The PageRank Vector of a Scale-Free Web Network Growing by Preferential Attachment

Natalia M. Markovich[1(✉)] and Udo R. Krieger[2]

[1] V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Profsoyuznaya Street 65, 117997 Moscow, Russia
markovic@ipu.rssi.ru

[2] Fakultät WIAI, Otto-Friedrich-Universität, An der Weberei 5, 96047 Bamberg, Germany
udo.krieger@ieee.org

Abstract. We consider a scale-free model of the Web network that is evolving by preferential attachment schemes and derive an explicit formula of its PageRank vector. Its i^{th} element indicates the probability that a surfer resides at a related Web page i in a stationary regime of an associated random walk. Considering the growth of a directed Web graph, we apply linear preferential attachment schemes proposed by Samorodnitsky et al. (2016). To express the probability of a connection between two nodes of this Web graph, our derivation allows us to avoid the consideration of complicated paths with random lengths and to cover both self-loops and multiple edges between nodes. An algorithm of the PageRank vector calculation for graphs without loops is provided. The approach can be extended in a similar way to graphs with loops. In this way, our approach enhances existing analysis schemes. It provides a better insight on the PageRank of growing scale-free Web networks and supports the adaptation of the model to gathered network statistics.

Keywords: PageRank vector · Scale-free network · Linear preferential attachment

1 Introduction

Let $G = (V, E)$ be the directed graph of a scale-free Web network with $n = |V|$ vertices $v \in V$, i.e., Web pages, and a growing number $|E|$ of edges $e \in E$, i.e., hyperlinks among these pages. The PageRank is accepted as a popular basic measure to rank the Web pages that are provided by a search engine after a request to the Web. Google's PageRank vector $R = (R_1, ..., R_n)^T \in (0, \infty)^n$ [1] is the unique solution of the following system of linear equations

$$R_i = c \sum_{j:(j,i)\in E} \frac{R_j}{D_j} + (1-c)q_i, \qquad i \in \{1, \ldots, n\}.$$

The sum is taken over a number of pages j with incoming links to page i (in-degree), D_j is the number of outgoing links of page j (out-degree), and $c \in (0,1)$ is a damping factor which was originally set equal to 0.85 by Google. The probability vector $q = (q_1, q_2, ..., q_n)^T$ is a personalization vector of a user's preferences such that $q_i \geq 0$ and $\sum_{i=1}^{n} q_i = 1$ hold.

The in-degree is the simplest guidance to calculate the PageRank since the distributions of both entities are similar, [2]. Given the weak degree correlations in the Web graph, the approximation of the PageRank vector by the in-degrees of nodes can be relatively accurate [2]. A data analysis of Web graphs has revealed that the tails of both the PageRank and in-degree distributions follow power laws $1 - F(x) = c \cdot x^{-\alpha}$ with the same exponent α which is about 1.1, [3]. As a search engine can easily monitor the in-degree arising from the evolution of the Web and the in- and out-degree are available gathered statistics, the latter approximation is useful in practice.

Among other approaches to calculate the PageRank R_i of a randomly chosen Web page are the iteration procedure [4] and the mean field approach [5].

We focus on the PageRank formula proposed in [6, Lemma 3.1] and recalled here as Lemma 1. This formula is only valid for trees and directed acyclic Web graphs and the assumption that the out-degree m for each node is fixed. It is our objective to obtain an explicit formula of the PageRank vector without the constraints of [6]. Our derivation allows us to avoid the consideration of complicated paths with random lengths and to cover both self-loops and multiple edges between nodes. To this end, we use the probabilities of the edge creation by the $\alpha-$, $\beta-$ and $\gamma-$schemes of the linear preferential attachment (PA) proposed in [7] and [8] to determine the probability distribution of accessing the Web pages. The latter schemes are considered as basic growth model of the network. We propose a new formula for the PageRank of a node with a random out-degree m which is obtained after appending new nodes to the scale-free network.

The paper is organized as follows. In Sect. 2 the dynamics of an evolving Web graph is described and in Sect. 3 the calculation of its PageRank vector. The new computational formula for the PageRank vector of evolving Web networks is presented in Sect. 4. The exposition is finalized by some conclusions.

2 The Linear Preferential Attachment Schemes

Let $G(n) = (V(n), E(n))$ denote a directed graph with sets of nodes $V(n)$ and edges $E(n)$, n be the number of edges and $N(n)$ denote the number of nodes in $G(n)$. Let us recall the $\alpha-$, $\beta-$ and $\gamma-$PA-schemes given in [7,8].

A finite directed graph $G(n_0)$ is used as a seed network. It consists of at least one node v_0 and n_0 edges. A new node v is appended to the existing graph $G(n-1), n > n_0$, by adding a single edge to $G(n-1)$. The edge creation is provided by flipping a three-sided coin with probabilities α, β and γ. To this end, an i.i.d. sequence of trinomial r.v.s with values 1, 2 and 3 and the corresponding probabilities α, β and γ are generated to select schemes. $I_n(v)$ and $O_n(v)$ denote the in- and out-degree of v. A new node v is selected among nodes of the network and appended to $G(n-1)$.

The edge $v \to w \equiv (v, w) \in E(n)$ directed from the new node $v \in V(n)$ to an existing node $w \in V(n-1)$ is created with probability α. The existing node $w \in V(n-1)$ is chosen with probability

$$(P_\alpha)_{v,w} = P\{v \to w\} = \frac{I_{n-1}(w) + \delta_{in}}{n - 1 + \delta_{in} N(n-1)} \tag{1}$$

by the $\alpha-$scheme.

An edge (v, w) is added to $E(n-1)$ with probability β and the existing nodes $v \in V(n-1) = V(n), w \in V(n-1)$ are chosen independently from $G(n-1)$ with probability

$$(P_\beta)_{v,w} = P\{v \to w\} = \left(\frac{I_{n-1}(w) + \delta_{in}}{n - 1 + \delta_{in} N(n-1)}\right)\left(\frac{O_{n-1}(v) + \delta_{out}}{n - 1 + \delta_{out} N(n-1)}\right) \tag{2}$$

by the $\beta-$scheme.

An edge (w, v) from the existing node $w \in V(n-1)$ to v is created and the node w is chosen with probability

$$(P_\gamma)_{w,v} = P\{w \to v\} = \frac{O_{n-1}(w) + \delta_{out}}{n - 1 + \delta_{out} N(n-1)} \tag{3}$$

by the $\gamma-$scheme.

The parameters of the PA method δ_{in} and δ_{out} can be estimated by the semi-parametric extreme value method (EV) based on the maximum-likelihood method, [8]. It holds $\alpha + \beta + \gamma = 1$.

3 Calculating the PageRank Vector of a Web Network

There are numerous approaches to calculate the PageRank R_i of a randomly chosen page $v = i \in V$ in a Web graph $G = (V, E)$. One of them is determined by the following iteration

$$\widehat{R}_i^{(n,0)} = 1, \quad \widehat{R}_i^{(n,k)} = \sum_{j \to i} \frac{c}{D_j} \widehat{R}_j^{(n,k-1)} + (1 - c), \quad k \in \mathbb{N}, \tag{4}$$

proposed in [4] for a given uniform personalization vector $q_i = 1/n, 1 \le i \le n = |V|$. Then the scale-free PageRank of a node $v = i$ is denoted by $R_i^{(n)} = nR_i$. This iteration (4) is proceeding until the difference between two consecutive iterations $|\widehat{R}_i^{(n,k)} - \widehat{R}_i^{(n,k-1)}|$ will be small enough to reach approximately its limit $R_i^{(n)} = \lim_{k \to \infty} \widehat{R}_i^{(n,k)}$ which is sufficient for a moderate number of iterations k. Here, $j \to i$ implies that node j is linked to node i, i.e. $(j, i) \in E$.

Another important method is the mean field approach [5]. Its idea is to average the PageRanks of nodes that are aggregated within in- and out-degree classes (k_{in}, k_{out}). Such class contains nodes with the same in-degree k_{in} and the same out-degree k_{out}. We focus on the PageRank formula (5) proposed in [6, Lemma 3.1] and recalled further in Lemma 1.

In [6] the following notations are adopted for a Web graph $G = (V(n), E(n))$: $V(n) = \{0, \ldots, n\}$, $E(n) \subseteq V(n) \times V(n)$, $n \in \mathbb{N}$.

Let $\pi_v(n)$ be the PageRank of a node v after the n^{th} step of the network's evolution, $P_v(n)$ be the set of all paths from nodes $v+1, \ldots, n$ to v and $\ell(p)$ be the length of such a path p.

Lemma 1. *[6] The PageRank of node $v \in V(n)$, $v > 0$ within the realization of a growing network at time step $n > 0$ is given by*

$$\pi_v(n) = \frac{1-c}{n+1}\left(1 + \sum_{p \in P_v(n)} \left(\frac{c}{m}\right)^{\ell(p)}\right) \tag{5}$$

and the PageRank of the initial node $v = 0$ is given by

$$\pi_0(n) = \frac{1}{n+1}\left(1 + \sum_{p \in P_0(n)} \left(\frac{c}{m}\right)^{\ell(p)}\right).$$

Note that $m = 1$ implies a chain as special case of a tree. The proof of the Lemma is based on a formula of the PageRank vector in [9]

$$\pi(n) = \frac{1-c}{n+1}\underline{\mathbf{1}}^T[I - cP]^{-1} \tag{6}$$

with a scaling term $n+1$ due to a uniform personalization vector $q = 1/(n+1)\underline{\mathbf{1}} \in (0,1]^{n+1}$. Here, $\underline{\mathbf{1}}$ is the column vector of all ones.

$P \in [0,1]^{(n+1)\times(n+1)}$ is the square hypermatrix of a random walk defined by the transition probabilities of a Web surfer in the following equation

$$\widetilde{P} = c \cdot P + \frac{1-c}{n+1} \cdot E, \tag{7}$$

$E = \underline{\mathbf{1}} \cdot \underline{\mathbf{1}}^T$ is a matrix whose entries are all equal to one. The Web pages can be considered as states of a Markov chain. The $(i,j)^{th}$ element p_{ij} of P is the probability of a surfer moving from Web page i to j in one time step [9]. If there is no outgoing link from page i to j, then $p_{ij} = 0$. If a surfer follows a random walk corresponding to $\widetilde{P}$, then the i^{th} coordinate π_i of the PageRank vector π coincides with the probability in a stationary regime that the surfer stays at page i [6].

This formula (5) is only valid for trees and directed acyclic Web graphs. Other constraints of (5) are as follows. Firstly, the PageRank expression is addressed to a growing directed graph with a fixed value m of the out-degree for each node which is not plausible for real-world networks. The fixed m allows to start from any node (Web page) equally likely with a probability c/m and to follow any of the outgoing links to arrive to another node. Secondly, the PA is considered as a basic growth model of the network's evolution. A new node is appended to the network at each time step and adds m incoming links to existing nodes.

Indeed, not all pairs of nodes may be connected by a direct edge and the path length between them is random. In [6] the probability of a connection between two nodes is taken equal to $(c/m)^{\ell(p)}$. To this end, it is assumed that all per-hop links on the paths from node i to node j (i.e. $i \Rightarrow j$) are independent. This feature may be unrealistic since according to a linear PA each existing node j is chosen randomly from the current network state with a probability proportional to its degree k_j. This means that new nodes prefer to become attached to an existing node with a large node degree. The existence of superstar nodes to which a large proportion of nodes is attached may create a dependence between the links of paths between nodes. The length of the path $i \Rightarrow j$ is also determined by the direction of these link attachments.

4 The PageRank Vector of an Evolving Web Graph

We use (1)–(3) to define the elements of the hypermatrix P in (6). Let us start the PA from a single node and consider elementary graphs like trees and chains to derive the PageRank vector of an arbitrary graph obtained after the nth step of the network's evolution. Nodes are numerated in the order of appending them to the evolving graph. The probability to append a new edge leading from the newly appearing node i to the existing node j is

$$p_{ij} = P\{i \to j\} = \begin{cases} \alpha(P_\alpha)_{ij}, \, i > j, \\ \gamma(P_\gamma)_{ij}, \, i < j, \\ \beta(P_\beta)_{ij}, \, i \in V(n-1) = V(n). \end{cases} \tag{8}$$

For simplicity, graphs without loops are considered. The approach can be extended the same way for graphs with loops and multiple edges.

To obtain a PageRank vector we use (6) and the expansion of the inverse matrix

$$[I - cP]^{-1} = I + cP + c^2P^2 + ... \tag{9}$$

by a power series. Let us consider examples to explain our approach.

Example 1. Let the node 2 be appended to the initial node 1 by adding a single edge leading from 1 to 2, Fig. 1(a). The self-loops in the node 2 are impossible on Step 2 due to the application of the PA schemes (1)–(3). If $p_{12} = P\{1 \to 2\} \neq 0$ holds, then the hyperlink matrix at Step 1 is given by $P = \begin{pmatrix} 0 & p_{12} \\ 0 & 0 \end{pmatrix}$, and we get $P^2 = P^3 = ... = 0$. By (6) and (9) the PageRank vector at Step 1 is

$$\pi(1) = \frac{1-c}{2}((1,1) + (0, cp_{12})).$$

If $p_{21} = P\{2 \to 1\} \neq 0$ holds (see Fig. 1(b)), then

$$\pi(1) = \frac{1-c}{2}((1,1) + (cp_{21}, 0)).$$

If $p_{12} \neq 0$ and $p_{21} = \beta(P_\beta)_{21}$ hold (see Fig. 1(c)), that means a loop, then we get $P = \begin{pmatrix} 0 & p_{12} & 0 \\ p_{21} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, and by (6) and (9) it holds

$$\pi(2) = \frac{1-c}{3}(1 + \sum_{k=1}^{\infty} c^{2k-1} p_{12}^{k-1} p_{21}^{k}(cp_{12}+1), 1 + \sum_{k=1}^{\infty} c^{2k-1} p_{12}^{k} p_{21}^{k-1}(cp_{21}+1), 0)$$

at Step 2. Here, zero relates to the node 3 that is absent.

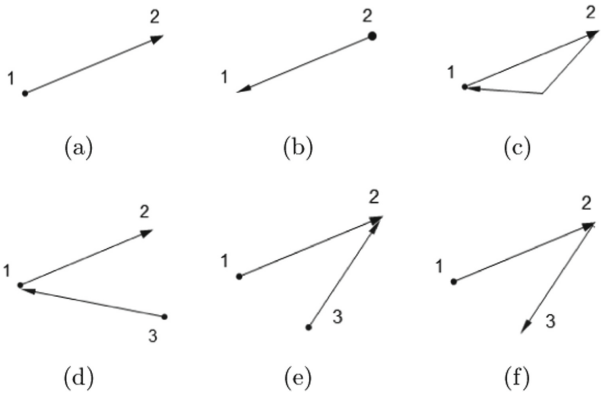

Fig. 1. Simplest graphs on Steps 1 and 2 of the graph evolution.

In the same way we have

$$\pi(2) = \frac{1-c}{3}((1,1,1) + c \cdot (p_{31}, p_{12} + cp_{31}p_{12}, 0)), \tag{10}$$

$$\pi(2) = \frac{1-c}{3}((1,1,1) + c \cdot (0, p_{12} + p_{32}, 0)), \tag{11}$$

$$\pi(2) = \frac{1-c}{3}((1,1,1) + c \cdot (0, p_{12}, p_{23} + cp_{12}p_{23})) \tag{12}$$

corresponding to graphs in Fig. 1(d)–(f). Note that root nodes which have outgoing edges only (see Fig. 1(a)–(f) except Fig. 1(c)) correspond to zeros in the brackets in formulae of $\pi(n)$. If the node i has only incoming edges, e.g. the node 2 in Fig. 1(e), then the probabilities $\{p_{ji}\}$ corresponding to these edges are summarized. The chains incoming to a node correspond to sums like in (10) and (12) reflecting the order of the edges in incoming chains.

Then the subsequent Lemma follows.

Lemma 2. *Let the graph be evolving by a linear PA model (1)–(3) and p_{ij} be defined by (8). Then the PageRank vector at time step n is given by*

$$\pi(n) = \frac{1-c}{n+1}(\underline{\mathbf{1}}^T + c \cdot (p_{1j_1}(1+cp_{i1}), p_{1j_2}(1+cp_{i2}), ..., \\ p_{1j_{i-1}}(1+cp_{i(i-1)}), 0, p_{1j_{i+1}}(1+cp_{i(i+1)}), ..., p_{N_i j_{N_i}}(1+cp_{iN_i}), ...))$$

for the Galton-Watson tree (Fig. 2(a)) with the root node i, and

$$\pi(n) = \frac{1-c}{n+1}\left(\underline{\mathbf{1}}^T + c\cdot(0, p_{12}, p_{23} + cp_{12}p_{23}, ..., p_{n(n+1)} + \sum_{t=1}^{n-1} c^t \prod_{j=n-t}^{n} p_{j(j+1)})\right)$$

for chains $(1 \to 2 \to \ldots \to n)$ beginning at node 1.

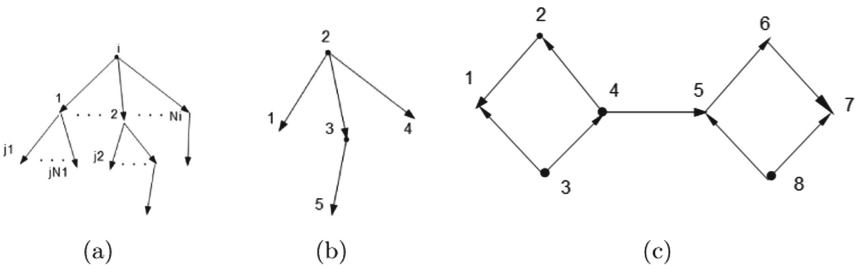

Fig. 2. Examples of Galton-Watson trees and a directed graph.

By analyzing incoming chains to each node one can easily obtain the PageRank vector of any graph without loops.

Algorithm 1. Calculation of the PageRank vector of graphs without loops

- Let a node 1 initiate a seed graph and n be the number of steps.
- Define a PageRank vector as $\pi(n) = \frac{1-c}{n+1}((1, ..., 1) + c\cdot(a_1, a_2, \ldots, a_n, a_{n+1}))$, where a_i corresponds to the node i.
- Find all incoming chains of the graph to a node i, $i \in \{1, 2, \ldots, n, n+1\}$, and determine $a_i = \sum_{j\to i} p_{ji} + \sum_{t=1}^{\ell_i - 1} c^t \sum_{j_t \to \ldots \to i} \underbrace{p_{j_t j_{t-1}} \cdot \ldots \cdot p_{ji}}_{t+1 \text{ multipliers}}$, where ℓ_i is a maximum length of the incoming chains to the node i.
- If the node i is a root, i.e. it has only outgoing edges, then determine $a_i = 0$.

Example 2. Let us consider the tree (Fig. 2(b)) and the directed graph (Fig. 2(c)). Their corresponding PageRank vectors are the following

$$\pi(4) = \frac{1-c}{5}((1,1,1,1,1) + c(p_{21}, 0, p_{23}, p_{24}, p_{35}(1 + cp_{23}))),$$

$$\pi(8) = \frac{1-c}{9}(\underline{\mathbf{1}}^T + c(p_{21} + p_{31} + cp_{21}p_{42}, p_{42} + cp_{42}p_{34}, 0, p_{34}, p_{45} + cp_{34}p_{45}, p_{56}$$

$$+ cp_{45}p_{56} + c^2 p_{34}p_{45}p_{56}, p_{87} + p_{67} + cp_{56}p_{67} + c^2 p_{45}p_{56}p_{67} + c^3 p_{34}p_{45}p_{56}p_{67}, 0)).$$

The PageRank vector of a graph with loops like in Fig. 1(c) may be obtained in a similar way, but this item is out of scope of the paper.

5 Conclusions

Regarding a free-scale directed Web network without loops evolving by the $\alpha-$, $\beta-$ and $\gamma-$linear PA-schemes given in [7,8], the associated PageRank vector at time step n is computed. The PageRank vector is derived avoiding constraints of [6, Lemma 3.1], particularly, a fixed out-degree of all nodes. Our result allows us to avoid the consideration of random length paths to express the probability of a connection between two nodes. The algorithm to calculate the PageRank vector of directed graphs without loops is provided.

Since the $\alpha-$, $\beta-$ and $\gamma-$linear PA-schemes allow us to generate graphs with loops and multiple edges, PageRank vectors of graphs with loops and multiple edges are a subject of our further research.

Acknowledgments. The first author was partly supported by Russian Foundation for Basic Research (grant 19-01-00090).

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
2. Fortunato, S., Boguñá, M., Flammini, A., Menczer, F.: Approximating PageRank from in-degree. In: Aiello, W., Broder, A., Janssen, J., Milios, E. (eds.) WAW 2006. LNCS, vol. 4936, pp. 59–71. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78808-9_6
3. Litvak, N., Scheinhardt, W.R.W., Volkovich, Y.: In-degree and PageRank: why do they follow similar power laws? Internet Math. **4**(2–3), 175–198 (2007)
4. Chen, N., Litvak, N., Olvera-Cravioto, M.: PageRank in scale-free random graphs. In: Bonato, A., Graham, F.C., Prałat, P. (eds.) WAW 2014. LNCS, vol. 8882, pp. 120–131. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13123-8_10
5. Fortunato, S., Boguñá, M., Flammini, A., Menczer, F.: On local estimations of PageRank: a mean field approach. Internet Math. **4**(2–3), 245–266 (2007)
6. Avrachenkov, K., Lebedev, D.: PageRank of scale-free growing networks. Internet Math. **3**(2), 207–231 (2006)
7. Samorodnitsky, G., Resnick, S., Towsley, D., Davis, R., Willis, A., Wan, P.: Non-standard regular variation of in-degree and out-degree in the preferential attachment model. J. Appl. Prob. **53**(1), 146–161 (2016)
8. Wan, P., Wang, T., Davis, R.A., et al.: Are extreme value estimation methods useful for network data? Extremes **23**, 171–195 (2020). https://doi.org/10.1007/s10687-019-00359-x
9. Langville, A.N., Meyer, C.D.: Deeper inside PageRank. Internet Math. **1**(3), 335–380 (2005)