Pr\{X_k = i\}$$

and in particular $\boldsymbol{p}_{(0)} = \boldsymbol{p}$. As it is easy to verify that $\boldsymbol{p}_{(k)}^T = \boldsymbol{p}^T P^k$, the entire behaviour of the chain is established by its initial distribution and transition matrix.

Note that $\|\boldsymbol{p}\| = 1$, and that P is *stochastic*, that is, all its rows are distributions or, equivalently, $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the vector whose components are all 1's. Such a matrix P naturally defines a separated graph with node set S, and with arcs coloured on $(0\,..\,1]$ corresponding to non-null transitions.

More generally, any non-negative square matrix M naturally defines a separated $\mathbf{R}^+$-coloured[4] graph G where N_G is the set of indices of M whereas $A_G = \{\,\langle i,j\rangle \mid M_{ij} > 0\,\}$, $s(\langle i,j\rangle) = i$, $t(\langle i,j\rangle) = j$ and $c_G(\langle i,j\rangle) = M_{ij}$. Conversely, given an $\mathbf{R}^+$-coloured graph G, one can consider the matrix M having N_G as set of indices, and $M_{ij} = \sum_{a\in G(i,j)} c_G(a)$. This correspondence restricts to a bijection between the set of separated $\mathbf{R}^+$-coloured graphs and the set of non-negative square matrices. In the following, when no confusion is possible we will denote both a matrix and the corresponding graph with the same letter and we will say that a graph $\mathbf{R}^+$-coloured is *stochastic* iff the associated matrix is.[5]

[3] In this paper, all vectors are *column* vectors, the vector norm $\|-\|$ is L_1 and a non-negative vector with norm 1 is a *distribution* over the set of its indices.

[4] We use $\mathbf{R}^+$ to denote the set of positive real numbers.

[5] We observe that sometimes “stochastic graph” is used to mean a graph generated by some stochastic process; here, on the other hand, the graph, and its colouring, are fixed.

Essentially, a stochastic graph is a convenient way to represent the transition matrix of a Markov chain, with the additional freedom of being able to specify multiple arcs between states. Note that traditionally graphs have been used to define *random walks* [20] – a typical example of a Markov chain: in that case, the states of the chain are the nodes of an undirected graph, and the transitions from a node to its neighbours are equiprobable. In our setting, this is equivalent to representing the undirected graph as a a symmetric graph and setting the colour of an arc a to $1/d^+(s(a))$. More generally, every (not necessarily symmetric) graph G without sinks[6] can be coloured as above. We call the colouring so obtained the *natural random-walk colouring* and the associated Markov chain the *natural random walk on G*.

As we have already observed, the behaviour of a Markov chain mainly depends on the properties of its transition matrix. For this reason, in the next section we will recall some basic facts about non-negative matrices.

3.1. Non-negative matrices

Given a non-negative matrix M, we say that M is *primitive* if there exists a positive integer k such that all entries of M^k are positive and that M is *irreducible* if, for any i and j, there exists a positive integer k such that $(M^k)_{ij} > 0$. It is easy to verify that M is irreducible iff its graph is strongly connected (*i.e.*, iff it has one single strongly connected component). The *period* of an index i is defined as $\gcd\{\, k > 0 \mid (M^k)_{ii} > 0 \,\}$; an index is said to be *aperiodic* if its period is 1. It is well known that indices in the same strongly connected component have the same period, so that, in particular, it is possible to define the period of an irreducible matrix. Moreover, if $M_{ii} > 0$ (*i.e.*, if there is a loop at i) then the strongly connected component including i is aperiodic. A matrix is primitive iff it is irreducible and aperiodic.

An important result on irreducible matrices (see, *e.g.*, [30]) is the Perron–Frobenius Theorem, stating that every non-negative irreducible matrix M has a positive eigenvalue, equal to its spectral radius $\rho(M)$, associated with a positive eigenvector. If M is reducible a weaker statement holds: $\rho(M)$ is a (possibly null) eigenvalue of M and there exists a non-negative eigenvector associated with it.

Indices can be classified as follows. An index i is said to be *inessential* if there exists a j such that $i \rightsquigarrow j$, but $j \not\rightsquigarrow i$, or if i leads to no index at all (this happens if the i-th row of M is null), otherwise it is said to be *essential*. It is easy to check that indices in the same strongly connected component are all of the same kind, and that the essential components are the maximal elements of the partial order induced by $\rightsquigarrow$. In particular, the indices of an irreducible matrix are all essential. Note also that if every row of M has at least a positive entry, then there exists at least one essential index (hence, one essential component). In particular, this implies that any stochastic matrix has at least one essential index.

[6] A node i of a graph G is a sink if $G(i, -) = \varnothing$.

Finally, a non-negative matrix is said to be *unichain* if all its essential indices form one single strongly connected component; observe that a unichain matrix with no inessential indices is irreducible.

3.2. Invariant and limit distribution

Going back to Markov chains, a chain is said to be *irreducible*, *primitive*, *cyclic* or *unichain* according to whether the transition matrix P is of this kind. Analogously, the states in S can be classified into *periodic, essential* or *inessential* according to the properties of the indices of the corresponding matrix.

A distribution $\boldsymbol{p}$ is *invariant* for P if $\boldsymbol{p}^T P = \boldsymbol{p}^T$ (that is, if $\boldsymbol{p}$ is a left eigenvector of P). Also, given a distribution $\boldsymbol{p}$, when $\lim_{k\to\infty} \boldsymbol{p}^T P^k$ is defined, it is called the *limit distribution* of $\boldsymbol{p}$ under P. Observe that if $\lim_{k\to\infty} P^k$ is defined, then $\lim_{k\to\infty} \boldsymbol{p}^T P^k = \boldsymbol{p}^T(\lim_{k\to\infty} P^k)$, but the left-hand side of previous identity can be defined even if the right hand side is not. A way to understand the limit behaviour of the chain is to consider the *Cesàro limit*

$$P^* = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k,$$

that, as it is well known [16], is always defined and is equal to $\lim_{k\to\infty} P^k$ whenever the latter is defined (in such a case, the latter is denoted by P^∞). In particular, it holds that

$$P^*P = PP^* = (P^*)^2 = P^*$$

a fact that allows to draw very general conclusions about the invariant and limit distribution of the chain, as summarised by the following

Proposition 3.1. *Let P be a stochastic matrix. A distribution $\boldsymbol{p}$ is invariant for P iff $\boldsymbol{p}^T = \boldsymbol{q}^T P^*$ for some $\boldsymbol{q}$. Moreover $\boldsymbol{p}$ is the limit distribution of a given $\boldsymbol{q}$ under P iff $\boldsymbol{p}^T = \boldsymbol{q}^T P^*$. Finally, the limit $\lim_{k\to\infty} \boldsymbol{q}^T P^k$ is defined for every distribution $\boldsymbol{q}$ iff $\lim_{k\to\infty} P^k$ is defined.*

Notice that a simple consequence of the previous proposition is that there is always at least one invariant distribution.

Due to the previous considerations, the long-term behaviour of a Markov chain is completely specified by the Cesàro limit of its transition matrix. The limit depends on the properties of P. Indeed, P is unichain iff the Cesàro limit satisfies $P^* = \mathbf{1}\boldsymbol{p}^T$, where the positive entries of $\boldsymbol{p}$ correspond to the normalised left Perron eigenvector of the irreducible submatrix of P corresponding to essential indices [24]. By Proposition 3.1, this is equivalent to the following two conditions: there is a unique invariant distribution; there is a unique limit distribution (albeit it might happen that for some distribution $\boldsymbol{q}$ the limit $\lim_{k\to\infty} \boldsymbol{q}^T P^k$ is not defined). If we consider also the periodicity of indices, this leads to

Proposition 3.2. *If P is a unichain stochastic matrix such that its essential indices are aperiodic, then $\lim_{k\to\infty} P^k = \mathbf{1}\boldsymbol{p}^T$, where $\boldsymbol{p}$ is the unique invariant distribution of P, and $\lim_{k\to\infty} \boldsymbol{q}^T P^k = \boldsymbol{p}$ for every distribution $\boldsymbol{q}$.*

We conclude by noting in passing that the literature on stochastic processes uses a slightly different terminology. Since

$$\Pr\{X_k = i \text{ for infinitely many } k \mid X_0 = i\}$$

is equal to 1 or 0 according to whether i is essential or inessential, respectively, it is usual to call essential states *recurrent* and inessential ones *transient.* Moreover, if the initial distribution of a Markov process is invariant, then the process is stationary. On the other hand, if the transition matrix is irreducible and the Markov chain is stationary, then the initial distribution is the only invariant distribution. For this reason, the invariant distribution of an irreducible Markov chain is also named *stationary.*

4. Markov chains with restart

We are finally going to introduce formally the *raison d'être* of this paper: PageRank. First note that the link structure of the web can be represented by the *web graph*, whose nodes are web pages and arcs correspond to links. One could try to assign a greater rank to pages that have a higher component in the limit distribution of the natural random walk on the web graph. However, such an approach presents some problems: what initial distribution should be chosen? Will the limit distribution be unique? How fast will the process converge to the limit? A way [27] to overcome all these problems is to perturb the random walk so to make it unichain and to be able to tune its convergence speed.

Here we extend this idea from the random walk related to PageRank to any Markov chain; in this way we can highlight several connections of PageRank and derived ranking schemes with previous research on perturbed Markov chains, providing easy and structured proofs of several useful results.

Perturbation theory of linear operators is a classic field [19] and several results are known for the case of Markov chains [29]. A case of particular interest regards *analytic perturbations*, *i.e.*, the study of $P(\varepsilon) = P + \varepsilon P_1 + \varepsilon^2 P_2 + \varepsilon^3 P_3 + \dots$, where P and $P(\varepsilon)$ are stochastic matrices, for a small enough $\varepsilon > 0$ and for some matrices $P_1, P_2, P_3, \dots$; when $0 = P_2 = P_3 = \dots$, the perturbation is said to be *linear.*

Given a stochastic matrix P, a distribution $\boldsymbol{v}$, and a real $\alpha \in [0\,..\,1)$, we define the matrix

$$\mathscr{R}(P, \boldsymbol{v}, \alpha) = \alpha P + (1 - \alpha)\mathbf{1}\boldsymbol{v}^T.$$

It is easy to see that $\mathscr{R}(P, \boldsymbol{v}, \alpha)$ constitutes a linear perturbation of P for $P_1 = \mathbf{1}\boldsymbol{v}^T - P$ and $\varepsilon = 1 - \alpha$. A Markov chain with transition matrix $\mathscr{R}(P, \boldsymbol{v}, \alpha)$ has the following interpretation as a stochastic process: at every time step the next state is chosen with probability α according to the transition probabilities given by P or, with probability $1 - \alpha$, the chain is "restarted" at state i with probability v_i. For this reason we call such a process a *Markov chain with restart.* As anticipated, the introduction of the perturbation is justified by the following

Theorem 4.1. *For every stochastic matrix P and distribution $\boldsymbol{v}$, if $\alpha \in [0\,..\,1)$, then*

- *for every j such that $v_j > 0$, j is essential and aperiodic for $\mathscr{R}(P, \boldsymbol{v}, \alpha)$;*
- *$\mathscr{R}(P, \boldsymbol{v}, \alpha)$ is unichain and all its essential indices are aperiodic.*

Proof. Let $R = \mathscr{R}(P, \boldsymbol{v}, \alpha)$. If j is such that $v_j > 0$, then due to the contribution of $(1-\alpha)\mathbf{1}\boldsymbol{v}^T$ it is immediate to conclude that $R_{ij} > 0$ for every index i, hence j is essential; moreover there is a loop in j, so it is also aperiodic. Let now i be an essential index of R and consider a j such that $v_j > 0$ (such index must exist, since $\boldsymbol{v}$ is a distribution); again, due to the contribution of $(1-\alpha)\mathbf{1}\boldsymbol{v}^T$, there is an arc from i to j, but $j \rightsquigarrow i$ in R, otherwise i would be inessential; hence i and j are in the same strongly connected component. □

Theorem 4.1 together with Proposition 3.2 imply that, for every stochastic matrix P, any Markov chain with restart having transition matrix $\mathscr{R}(P, \boldsymbol{v}, \alpha)$ has a unique limit (and invariant) distribution $\boldsymbol{r}(P, \boldsymbol{v}, \alpha)$.

In this setting, PageRank as defined in [27] is the limit distribution $\boldsymbol{r}(W, \mathbf{1}/|N_W|, 0.85)$, where W is the web graph endowed with the natural random-walk colouring[7]. More generally, PageRank has been studied as the limit distribution $\boldsymbol{r}(W, \boldsymbol{v}, \alpha)$ when $\boldsymbol{v}$ is an arbitrary *preference vector* [12] or considering the *damping factor* α as a real parameter [1].

We now recast some known results, originally obtained studying PageRank, in our more general framework. First of all, observe that Theorem 4.1 can be obtained (albeit in a more algebraic and less intuitive way) noting that the perturbation induces a strong separation in the spectrum of the matrix: if $1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ is the spectrum of P, then the spectrum of $\mathscr{R}(P, \boldsymbol{v}, \alpha)$ is known [6, 13] to be

$$1 > \alpha\lambda_2 \geq \cdots \geq \alpha\lambda_k.$$

Moreover, the fact that the second largest eigenvalue of $\mathscr{R}(P, \boldsymbol{v}, \alpha)$ is less then or equal to $\alpha < 1$ implies

$$\left\| \boldsymbol{q}^T \mathscr{R}(P, \boldsymbol{v}, \alpha^k) - \boldsymbol{r}(P, \boldsymbol{v}, \alpha^T) \right\| = O(\alpha^k)$$

independently of $\boldsymbol{q}$. This is a very relevant fact from the application point of view, as the limit distribution can be efficiently obtained by successive (left) multiplication of $\boldsymbol{q}_{(k)}$ by $\mathscr{R}(P, \boldsymbol{v}, \alpha)$ (the well-known Power Method [24]).

Viewing $\boldsymbol{r}(P, \boldsymbol{v}, \alpha)$ as the invariant distribution, one can also obtain the closed form [13]

$$\boldsymbol{r}(P, \boldsymbol{v}, \alpha) = (1-\alpha)\boldsymbol{v}^T(I - \alpha P)^{-1}, \qquad (1)$$

where $(I - \alpha P)^{-1}$ is defined since $I - \alpha P$ is non-singular for every $\alpha < 1$.

[7] Provided that W has no sinks; otherwise, sinks must be patched by adding links to all other nodes.

The behaviour of such invariant distribution and its relationship with α has been deeply investigated in [1], where the following Maclaurin expansion was obtained

$$\boldsymbol{r}(P, \boldsymbol{v}, \alpha) = \boldsymbol{v}^T + \sum_{k=1}^{\infty} \alpha^k \boldsymbol{v}^T (P^k - P^{k-1})$$

together with a closed form for the derivatives of any order with respect to α. Incidentally, we observe that such results could also be directly obtained from [29] where a Maclaurin expansion is given for the more general perturbation $\alpha P + (1-\alpha)P_1$, where P_1 is stochastic and unichain.

The behaviour of $\boldsymbol{r}(P, \boldsymbol{v}, \alpha)$ at the boundary values of α is given by

$$\boldsymbol{r}(P, \boldsymbol{v}, 0) = \boldsymbol{v}^T \qquad \text{and} \qquad \lim_{\alpha \to 1^-} \boldsymbol{r}(P, \boldsymbol{v}, \alpha) = \boldsymbol{v}^T P^* \tag{2}$$

where the first identity is trivial and the second one (which, by the way, confirms a conjecture stated in [1]) can be obtained as follows. The *resolvent* of a stochastic matrix P is the linear operator $R(\mu, P) = (\mu I - P)^{-1}$, defined for every μ which is not an eigenvalue of P; it can be expanded into a *Laurent series* around every eigenvalue of P [33, Chapter VIII, Section 8]. In particular, the expansion around 1 is

$$R(\mu, P) = \frac{P^*}{\mu - 1} + \sum_{k=0}^{\infty} (\mu - 1)^k Q^{k+1}$$

for a suitable matrix Q. This implies that

$$\lim_{\mu \to 1^+} (1 - \mu) R(\mu, P) = P^*,$$

whence, by applying (1), we get the limit (2).

5. Markov chains with restart and fibrations

It is now time to present our main result. We are going to relate fibrations and Markov chains with restart, by showing that the limit distribution (and thus PageRank values) along a fibre must be constant. This provides, by means of a purely combinatorial construction, an exact constraint on a limit process.

Given an $\mathbf{R}^+$-coloured graph G and a non-negative vector $\boldsymbol{v}$ over N_G, we define the formal power series vector $\boldsymbol{z}(G, \boldsymbol{v}, \alpha)$ as

$$\boldsymbol{z}(G, \boldsymbol{v}, \alpha) = (1 - \alpha) \boldsymbol{v}^T \sum_{k=0}^{\infty} \alpha^k G^k.$$

If the spectral radius of G is not greater than 1 and $\alpha \in [0\,..\,1)$, then the series converges to $(1 - \alpha)\boldsymbol{v}^T (I - \alpha G)^{-1}$. Recalling (1), this implies the following

Theorem 5.1. *For every stochastic graph G, distribution $\boldsymbol{v}$ and $\alpha \in [0\,..\,1)$, the vector $\boldsymbol{z}(G, \boldsymbol{v}, \alpha)$ is the unique invariant distribution of the matrix $\mathscr{R}(G, \boldsymbol{v}, \alpha)$, that is, $\boldsymbol{z}(G, \boldsymbol{v}, \alpha) = \boldsymbol{r}(G, \boldsymbol{v}, \alpha)$.*

The series $\boldsymbol{z}(G, \boldsymbol{v}, \alpha)$ can be expressed in terms of the paths in G. First notice that the colour function c of the graph can be extended by multiplication to all paths by setting

$$c(\langle i_0 a_1 i_1 \cdots i_{k-1} a_k i_k \rangle) = \prod_j c(a_j).$$

The definition is motivated by the fact that, if G is separated and stochastic, then

$$c(\langle i_0 a_1 i_1 \cdots i_{k-1} a_k i_k \rangle) = \Pr\{X_k = i_k, X_{k-1} = i_{k-1}, \ldots, X_1 = i_1 \mid X_0 = i_0\}$$

where $(X_k)_{k\in\mathbf{N}}$ is any Markov chain with transition matrix G. Note that, in particular, for a 0-length path π, we have $c(\pi) = 1$. Then, for every $i, j \in N_G$ and $k \in \mathbf{N}$ we have

$$\left(G^k\right)_{ij} = \sum_{\pi \in G^*(i,j), |\pi| = k} c(\pi),$$

and hence

$$z_j(G, \boldsymbol{v}, \alpha) = (1-\alpha) \sum_{\pi \in G^*(-,j)} \alpha^{|\pi|} v_{s(\pi)}\, c(\pi) \tag{3}$$

for every node $j \in N_G$.

The values of the formal series $\boldsymbol{z}(-,-,-)$ are preserved by fibrations, provided that the vectors involved are suitably transformed. Given a fibration $f : G \to B$ and a non-negative vector $\boldsymbol{u}$ over N_B, we define the *lifting* of $\boldsymbol{u}$ along f as the vector $\boldsymbol{u}^f$ over N_G such that $\left(\boldsymbol{u}^f\right)_i = u_{f(i)}$.

Theorem 5.2. *For every colour-preserving fibration $f : G \to B$ and non-negative vector $\boldsymbol{u}$ over N_B,*

$$\boldsymbol{z}(G, \boldsymbol{u}^f, \alpha) = \boldsymbol{z}(B, \boldsymbol{u}, \alpha)^f.$$

Proof. We must prove that $z_i(G, \boldsymbol{u}^f, \alpha) = z_{f(i)}(B, \boldsymbol{u}, \alpha)$ for every node node i of G. Let us extend f to paths and consider its restriction to $G^*(-,i)$, which maps paths in $G^*(-,i)$ to paths in $B^*(-, f(i))$. The restriction is a bijection because f is a fibration, and thus paths lift uniquely. Hence,

$$\begin{aligned}\sum_{\pi \in G^*(-,i)} \alpha^{|\pi|} \left(\boldsymbol{u}^f\right)_{s(\pi)} c(\pi) &= \sum_{\pi \in G^*(-,i)} \alpha^{|f(\pi)|} u_{f(s(\pi))}\, c(f(\pi)) \\ &= \sum_{\xi \in B^*(-,f(i))} \alpha^{|\xi|} u_{s(\xi)}\, c(\xi),\end{aligned}$$

and the result follows from (3). □

In particular, this implies that $\boldsymbol{z}(G, \boldsymbol{u}^f, \alpha)$ is fibrewise constant. We conclude that the same must be true of the limit distribution:

Theorem 5.3. *Let G be stochastic, $f : G \to B$ be a colour-preserving fibration and $\boldsymbol{u}$ a non-negative vector over N_B such that $\boldsymbol{u}^f$ is a distribution over N_G. Then the limit distribution $\boldsymbol{r}(G, \boldsymbol{u}^f, \alpha)$ is fibrewise constant for every $\alpha \in [0\,..\,1)$.*

5.1. Applications to PageRank

The most immediate application of these results to PageRank is an easy consequence of Theorem 5.3: if a graph G contains two nodes that are in the same fibre of some colour-preserving fibration, then they have the same PageRank for all α, provided that the preference vector is fibrewise constant and G is endowed with its the natural random-walk colouring.

A more general result can indeed be obtained. Suppose you have two graphs G_1 and G_2 with n_1 and n_2 nodes respectively, colour them with their natural random-walk colouring, and suppose they are epimorphically fibred over the same graph B with two fibrations $f_1 : G_1 \to B$ and $f_2 : G_2 \to B$ that respect the colouring. For instance, if f_1 and f_2 are outdegree-preserving fibrations and B is not coloured, you can colour B so that f_1 and f_2 respect the colouring: this is possible, because if $f_1(a_1) = f_2(a_2)$ (for some $a_1 \in A_{G_1}$ and $a_2 \in A_{G_2}$) then the outdegree of $s(a_1)$ is the same as the outdegree of $s(a_2)$, so a_1 and a_2 must have the same colour.

Assume now that $i_1 \in N_{G_1}$ and $i_2 \in N_{G_2}$ are two nodes that are identified by the two fibrations (*i.e.*, $f_1(i_1) = f_2(i_2)$), and consider the vectors $\boldsymbol{u}_1 = \mathbf{1}/n_1$ and $\boldsymbol{u}_2 = \mathbf{1}/n_2$ over N_B. By Theorem 5.1 and Theorem 5.2,

$$\begin{aligned} r_{i_1}\big(G_1, \boldsymbol{u}_1^{f_1}, \alpha\big) &= z_{i_1}\big(G_1, \boldsymbol{u}_1^{f_1}, \alpha\big) = z_{f(i_1)}(B, \boldsymbol{u}_1, \alpha) \\ &= \frac{n_1}{n_2}\, z_{f(i_2)}(B, \boldsymbol{u}_2, \alpha) = \frac{n_1}{n_2}\, z_{i_2}\big(G_2, \boldsymbol{u}_2^{f_2}, \alpha\big) = \frac{n_1}{n_2}\, r_{i_2}\big(G_2, \boldsymbol{u}_2^{f_2}, \alpha\big). \end{aligned}$$

In particular, the PageRank values of i_1 and i_2, computed with uniform preference vector, coincide up to the multiplicative constant n_1/n_2.

6. Fibrations and stochastic graphs

The results of the previous section make it clear that it is interesting to build fibrations having a stochastic graph as total graph; however, if $f : G \to B$ is a colour-preserving fibration, and G is stochastic, B needs not be stochastic itself. Since we are interested in stochastic graphs that are fibred over a common base, we approach the problem of characterising $\mathbf{R}^+$-coloured graphs B over which stochastic graphs can be fibred. To this aim, initially we shall require the fibration to preserve colours only up to a multiplicative constant.

Formally, given an $\mathbf{R}^+$-coloured graph B, we want to establish necessary and sufficient conditions under which there exists a stochastic graph G and an epimorphic fibration $f : G \to B$ that preserves colours up to a multiplicative constant

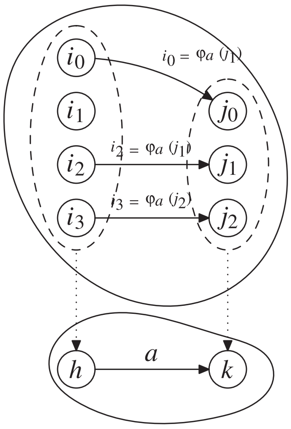

FIGURE 3. An example of the construction of G (upper part) from B (lower part); fibers F_h and F_k are enclosed in dotted ellipses ($n_h = 4$ and $n_k = 3$).

$\lambda > 0$, that is, $c(f(a)) = \lambda \cdot c(a)$ for all $a \in A_G$. The constraints we provide will turn out to be a special consequence of a more general property on the eigenvectors of the matrices G and B, that will be discussed in Section 6.1.

The first step towards this goal is a combinatorial description of all possible fibrations over B. There is a standard representation for fibrations [2, 14] that extends the results about the classical representation of coverings by voltage assignments [9]: an epimorphic fibration over B whose fibre over h has cardinality n_h is described by:

(1) a nonempty set F_h of cardinality n_h for each node h of B;
(2) a function $\varphi_a : F_k \to F_h$ for each arc $a \in B(h, k)$.

Essentially, for each node of B we fix a fibre F_h. Then, we know that we must add to G exactly $|F_k|$ copies of each arc of B ending in k, and each copy must end in a distinct element of F_k (as we need to lift uniquely that arc to each element of F_k). Our only freedom now is to decide which node will be the *source* of each copy, and the source is provided exactly by the function φ_a, which defines the source of the copy associated with each element of F_k. We assume for simplicity that the F_h's are pairwise disjoint.

Geometrically, we are stacking $|F_h|$ nodes of the graph G over h, as illustrated in the example of Figure 3. Then, for each arc a of B going from h to k and every node $j \in F_k$, we add an arc $\widetilde{a}^j$ in G, setting its target to j, and we freely choose its source $\varphi_a(j)$ in F_h.[8] Clearly, if we want to preserve colours up to multiplication by λ, a copy of arc a will have to be coloured by $c(a)/\lambda$.

[8] The data defining a fibration actually induce a presheaf on B^*, and this correspondence extends to an equivalence between the category of fibrations over B and the category of presheaves on B^*; see [2].

These data define a total graph G that has nodes $\bigcup_h F_h$, and arcs $\bigcup_k B(-,k) \times F_k$. An arc $\langle a, j\rangle$ goes from $\varphi_a(j)$ to j. Finally, we define the fibration $f : G \to B$ that maps every node $i \in F_h$ to h, and every arc $\langle a, j\rangle$ to a.

If we additionally impose that G is stochastic, we must require that for all nodes of G, the sum of the colours of the outgoing arcs is exactly 1, that is, for all nodes h of B and all $i \in F_h$ we require

$$\sum_{k\in N_B} \sum_{a\in B(h,k)} \sum_{j\in F_k} c(a)\,[\varphi_a(j) = i]/\lambda = 1, \tag{4}$$

where we used Iverson's notation (a predicate between brackets has value 1 if true, 0 if false).

The condition we have given is not easily manageable. However, we can derive a much more interesting necessary condition. Let us sum over $i \in F_h$:

$$\sum_{i\in F_h} \sum_{k\in N_B} \sum_{a\in B(h,k)} \sum_{j\in F_k} c(a)\,[\varphi_a(j) = i] = \lambda n_h.$$

Rearranging the summation order, we get to

$$\sum_{k\in N_B} \sum_{a\in B(h,k)} c(a) \sum_{i\in F_h} \sum_{j\in F_k} [\varphi_a(j) = i] = \lambda n_h.$$

Now, the double internal summation is trivially n_k, and once we move it outside, the summation over a just gives B_{hk}. Thus, we arrive at

$$\sum_{k\in N_B} B_{hk} n_k = \lambda n_h, \tag{5}$$

which is an eigenvalue problem of the form $B\boldsymbol{n} = \lambda\boldsymbol{n}$, where $\boldsymbol{n}$ is a vector of fibre cardinalities. In other words, we have proved the following

Theorem 6.1. *Given an $\mathbf{R}^+$-coloured graph B, if there exist a stochastic graph G and an epimorphic fibration $f : G \to B$ that preserves colours up to a multiplicative constant $\lambda > 0$, then B has a positive integer eigenvector associated with λ whose h-th component is the cardinality of the fibre over h (i.e., $|f^{-1}(h)|$).*

Observe that the necessary condition of the previous theorem is satisfied, for instance, when B is irreducible and has rational entries. On the other hand, the condition is not sufficient: the existence of a positive integer eigenvector of B does not guarantee the existence of a fibration from a stochastic graph, as equations (4) on sets may not be satisfiable. Consider for example the graph B in Figure 4. The vector $\boldsymbol{n} = \langle 3, 1\rangle^T$ is a solution of equation (5) for $\lambda = 1$. However, there is no way to define the functions φ_a so to satisfy (4).

Finally, we note that B cannot have two positive eigenvectors associated with distinct λ's, since the following proposition holds.

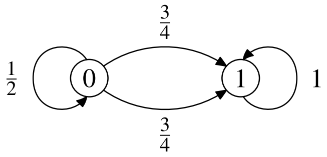

FIGURE 4. A graph over which no stochastic graph can be fibred.

Proposition 6.2. *Let A be a non-negative matrix, and assume $A\boldsymbol{x} = \lambda\boldsymbol{x}$ for some $\boldsymbol{x} > 0$. Then λ equals the spectral radius $\rho(A)$ of A.*

Proof. Note that necessarily λ is real and non-negative. Let $\boldsymbol{v}$ be a non-negative eigenvector associated with the spectral radius $\rho = \rho(A)$. Then there is an $\varepsilon > 0$ such that $\boldsymbol{x} - \varepsilon\boldsymbol{v} > 0$, so $A^n(\boldsymbol{x} - \varepsilon\boldsymbol{v}) \geq 0$. This entails

$$\lambda^n\|\boldsymbol{x}\| = \|A^n\boldsymbol{x}\| = \|A^n(\boldsymbol{x} - \varepsilon\boldsymbol{v}) + \varepsilon A^n\boldsymbol{v}\| \geq \|\varepsilon A^n\boldsymbol{v}\| = \varepsilon\rho^n\|\boldsymbol{v}\|,$$

so $\rho^n = O(\lambda^n)$, and since $\lambda \leq \rho$ we have $\lambda = \rho$. □

So, in particular, if $f : G \to B$ and $g : H \to B$ are two epimorphic fibrations that respect colours up to factors λ and μ, respectively, then necessarily $\lambda = \mu$. This fact allows one to rescale the colouring of B in a unique way (dividing by $\lambda = \mu$): this observation explains why we consider only fibrations that do respect colours.

6.1. A DEEPER LOOK

Actually, the computation we carried over has a much more general meaning when we look at it the other way around: if $\boldsymbol{w}$ is a right eigenvector of G associated with the eigenvalue λ, the equation $G\boldsymbol{w} = \lambda\boldsymbol{w}$ can be rewritten as the system of equations

$$\sum_{j \in N_G} \sum_{a \in G(i,j)} c(a)w_j = \lambda w_i,$$

where $i \in N_G$. Summing over $i \in f^{-1}(h)$ for any node h of B we obtain (after a rearrangement)

$$\sum_{j \in N_G} \sum_{i \in f^{-1}(h)} \sum_{a \in G(i,j)} c(a)w_j = \lambda \sum_{i \in f^{-1}(h)} w_i.$$

The two internal summations actually correspond (because of the lifting property) to a summation over arcs of B:

$$\sum_{j \in N_G} \sum_{a \in B(h,f(j))} c(a)w_j = \lambda \sum_{i \in f^{-1}(h)} w_i.$$

If we now break the summation over j as a double summation over the nodes of B and over their fibres, we obtain

$$\sum_{k\in N_B}\sum_{j\in f^{-1}(k)}\sum_{a\in B(h,f(j))} c(a)w_j = \lambda \sum_{i\in f^{-1}(h)} w_i,$$

and this finally leads us to

$$\sum_{k\in N_B}\sum_{a\in B(h,k)} c(a) \sum_{j\in f^{-1}(k)} w_j = \sum_{k\in N_B} B_{hk} \sum_{j\in f^{-1}(k)} w_j = \lambda \sum_{i\in f^{-1}(h)} w_i.$$

The last equation exactly states that the vector $\boldsymbol{u}$ over N_B defined by $u_h = \sum_{i\in f^{-1}(h)} w_i$ is an eigenvector of B associated with the eigenvalue λ, provided that it is nonzero. In other words,

Theorem 6.3. *Given an* $\mathbf{R}^+$*-coloured graph* G *and a colour-preserving fibration* $f : G \to B$*, if* $\boldsymbol{w}$ *is a right eigenvector of* G *associated with the eigenvalue* λ*, then the vector* $\boldsymbol{u}$ *defined by*

$$u_h = \sum_{i\in f^{-1}(h)} w_i,$$

is a right eigenvector of B *for* λ*, provided that* $\boldsymbol{u} \neq 0$*.*

The previous theorem is a dual counterpart of the classic result about lifting of eigenvectors used in spectral graph theory [28], which states that a left eigenvector $\boldsymbol{u}$ of B associated with the eigenvalue λ can be lifted to an eigenvector for the same eigenvalue by copying its coordinates fibrewise. Now, Theorem 6.1 can be obtained as a special consequence of Theorem 6.3, noting that being stochastic is equivalent to having $\mathbf{1}$ as a right eigenvector associated with the eigenvalue 1.

6.2. Computing over the base

Theorem 5.2 gives a precise relation between the formal series $\boldsymbol{z}(G, \boldsymbol{u}^f, \alpha)$ and $\boldsymbol{z}(B, \boldsymbol{u}, \alpha)$ whenever $f : G \to B$ is a colour-preserving fibration. If G is stochastic and $\boldsymbol{u}^f$ is a distribution over N_G, then $\boldsymbol{z}(G, \boldsymbol{u}^f, \alpha)$ is indeed the limit distribution of any Markov chain with restart having transition matrix $\mathscr{R}(G, \boldsymbol{u}^f, \alpha)$; hence, if we are interested in computing such a distribution, we can actually perform the computation over B, which might be much smaller. Here we must be careful, however: B is not itself stochastic, and $\boldsymbol{u}$ is not a distribution, so $\boldsymbol{z}(B, \boldsymbol{u}, \alpha)$ does not admit a stochastic interpretation. In particular, algorithms that are commonly used to compute $\boldsymbol{r}(-, -, -)$, like [7,11,17,18,21], cannot in general be applied to this case. However, by Theorem 6.1, the matrix B admits a positive eigenvector associated with the eigenvalue 1 and this allows us to proceed as follows.

It is possible to transform any $\mathbf{R}^+$-coloured graph B to obtain a stochastic graph B' (having the same underlying graph as B), whenever the matrix associated with B admits a positive eigenvector $\boldsymbol{w}$. It is sufficient to define the colouring function

c' of B' by setting for each arc a

$$c'(a) = \frac{1}{\rho}\frac{w_{t(a)}}{w_{s(a)}}c(a),$$

where ρ is the spectral radius of B and c is its colouring function. The resulting graph is indeed stochastic:

Proposition 6.4. *If B is an $\mathbf{R}^+$-coloured graph whose matrix admits an eigenvector $\boldsymbol{w} > 0$, then the graph B' is stochastic.*

Proof. By Proposition 6.2, the eigenvalue associated with $\boldsymbol{w}$ is the spectral radius ρ of the matrix B. Moreover, it is easy to show that

$$B' = \frac{1}{\rho}\mathrm{Diag}(\boldsymbol{w})^{-1}\, B\, \mathrm{Diag}(\boldsymbol{w})$$

where B' denotes, as usual, the matrix associated with the graph B' and, as it is immediate to see, is stochastic. □

The following theorem illustrates how the formal series $\boldsymbol{z}(-,-,-)$ changes when transforming a graph B (whose matrix has a positive eigenvector) into the stochastic graph B'.

Theorem 6.5. *If B is an $\mathbf{R}^+$-coloured graph such that its matrix admits an eigenvector $\boldsymbol{w} > 0$, then, for every non-negative vector $\boldsymbol{u}$ over N_B,*

$$\boldsymbol{z}(B, \boldsymbol{u}, \alpha) = \frac{1-\alpha}{1-\alpha\rho}\mathrm{Diag}(\boldsymbol{w})^{-1}\ \boldsymbol{z}(B', \mathrm{Diag}(\boldsymbol{w})\boldsymbol{u}, \alpha\rho),$$

where ρ is the spectral radius of B.

Proof. First of all, notice that for every path π

$$c_{B'}(\pi) = \frac{1}{\rho^{|\pi|}}\frac{w_{t(\pi)}}{w_{s(\pi)}}\, c_B(\pi).$$

Thus, by equation (3), we get

$$\begin{aligned}\frac{z_h(B', \boldsymbol{u}, \alpha)}{(1-\alpha)} &= \sum_{\pi\in B^*(-,h)} \alpha^{|\pi|} u_{s(\pi)}\, c_{B'}(\pi)\\ &= w_h \sum_{\pi\in B^*(-,h)} \left(\frac{\alpha}{\rho}\right)^{|\pi|} \frac{u_{s(\pi)}}{w_{s(\pi)}}\, c_B(\pi) = w_h\frac{z_h(B, \mathrm{Diag}(\boldsymbol{w})^{-1}\boldsymbol{u}, \alpha/\rho)}{1-\alpha/\rho},\end{aligned}$$

hence the result. □

As a consequence of the previous observations, we obtain the following

Theorem 6.6. *Let G be a stochastic graph, $f : G \to B$ be a colour-preserving epimorphic fibration, and $\boldsymbol{v}$ be a fibrewise constant distribution over N_G. Let $\boldsymbol{w} = (w_k)_{k \in N_B}$ be the vector whose k-th entry is $w_k = |f^{-1}(k)|$. Then, for every node i of G,*

$$r_i(G, \boldsymbol{v}, \alpha) = \frac{1}{w_{f(i)}} r_{f(i)}(B', \mathrm{Diag}(\boldsymbol{w})\boldsymbol{u}, \alpha),$$

where $\boldsymbol{u}$ is such that $\boldsymbol{v} = \boldsymbol{u}^f$.

Proof. First note that B' is a stochastic graph (since f being epimorphic implies $\boldsymbol{w} > 0$) and $\mathrm{Diag}(\boldsymbol{w})\boldsymbol{u}$ is a distribution over N_B, hence the right-hand side is well-defined. Applying Theorem 5.1 and Theorem 5.2, we have

$$r_i(G, \boldsymbol{u}^f, \alpha) = z_i(G, \boldsymbol{u}^f, \alpha) = z_{f(i)}(B, \boldsymbol{u}, \alpha),$$

whence the result is obtained by applying Theorem 6.5 (with $\rho = 1$) and again Theorem 5.1 to B'. □

Given a stochastic graph G and a distribution $\boldsymbol{v}$ over N_G, the previous theorem suggests that we should try to compute a fibration $f : G \to B$ such that $\boldsymbol{v}$ is fibrewise constant, so to compute the limit distribution on a smaller graph. Indeed, there is a *minimum* B with such a property, as the theorems given in Section 2.1 extend immediately [2] to graphs coloured on the nodes (where colours of the nodes represent the preference vector $\boldsymbol{v}$ of the Markov chain with restart) and fibrations preserving all colours. We shall see in Section 7 that B can actually be computed quite efficiently.

6.3. A worked-out example

Consider the graph G shown in Figure 5 (left), endowed with its natural random-walk colouring. The graph is fibred over B, shown in Figure 5 (centre), *via* the colour-preserving fibration $f : G \to B$ defined on the nodes by $f^{-1}(0) = \{0, 1\}$, $f^{-1}(1) = \{2, 3\}$ and $f^{-1}(2) = \{4, 5, 6, 7\}$ (any definition on the arcs is fine). Now suppose you want to compute the PageRank corresponding to the preference vector

$$\boldsymbol{v} = \langle 1/20, 1/20, 1/20, 1/20, 1/5, 1/5, 1/5, 1/5, 1/5 \rangle^T.$$

Since this vector is fibrewise constant, we can apply Theorem 6.6, with $\boldsymbol{u} = \langle 1/20, 1/20, 1/5 \rangle^T$, $\boldsymbol{w} = \langle 2, 2, 4 \rangle^T$ and B' obtained by recolouring B, as shown in Figure 5 (right).

In other words, we apply the standard PageRank computation to the graph B' using the preference vector $\mathrm{Diag}(\boldsymbol{w})\boldsymbol{u} = \langle 1/10, 1/10, 4/5 \rangle$; a direct computation gives

$$\boldsymbol{r}(B', \mathrm{Diag}(\boldsymbol{w})\boldsymbol{u}, \alpha) = \frac{\langle \alpha^2 + 8\alpha + 1, 8\alpha^2 + \alpha + 1, \alpha^2 + \alpha + 8 \rangle}{10(\alpha^2 + \alpha + 1)}.$$

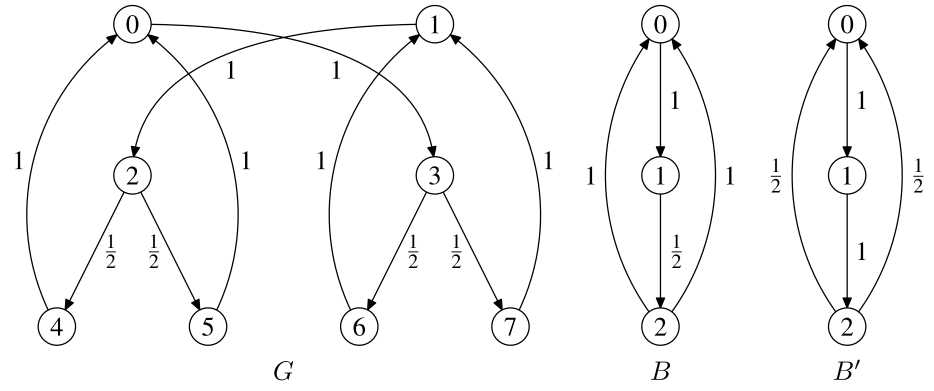

Figure 5. Two graphs G and B (with a fibration $f : G \to B$), and the corresponding stochastic graph B'.

Hence, for example, applying Theorem 6.6,

$$r_2(G, \boldsymbol{v}, \alpha) = \frac{1}{2} r_1(B', \mathrm{Diag}(\boldsymbol{w})\boldsymbol{u}, \alpha) = \frac{8\alpha^2 + \alpha + 1}{20(\alpha^2 + \alpha + 1)}.$$

7. Computation of the minimum base and graph isomorphism

As observed at the end of Section 6.2, Theorem 6.6 suggests that, given a Markov chain with restart specified by a stochastic graph G and a distribution $\boldsymbol{v}$, we should try to obtain a minimal fibration $f : G \to B$ such that $\boldsymbol{v}$ is fibrewise constant, so to compute the limit distribution on a smaller graph. Formally, we are considering the following problem

Problem: Minimum base labelling
Input: A graph G and a preference vector $\boldsymbol{v}$ over N_G
Output: A labelling $\ell : N_G \to \{0, 1, \dots, k-1\}$ such that, for any two nodes $i, j \in N_G$, $\ell(i) = \ell(j)$ iff i and j are in the same fibre of any minimal fibration for which $\boldsymbol{v}$ is fibrewise constant.

An algorithm for Minimum base labelling produces *canonical labellings* iff its output is automorphism invariant.

This problem is actually equivalent to the well-known *partition refinement* problem, assuming that the initial partition is the one induced by the preference vector $\boldsymbol{v}$. Starting from ideas appearing in Hopcroft's minimal-automaton construction [15], Cardon and Crochemore [4] devised a partition-refinement algorithm that in our terminology computes the minimum base labelling of a directed graph. Their algorithm does not *per se* provide canonical labellings, but it can be easily adapted to do so, and works in time $O((n+m)\log n)$, for an uncoloured graph with n nodes and m arcs[9]).

Indeed, this idea is not new, and has been used earlier [5, 23] by people working on graph isomorphism. This comes as no surprise, since graph isomorphism between fibration-prime graphs can be solved in polynomial time. As a matter of fact, by Proposition 2.3, if we establish a canonical order on labelled trees, we can canonically sort the nodes of a fibration prime graph; once the nodes are sorted, the isomorphism problem for fibration-prime graphs can be solved with a linear check.

For general graphs, in one of the seminal papers on the subject [5], Corneil and Gottlieb essentially propose to build the minimum base first, and then to reason on the fibres separately. If the minimum bases are not isomorphic, of course, isomorphism is impossible. Brendan McKay pushed this idea much further writing one of the fastest graph-isomorphism solvers, `nauty` [23], which is also able to compute generators for the automorphism group. `nauty` starts by building the *coarsest equitable partition* of the nodes, which in our language is the minimum base for the symmetric representation of the graph. If the fibres are trivial, the algorithm can canonically sort the vertices and perform a check. Otherwise, the algorithm starts a backtracking procedure, trying to break fibres by choosing selected elements. The time required for the construction of the minimum base is $O(n^2 \log n)$ for a graph with n nodes.

The connection between graph isomorphism and minimum bases has recently resurfaced, albeit unnoticed, in a paper by Gori, Sarti and Maggini [8]. They propose a polynomial isomorphism algorithm for a class of graphs defined in terms of PageRank: a graph with n nodes is said to be *Markovian spectrally distinguishable* if there are values α_0, α_1, …, α_{n-1} of the damping factor whose associated PageRank vectors form an invertible matrix. This class is in fact subsumed by fibration-prime graphs, since

Theorem 7.1. *A Markovian spectrally distinguishable graph is fibration prime.*

Proof. A nonprime graph is nontrivially fibred over its minimum base, and by Theorem 5.3 it contains at least two nodes whose PageRank values are the same for all values of the damping factor. As a consequence, it is impossible to build an invertible matrix using a set of PageRank vectors (at least two columns will always be equal). □

The converse of the previous theorem does not hold. The graph shown in Figure 6 is fibration prime (check the universal total graphs at depth three), and nonetheless the PageRank vector is

$$\left\langle \frac{\alpha+1}{2(2+\alpha)}, \frac{1}{2(2+\alpha)}, \frac{1}{4}, \frac{1}{4} \right\rangle^T ,$$

[9] We observe that in [4] the authors give an $O(m \log n)$ bound, but there seems to be a mistake in the computation of the bound, which is more correctly $O((n+m)\log n)$.

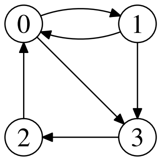

FIGURE 6. A fibration-prime graph that is not Markovian spectrally distinguishable. Nodes 2 and 3 have the same PageRank, independently of α.

so node 2 and 3 have the same PageRank (independently of α). We conclude that the class of Markovian spectrally distinguishable graphs is strictly smaller than the class of fibration-prime graphs.

All in all, we conclude that the isomorphism of fibration-prime graphs (and *a fortiori* of Markovian spectrally distinguishable graphs) is decidable in time $O((n+m)\log n)$ using purely discrete means. The space used by the above algorithms is $O(m+n)$. The PageRank-based algorithm proposed by Gori *et al.* comes with no detailed complexity analysis (the authors just notice that the overall algorithm must run in polynomial time), whereas we can obtain an almost linear upper bound.

The algorithms considered so far actually deal with uncoloured graphs only. The original paper about `nauty` does not discuss coloured graphs, whereas Cardon and Crochemore's does; however, the authors assume to be able to enumerate in linear time all arcs *with a given colour*. This is in contrast with a more realistic model in which arcs are enumerated in linear time, and then a constant-time colouring function provides the colour for each arc. Cardon and Crochemore's algorithm can be easily patched to work with the latter model, but in this case the time bound becomes $O((n+m\log m)\log n)$. The algorithm used by `nauty` can be adapted similarly.

7.1. SPACE $O(n)$

When dealing with very large web graphs, maybe using a compressed representation (see, *e.g.*, [3]), it is not always possible to use space $O(m)$. In this section we discuss how to implement a minimum base algorithm in additional space $O(n)$ (besides the space required to store the graph), paying of course a price in terms of computation time.

Let us start with a simple informal description of the algorithm. Throughout the algorithm k is the number of labels, and $\ell : N \to \{\,0,1,\dots,k-1\,\}$ is a *surjective* labelling of nodes: at the end of the algorithm, two nodes will have the same label iff they have the same universal total graph or, equivalently, if they are in the same fibre of any minimal fibration.

Let G be a C-coloured graph with colouring function $c : A \to C$, and assume a linear order on the colours. The algorithm performs a *refinement step* until no more refinement is possible:

(1) Set $k = 1$ and ℓ to the unique function $N \to \{0\}$.
(2) For each node i, if $G(-,i) = \{a_0, a_1, \dots, a_{d^-(i)-1}\}$, then let $\mathfrak{m}(i)$ be the multiset

$$\mathfrak{m}(i) = \left\{ \langle c(a_0), \ell(s(a_0)) \rangle, \langle c(a_1), \ell(s(a_1)) \rangle, \dots, \langle c(a_{d^-(i)-1}), \ell(s(a_{d^-(i)-1})) \rangle \right\} ;$$

update ℓ so that two nodes i and j have the same label iff $\mathfrak{m}(i) = \mathfrak{m}(j)$.
(3) If $k = n$ or the codomain of ℓ has not changed, stop; otherwise, set k to the cardinality of the codomain of ℓ and restart from (2).

We assume a standard model in which it is possible to iterate over the list of incoming arcs in linear time. Thus, the crux of the algorithm is the update of the labelling function ℓ. Since we have a problem of uniqueness, a possible approach is enumerating the multisets $\mathfrak{m}(i)$ in sorted order: equal elements will be enumerated consecutively, making it trivial to update ℓ.

To accomplish the task, we must define an easily computable order on multisets of pairs in $C \times \mathbf{N}$. The pairs themselves can be easily ordered lexicographically, since C is ordered, and by choosing for each multiset a *sorted* canonical representative we will be able to compare multisets in a lexicographic fashion.

The first point to examine, thus, is the time required to compute the canonical representative for all multisets. Note that we are sorting, for each node x, a list of $d^-(i)$ elements, which can be easily done in time $O(d^-(i) \log d^-(i))$. This gives an overall bound of

$$\sum_{i \in N} O(d^-(i) \log d^-(i)) = O(n + m \log m)$$

for the construction of canonical representatives.

We now sort the canonical representatives using a merge sort, carefully counting the number of times a canonical representative is used: after each comparison one element is emitted, and each element is emitted exactly once. We conclude that no more than m list elements are ever compared, so the most important cost in a merging phase is the time required to build all canonical forms (since we want to obtain space $O(n)$, we cannot build all representatives and reuse them for all phases): multiplying by $\log n$ (the number of merging phases) and recalling Theorem 2.2 we get the following

Theorem 7.2. *A mergesort-based algorithm for the construction of the minimum base with on-demand canonical representative construction uses time $O(q(n+m \log m) \log n)$ and space $O(n + \Delta^-)$, where Δ^- is the maximum indegree of G and q is the number of refinement steps.*

The presence of Δ^- is due to the fact that at some point the list for the node with largest indegree will have to be built. If the graph is separated, $\Delta^- \leq n$, and

Table 1. Experimental results about the dimension of the minimum base for some real-world Web snapshots.

Dataset	Number of nodes	Number of fibres	Average fibre size
WebBase	118,142,155	41,705,767	2.83
`.it`	41,291,594	15,245,587	2.71
`.uk`	39,459,925	14,154,663	2.79

the bound reduces to $O(n)$. A similar bound can be obtained for quicksort if the implementation uses actual medians as pivots. The number of refinement steps q in the worst case is $O(n)$, but experimentation with actual web graphs shows that it is actually much smaller, making the algorithm feasible even for very large graphs.

We remark that the labels assigned by this algorithm are canonical, as they correspond to the lexicographic order of the universal total graphs.

8. Experimental results and conclusions

The discussion of the previous section highlights a rather interesting fact: there is an entirely discrete algorithm (the construction of the minimum base) whose output, by Theorem 5.2, imposes constraints on the values of the limit distribution of a Markov chain. Thus, a limit process is constrained by a discrete process computable in polynomial time. Of course, the condition provided by Theorem 5.2 is sufficient only, but nonetheless it is fascinating that the discrete structure of the underlying graph can impose such a significant constraint on the limit distribution. Moreover, Theorem 6.6 can be used, at least in principle, to reduce the efforts required to compute the limit distribution by performing the actual computation on the base, which may be much smaller.

One may wonder whether this idea can be fruitfully applied for the computation of PageRank of real-world web graphs; actually, it might be the case that such graphs are themselves fibration-prime, which would make Theorem 6.6 useless in practice. On the contrary, some preliminary experiments performed on real datasets show that real web graphs exhibit a minimum base that is about 3 times smaller than the corresponding graph (see Tab. 1). Moreover, experiments show that the time required by our algorithm to compute the minimum base makes the use of Theorem 6.6 a viable option. As an aside, we observe that fibre sizes roughly follow a power-law distribution, as shown in Figure 7; we don't have any theoretical explanation of this fact, which certainly needs further investigation.

There is more, however: if we interrupt the minimum base construction algorithm at step k (in the $O(n)$ version) we obtain a partitioning of the nodes into classes sharing the first k levels of their universal total graph. Thus, in the case of a Markov chain with restart, the difference of the limit distribution for two nodes in the same class is bounded by α^k. Once again, we have a purely combinatorial computation that imposes constraints on the values of the limit distribution.

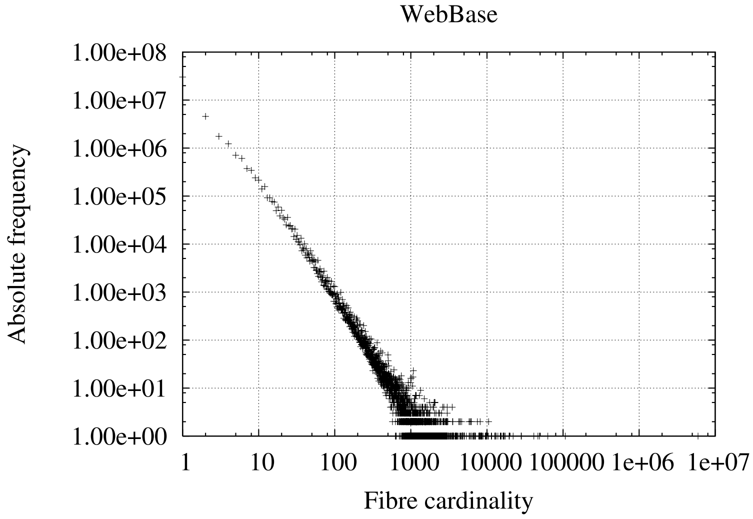

FIGURE 7. The distribution of fibre cardinalities.

Acknowledgements. We would like to acknowledge Dániel Fogaras who provided us a first proof of the limit in (2), for the special case of uniform $\boldsymbol{v}$ and primitive P, thus being the first confirming the conjecture stated in [1].

REFERENCES

[1] P. Boldi, M. Santini and S. Vigna, PageRank as a function of the damping factor, in *Proc. of the Fourteenth International World Wide Web Conference.* ACM Press. Chiba, Japan (2005) 557–566.

[2] P. Boldi and S. Vigna, Fibrations of graphs. *Discrete Math.* **243** (2002) 21–66.

[3] P. Boldi and S. Vigna, The WebGraph framework I: Compression techniques, in *Proc. of the Thirteenth International World Wide Web Conference.* ACM Press, Manhattan, USA (2004) 595–601.

[4] A. Cardon and M. Crochemore, Partitioning a graph in $O(|A| \log_2 |V|)$. *Theoret. Comput. Sci.* **19** (1982) 85–98.

[5] D.G. Corneil and C.C. Gotlieb, An efficient algorithm for graph isomorphism. *J. Assoc. Comput. Mach.* **17** (1970) 51–64.

[6] L. Eldén, The eigenvalues of the google matrix. Technical Report LiTH-MAT-R-04-01, Department of Mathematics, Linköping University, 2004. Available at `http://arxiv.org/abs/math.RA/0401177`.

[7] G.H. Golub and C. Greif, Arnoldi-type algorithms for computing stationary distribution vectors, with application to PageRank, *Technical Report SCCM-04-15*, Stanford University Technical Report (2004).

[8] M. Gori, M. Maggini and L. Sarti, Exact and approximate graph matching using random walks. *IEEETPAMI: IEEE Trans. Pattern Anal. Machine Intelligence* **27** (2005) 1100–1111.

[9] J.L. Gross and T.W. Tucker, *Topological Graph Theory.* Series in Discrete Mathematics and Optimization, Wiley-Interscience (1987).

[10] A. Grothendieck, Technique de descente et théorèmes d'existence en géométrie algébrique, I. Généralités. Descente par morphismes fidèlement plats. *Seminaire Bourbaki* **190** (1959–1960).

[11] T.H. Haveliwala, Efficient computation of PageRank. *Technical Report 31*, Stanford University Technical Report, October 1999. Available at `http://dbpubs.stanford.edu/pub/1999-31`.

[12] T.H. Haveliwala, Topic-sensitive pagerank, in *The eleventh International Conference on World Wide Web Conference.* ACM Press (2002) 517–526.

[13] T.H. Haveliwala and S.D. Kamvar, The second eigenvalue of the Google matrix. *Technical Report 20*, Stanford University Technical Report, March 2003. Available at `http://dbpubs.stanford.edu/pub/2003-20`.

[14] P. Híc, R. Nedela and S. Pavlíková, Front-divisors of trees. *Acta Math. Univ. Comenian. (N.S.)* **61** (1992) 69–84.

[15] J.E. Hopcroft, An $n \log n$ algorithm for minimizing states in a finite automaton, in *Theory of Machines and Computations*, edited by Z. Kohavi and A. Paz. Academic Press (1971).

[16] M. Iosifescu, *Finite Markov Processes and Their Applications.* John Wiley & Sons (1980).

[17] S.D. Kamvar, T.H. Haveliwala, C.D. Manning and G.H. Golub, Exploiting the block structure of the web for computing PageRank. *Technical Report 17*, Stanford University Technical Report, March 2003. Available at `http://dbpubs.stanford.edu/pub/2003-17`.

[18] S.D. Kamvar, T.H. Haveliwala, C.D. Manning and G.H. Golub, Extrapolation methods for accelerating PageRank computations, in *Proceedings of the twelfth international conference on World Wide Web.* ACM Press (2003) 261–270.

[19] T. Kato, *Perturbation Theory for Linear Operators.* Springer-Verlag, second edition (1976).

[20] L. László, Random walks on graphs: A survey, in *Combinatorics, Paul Erdős is Eighty*, Vol. 2, Bolyai Society Mathematical Studies, 1993, in *Proceedings of the Meeting in honour of P. Erdős*, Keszthely, Hungary **7** (1993) 1–46.

[21] C. Pan-Chi Lee, G.H. Golub and S.A. Zenios, A fast two-stage algorithm for computing PageRank and its extensions. *Technical Report SCCM-03-15*, Stanford University Technical Report (2003). Available at `http://www-sccm.stanford.edu/pub/sccm/sccm03-15_2.pdf`.

[22] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding.* Cambridge University Press, Cambridge UK (1995).

[23] B.D. McKay, Practical graph isomorphism. *Congressus Numerantium* **30** (1981) 45–87.

[24] C.D. Meyer, *Matrix analysis and applied linear algebra.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2000).

[25] M. Nasu, Constant-to-one and onto global maps of homomorphisms between strongly connect graphs. *Ergod. Th. & Dynam. Sys.* **3** (1983) 387–413.

[26] N. Norris, Universal covers of graphs: Isomorphism to depth $n-1$ implies isomorphism to all depths. *Discrete Appl. Math.* **56** (1995) 61–74.

[27] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank citation ranking: Bringing order to the web. *Technical Report 66*, Stanford University, 1999. Available at `http://dbpubs.stanford.edu/pub/1999-66`.

[28] D.M. Cvetković, M. Doob and H. Sachs, *Spectra of Graphs.* Academic Press (1978).

[29] J.P. Schweitzer. Perturbation theory and finite markov chains. *J. Appl. Probab.* **5** (1968) 401–413.

[30] E. Seneta, *Non-negative matrices and Markov chains.* Springer–Verlag, New York (1981).

[31] S.H. Unger, GIT – A heuristic program for testing pairs of directed line graphs for isomorphism. *Comm. ACM* **7** (1964) 26–34.

[32] S. Vigna, A guided tour in the topos of graphs. *Technical Report 199-97*, Università di Milano, Dipartimento di Scienze dell'Informazione, 1997. Available at `http://vigna.dsi.unimi.it/ftp/papers/ToposGraphs.pdf`.

[33] K. Yosida, *Functional Analysis.* Springer-Verlag, (1980), Sixth Edition.