Asymptotic Analysis for Personalized Web Search

Yana Volkovich and Nelly Litvak

Department of Applied Mathematics, University of Twente
PO Box 217, 7500 AE Enschede, The Netherlands
e-mail:{y.volkovich, n.litvak}@ewi.utwente.nl

Abstract

Personalized PageRank is used in Web search as an importance measure for Web documents. The goal of this paper is to characterize the tail behavior of the PageRank distribution in the Web and other complex networks characterized by power laws. To this end, we model the PageRank as a solution of a stochastic equation $R \stackrel{d}{=} \sum_{i=1}^{N} A_i R_i + B$, where R_i's are distributed as R. This equation is inspired by the original definition of the PageRank. In particular, N models the number of incoming links of a page, and B stays for the user preference. Assuming that N or B are heavy-tailed, we employ the theory of regular variation to obtain the asymptotic behavior of R under quite general assumptions on the involved random variables. Our theoretical predictions show a good agreement with experimental data.

Keywords: PageRank, Web, Regular variation, Stochastic equation, Tauberian theorems

AMS MSC: 68P10, 90B15, 40E05

1 Introduction

Today the World Wide Web is an important part of our life. Thus, understanding properties of the Web is one of the most essential research needs. Web has a complex structure with some notable features. Cardinally, it is huge: by some estimations indexed Web contains at least 27.5 billion pages*; and it continues to grow very fast. Moreover, it has linking or, more precisely, hyperlinking structure. A convenient way to analyze Web structure is to consider the Web as a graph, where pages are nodes, and links are edges. Then we can assign different characteristics for each node in such graph. The terms *in-degree* and *out-degree* are used for the number of incoming and outgoing links of a page, respectively. Further, *PageRank* is a widely accepted notion for characterizing an importance of each node in the graph. It is worth noting that in- and out-degree are natural characteristics of the graph structure while PageRank is a

*www.worldwidewebsize.com; (Accessed in July 2008).

popularity measure designed to enhance Web search. The PageRank as originally introduced by Google is one of significant characteristics that affects the listing of Web pages returned by a search engine in response to a query. We provide formal definition of the PageRank in the Section 1.1.

Most of experimental studies of the Web agree that in-degree, out-degree, and the PageRank on the Web follow power laws. In simple words, a random variable X has a power law distribution with exponent $\alpha > 0$, if its probability of obtaining a value grater than x is proportional to $x^{-\alpha}$. In the Web, the power law exponents can deviate depending on a data set and an estimator but are believed to satisfy $\alpha = 1.1$ for in-degree and PageRank, and $\alpha \approx 2$ for out-degree [5, 24, 28].

The goal of this paper is to provide mathematical evidence for the power law behavior of the PageRank and its relation to the different characteristics of the underlying graph. To this end we propose a stochastic model that is a considerable extension of our previous work [18, 27]. The PageRank is modeled as a solution of a distributional identity, and the tail behavior of such solution is obtained under various assumptions on involved parameters. The generality of our analytical model allows us to take into account many different factors affecting the PageRank, such as personalization of the PageRank as we define in the next section, and a possible dependence between personalized preference scores and in-degrees of the Web pages. The analyzed stochastic equation as described in Section 1.3, is of independent mathematical interest.

1.1 Personalized PageRank

With evolution of the Web, the first search engines quickly became insufficient because the underlying techniques were developed for document collections where all documents were assumed to have high quality, and to be homogeneous. This holds, for example, for collections of papers or books, where the number of citations is a good measure of popularity. However, the homogeneity assumption is definitely violated in a representative collection of Web pages, where the best text match does not implies the highest relevance, and incoming links can often be a spam. In the end of $90'$s Brin and Page with PageRank [23] and Kleinberg with HITS [16] proposed to use hyperlink analysis for measuring importance of pages in Web search. In this work we focus only on PageRank. Originally created for Web ranking, the PageRank has become a major method for evaluating popularity of nodes in various information networks. Besides its primary application in search engines, PageRank is successfully used for solving other important problems such as spam detection [10], graph partitioning [2], and finding gems in scientific citations [6], just to name a few.

Denote by w the number of nodes in the Web graph. The PageRank is defined as a stationary distribution of an 'easily bored surfer' random walk on the graph. At each step, with probability c, such random walk follows a randomly chosen outgoing link of a page, and with probability $[1 - c]$ the walk starts afresh from a page chosen at random according to some *teleportation* distribution. The constant c is called a *damping factor*, and is usually set

$c = 0.85$. If a page is a *dangling* node, i.e. it has no out-going link, then we assume that this page has links to all pages in the network. Denoting the total number of nodes in the Web graph by w, we get that the probability to follow a particular link from such page becomes $1/w$, and it is almost zero for large w.

We can summarize the PageRank definition in the next formula:

$$PR(i) = c \sum_{j \to i} \frac{1}{d_j} PR(j) + \frac{c}{w} \sum_{j \in \mathcal{D}} PR(j) + (1-c)T(i), \; i = 1, \ldots, w, \tag{1}$$

where $PR(i)$ is the PageRank of page i, d_j is out-degree of page j, the sum is taken over all pages j that link to page i, $\mathcal{D}$ is a set of dangling nodes, and $T(i)$ is the probability to start walk afresh in page i. It is clear that the PageRank values in (1) scale as $1/w$ with the number of pages. In our analysis, it is more convenient to deal with corresponding *scale-free PageRank* scores

$$R(i) = wPR(i), \quad i = 1, \ldots, w, \tag{2}$$

assuming that w goes to infinity. In this setting, it is easier to compare the probabilistic properties of PageRank and in- and out-degree, that are also scale-free. In the remainder of the paper, by PageRank we mean the scale-free PageRank scores (2). Then the original definition (1) can be written as

$$R(i) = c \sum_{j \to i} \frac{1}{d_j} R(j) + \frac{c}{w} \sum_{j \in \mathcal{D}} R(j) + (1-c)wT(i), \; i = 1, \ldots, w. \tag{3}$$

In the definition of *standard PageRank* [23], the teleportation distribution is assumed to be uniform, i.e. $T(i) = 1/w$ for every $i = 1, \ldots, w$. However, such approach does not reflect search preferences for deferent users. Page et al. [23] suggest to personalize PageRank by adjustment in the teleportation jumps with respect to the individual user tastes. The knowledge of the user preferences can be based on the usage data, such as browsing histories, or search engine logs; or/and on the user data, such as information about personal characteristics of the user, e.g., name, age, or geographic location [22]. However, the individual-personalized PageRank is computationally infeasible in practice. Then the idea is to build an approximation of such individual PageRank, that is still allows to achieve good level of personalization. Below we list several approaches for this approximation [12]. The Topic-Sensitive PageRank [11] restricts the interests of a user to the small number of topics, say $K = 20$. Then the teleportation jump can be defined as follows: $T(i) = \sum_{i \in J} p_J p_{i,J}$, where p_J is the teleportation probability to the topic J, $J = 1, \ldots, K$, and $p_{i,J}$ is a probability to teleport into particular page i within topic J. Intuitively, if some individuals like to surf for pages about sport, then their search result can be improved by enlarging the $T(i)'$s in (3) for the pages with sport content. Then, the Topic-Sensitive PageRank represents user preferences for the beneficial topics choice. Modular PageRank, that was proposed by Jeh and Widom in [13], is similar to the above approach. However, in this case the surfer teleports to the certain pages with high ranks instead of set of the topic-related pages. In the BlockRank [15] the

Web is considered to be combined from the blocks, for example, each block represents a host. Then, the teleportation jump can be defined as follows: $T(i) = p_J PR_J(i)$, where p_J is a probability to jump into block J, and $PR_J(i)$ is the local PageRank of page i in block J. We also mention next two approaches that personalize PageRank not through the teleportation. The first, the query-dependent PageRank [25], is based on the idea to replace $1/d_j$ in (3) with $p_q(j \to i)$, the probability that random walk follows the link to page i given that it is on page j and is searching for query q. In the second, Constantine and Gleich [7] suggest to modify the damping factor c accordingly to the user surfing properties.

With any of the above mentioned approaches to personalized ranking, the resulting distribution of the PageRank scores for a given Web graph, depends on local graph characteristics such as in-degree and out-degree. In the next section we discuss the tail behavior of the PageRank distribution, and its relations to different parameters in the Web.

1.2 Power law distributions in the Web

It has become a common knowledge that in-degree and PageRank in the Web follow a power law with the same exponent [8, 18, 24, 27]. However, as we saw above, the main idea of PageRank is that it depends not only on quantity but also on quality of incoming links of a page. Moreover, we emphasize that PageRank is a global characteristics of the Web while in-degree is a local one. Thus, the phenomena of asymptotic similarity between in-degree and PageRank is not trivial to justify. In [3, 9] authors verify asymptotic properties of PageRank distribution for the case of preferential attachment models [1], which are often used for simulating graphs with power-law distributed in-degree. In this work, as in [18, 27], we explain asymptotic behavior of PageRank distribution by modeling a personalized PageRank as a solution of a certain stochastic equation.

To obtain the asymptotic behavior of PageRank we employ the theory of regular variation that provides natural mathematical formalism for analyzing power laws. A non-negative random variable X is said to be *regularly varying* with index α, if $\mathbb{P}(X > x) \sim x^{-\alpha} L(x) \quad \text{as} \quad x \to \infty$, for some positive *slowly varying* function $L(x)$ (that is, by definition, for every $y > 0$ we have $L(yx)/L(x) \to 1$ as $x \to \infty$.) Here, as in the remainder of this paper, the notation $a(x) \sim b(x)$ means that $a(x)/b(x) \to 1$. We provide all necessary preliminaries on the theory of regular variation in Appendix A.

1.3 Stochastic equations

From mathematical point of view, this paper presents the analysis of the following distributional identity

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j + B, \tag{4}$$

where we assume that all random variables are positive; R_j's are independent and distributed as R; and A_j's are independent and distributed as some random variable A with $\mathbb{E}(A) = [1 - \mathbb{E}(B)]/\mathbb{E}(N) < 1$. We also set R_j's and A_j's to be independent, and to be independent of N and B. Moreover, it is essential that $\mathbb{E}(B) < 1$. We emphasis that N and B can be dependent.

The equations similar to (4) are well known in the literature. For instance, such equation can also describe the distribution of the busy period in the $M/G/1$ queue:

$$R \stackrel{d}{=} \sum_{j=1}^{N(S_1)} R_j + S_1,$$

where R is the distribution of the busy period (the time interval during which the queue is non-empty), S_1 is the service time of the customer that initiated the busy period, $N(S_1)$ is the number of Poisson arrivals during this service time, and R_j's are independent and distributed as R. We refer to [21, 29] for more details on the asymptotics of a busy period in queues with heavy tails.

Another version of (4) arises in the theory of branching processes. For $B = 0$ we can obtain the following equation:

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j,$$

that has been analyzed in detail by Liu [20, 19].

The rest of the paper is organized as follows. In Section 2 we describe the model for in- and out-degrees, and provide stochastic equation for PageRank in the form (4), where each random variable represents a certain parameter in the Web. In Section 3 we use a probabilistic approach to show that the proposed equation has a unique non-trivial solution with a finite mean. We introduce a recurrent stochastic model for the power iteration algorithm commonly used in PageRank computations [17], and we obtain the PageRank asymptotics after each iteration in Section 3.2. The tail behavior of the PageRank in our model is obtained in Section 4.1. To this end, we use Laplace-Stieltjes transforms and apply Tauberian theorem, see Theorem 8.1.6 in [4], or Theorem A.1 in Appendix A.

Our analysis reveals that the in-degree distribution is not the only determining factor for the asymptotic behavior of the personalized PageRank. It turns out that the teleportation distribution can play a significant role as well. In fact, the asymptotic properties of PageRank as a solution of (4), are defined by the distribution with the heaviest tail. We are also able to explicitly derive the constant multiplicative factor that quantifies the difference between the tail asymptotics of PageRank, in-degree, and teleportation distributions. In Section 5 we show that analytical results are in agreement with the Web data.

2 Model

We persuade the idea suggested at [18, 27] for the case of the personalized PageRank. We start with models for in- and out-degree distributions in the Web. Then, we define PageRank of a random page in the network as the solution of a stochastic equation in Section 2.2.

2.1 In- and out-degree

We start by modeling the in-degree distribution. It is a common believe that in-degree in the Web follows power law with exponent $\alpha_N \approx 1.1$. We set in-degree of a randomly chosen page to be distributed as an integer valued regularly varying random variable N with index $\alpha_N > 1$. One of the ways to model N is as follows: we assume that $N = N(X)$, where X is regularly varying with index α_N and $N(x)$ is the number of Poisson arrivals during the time interval $[0, x]$, when arrival rate is 1. Thus, if X is regularly varying then $N(X)$ is also regularly varying and asymptotically identical to X (see e.g. [18]):

$$\mathbb{P}(X > x) \sim x^{-\alpha_N} L_N(x) \Leftrightarrow \mathbb{P}(N(X) > x) \sim x^{-\alpha_N} L_N(x) \quad \text{as } x \to \infty. \tag{5}$$

Then $N(X)$ is indeed an integer and obeys the power law. We use this representation for N in Section 4.

Next, we model the weights $1/d_j$ in (3). Recall that d_j is the out-degree of page j that has a link to page i. As in [27] we consider a random variable D that represents the out-degree of a page that links to a particular randomly chosen page i. Note that D is *not* the same random variable as an out-degree of a random page since the additional information that a page has a link to i alters the out-degree distribution. This phenomenon is known as inspection paradox [26]. Thus, the number of out-links from a page containing a random link is stochastically larger than an out-degree of a random page. If p_j is a fraction of the pages with out-degree $j \geq 0$, then we can obtain

$$\lim_{w \to \infty} \mathbb{P}(D = j) = j p_j / \mathbb{E}(N), \quad j \geq 1. \tag{6}$$

where $\mathbb{E}(N)$ is the average in/out-degree, and w is the number of pages in the Web. For sufficiently large networks, we may assume that the distribution of D is equal to its limiting distribution as defined by (6). We refer to D as an *effective out-degree*. The term is motivated by the fact that the distribution of D is the one that participates in the PageRank formula (3).

2.2 Stochastic equation for PageRank

Now, we are ready to model the PageRank distribution. We view the PageRank of a random page as a random variable R with $\mathbb{E}(R) = 1$. Further, we assume that the PageRank of a random page does not depend on the fact of whether the page is dangling. We note that such independence immediately implies that in large networks, the fraction of the total PageRank mass concentrated in dangling nodes, equals to the fraction of dangling nodes p_0, simply by the law of large numbers: $p_0 = (1/w) \sum_{j \in \mathcal{D}} R(j)$.

Our goal is to analyze to what extent the tail probability $\mathbb{P}(R > x)$ for large enough x depends on the in-degree N, the effective out-degree D, the teleportation jump T and the fraction of dangling nodes p_0. To this end, we model PageRank R as a solution of a stochastic equation involving N, T and D. Inspired by the original formula (3), the stochastic equation for the PageRank is as follows:

$$R \stackrel{d}{=} c \sum_{j=1}^{N} \frac{1}{D_j} R_j + cp_0 + (1-c)wT. \tag{7}$$

Here R_j's and D_j's are independent and distributed as R and D, respectively. Moreover, we need to assume that R_j's and D_j's are independent and independent of N and T. As before, $c \in (0,1)$ is a damping factor. We emphasis that N and T are allowed to be depended, that is often the case for the personalized PageRank.

Hence, in stochastic equation (7) we generalize models from [18, 27] for the case of random out-degree, and random teleportation jump. Moreover, here we allow this personalization jump to be dependent on the in-degree. In the next section we will show that (7) has a unique solution R such that $\mathbb{E}(R) = 1$.

3 Probabilistic analysis

In the next two sections we will analyze the following stochastic equation

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j + B, \tag{8}$$

where we assume that all random variables are positive; R_j's are independent and distributed as R; and A_j's are independent and distributed as some random variable with $\mathbb{E}(A) = [1 - \mathbb{E}(B)]/\mathbb{E}(N)$. We also set R_j's and A_j's to be independent, and to be independent of N and B. Moreover, it is essential that $\mathbb{E}(B) < 1$. We emphasis that N and B can be dependent.

It is easy to see that the above equation corresponds to (7) for $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)nT$. The notations presented below are adopted from Liu [20].

Let $\{(N_u, A_{u_1}, A_{u_2}, \dots)\}_u$ be a family of independent copies of $(N, A_1, A_2, \dots)$ indexed by all finite sequences $u = u_1 \dots u_i$, where $u_j \in \{1, 2, \dots\}$, $j = 1 \dots i$. And let $\mathbb{T}$ be *the Galton-Watson tree* with defining elements $\{N_u\}$: we have $\emptyset \in \mathbb{T}$ and, if $u \in \mathbb{T}$ and $j \in \{1, 2, \dots\}$, then concatenation $uj \in \mathbb{T}$ if and only if $1 \le j \le N_u$. In other words, we indexed the nodes of the tree with root $\emptyset$ and the first level nodes $1, 2, \dots N_\emptyset$, and at every subsequent level, the jth offspring of u is termed uj (see Figure 1).

We use the next lemma to prove the existence of the solution (8). This lemma is a result mentioned in [20].

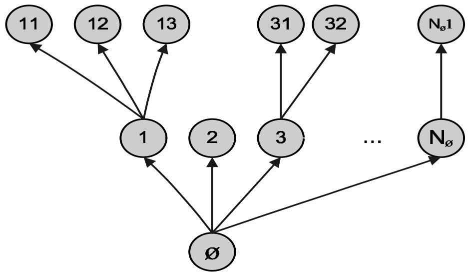


Figure 1: An example of Galton-Watson tree

Lemma 1. *If* $\mathbb{E}\left(\sum_{j=1}^{N} A_j\right) = 1$*, then the sequence* $\sum_{u_1\ldots u_i \in \mathbb{T}} A_{u_1} \ldots A_{u_1\ldots u_i}$ *is a martingale.*

Our main goal is to show how the asymptotics of R in (8) depends on the distribution of N and B. We divide this problem into three possible cases. In the first case, we assume that N is a regularly varying random variable, and B has some distribution with lighter tail. Then we recall that N is an integer valued regularly varying random variable

$$\mathbb{P}(N > x) \sim x^{-\alpha_N} L_N(x) \quad \text{as } x \to \infty.$$

In the second case, we take B to be regularly varying and N to have a lighter tail. Then, we have

$$\mathbb{P}(B > x) \sim x^{-\alpha_B} L_B(x) \quad \text{as } x \to \infty, \tag{9}$$

where $L_B(x)$ is slowly varying function. In the final case, we consider both variables to be regularly varying with the same indexes.

In the remainder of this section we establish the existence and the asymptotic properties of R in (8) using an iterative procedure.

3.1 Iterations

We start with initial distribution $R^{(0)}$, and for every $k \geq 1$, we define the result of the kth iteration of (8) through a distributional identity:

$$R^{(k)} = \sum_{j=1}^{N} A_j R_j^{(k-1)} + B, \tag{10}$$

where $R_j^{(k-1)}$ and A_j, $j \geq 1$, are independent and distributed as $R^{(k-1)}$ and A, respectively.

Repeatedly applying (10), we derive the following representation for $R^{(k)}$,

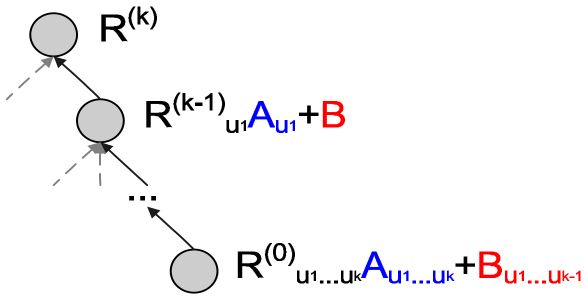


Figure 2: The kth iteration

$k \geq 1$:

$$R^{(k)} = \sum_{u_1 \dots u_k \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_k} R^{(0)}_{u_1 \dots u_k} + \sum_{i=0}^{k-1} \sum_{u_1 \dots u_i \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_i} B_{u_1 \dots u_i}, \tag{11}$$

where $\mathbb{T}$ is a notation for the Galton-Watson tree. In Figure 2 we display the graphic interpretation of $R^{(k)}$.

In the next theorem we show that iterations $R^{(k)}$, $k \geq 1$, converge to the unique solution of (8).

Theorem 1. *Equation (8) has the unique non-trivial solution with mean* 1 *given by*

$$R^{(\infty)} = \lim_{k \to \infty} R^{(k)} = \sum_{i=0}^{\infty} \sum_{u_1 \dots u_i \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_i} B_{u_1 \dots u_i}. \tag{12}$$

Proof. It is easy to verify that $R^{(\infty)}$ in (12) is a well-defined solution of (8). In particular, because all random variables are positive, we apply Fubini's theorem to obtain

$$\mathbb{E}\left(R^{(\infty)}\right) = \mathbb{E}\left[\sum_{i=0}^{\infty} \sum_{u_1 \dots u_i \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_i} B_{u_1 \dots u_i}\right]$$
$$= \mathbb{E}(B) \sum_{i=0}^{\infty} (1 - \mathbb{E}(B))^n \mathbb{E}\left[\sum_{u_1 \dots u_i \in \mathbb{T}} \frac{1}{1 - \mathbb{E}(B)} A_{u_1} \dots \frac{1}{1 - \mathbb{E}(B)} A_{u_1 \dots u_i}\right] = 1,$$

where the final equation holds since $\sum_{u_1 \dots u_i \in \mathbb{T}} (A_{u_1}/(1 - \mathbb{E}(B))) \dots (A_{u_1 \dots u_i}/(1 - \mathbb{E}(B)))$ is a martingale with mean 1 accordingly to Lemma 1. Here we can take $\mathbb{E}(B)$ outside of the summation since $B_{u_1 \dots u_i}$ comes from the $(i-1)$th step, and is independent of the number of incoming links at the level i. We refer to Figure 2 for illustration.

To prove the uniqueness, assume that there is another solution with mean 1 and take this solution as an initial distribution $R^{(0)}$ with $\mathbb{E}(R^{(0)}) = 1$. Consider

$R^{(k)}$, then the first part of (11) has a mean:

$$\mathbb{E}\left(\sum_{u_1\ldots u_k\in\mathbb{T}} A_{u_1}\ldots A_{u_1\ldots u_k} R^{(0)}_{u_1\ldots u_k}\right) = (\mathbb{E}(N))^k \left(\frac{(1-\mathbb{E}(B))}{\mathbb{E}(N)}\right)^k = (1-\mathbb{E}(B))^k,$$

and hence this part converges in probability to 0, as $k \to \infty$, because, by the Markov inequality, the probability that this term is greater than some $\epsilon > 0$ is at most $(1-\mathbb{E}(B))^k/\epsilon \to 0$ as $k \to \infty$. Moreover, the second part of (11) converges a.s. to $R^{(\infty)}$ as $k \to \infty$. It follows that (11) converges to $R^{(\infty)}$ in probability. We conclude that there is no other fixed point of (8) with mean 1 except $R^{(\infty)}$. □

3.2 Asymptotics for Iterations

In this section we will define the asymptotics for the result of every iteration based on the properties of N and B. At this point, we assume that $\mathbb{E}(N)\mathbb{E}(A^\alpha) < 1$, where $\alpha = \min(\alpha_N, \alpha_B)$. In the next theorem we consider the case when the initial distribution $R^{(0)}$ has lighter tail than N or B. This assumption makes sense since iterations usually start with $R^{(0)} \equiv 1$. For the other types of distribution of $R^{(0)}$ we refer to Remark 1.

Theorem 2. (i) *If* $\mathbb{P}(B > x) = o(\mathbb{P}(N > x))$ *and* $\mathbb{P}(R^{(0)} > x) = o(\mathbb{P}(N > x))$, *then for all* $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_N^{(k)}\mathbb{P}(N > x) \text{ as } x \to \infty,$$

where $C_N^{(k)} = (\mathbb{E}(A))^{\alpha_N} \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i$.

(ii) *If* $\mathbb{P}(N > x) = o(\mathbb{P}(B > x))$ *and* $\mathbb{P}(R^{(0)} > x) = o(\mathbb{P}(B > x))$, *then for all* $k \geq 1$,

$$\mathbb{P}(R^{(k)} > x) \sim C_B^{(k)}\mathbb{P}(B > x) \text{ as } x \to \infty,$$

where $C_B^{(k)} = \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_B})\right]^i$.

(iii) *If* $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$ *for some constant* C_{BN}, $\mathbb{P}(R^{(0)} > x) = o(\mathbb{P}(N > x))$, *and* $\mathbb{P}(N > x, B > x) = o(\mathbb{P}(N > x))$, *then for all* $k \geq 1$,

$$\mathbb{P}(R^{(k)} > x) \sim C^{(k)}\mathbb{P}(N > x) \text{ as } x \to \infty,$$

where $C^{(k)} = [C_{BN} + (\mathbb{E}(A))^{\alpha_N}] \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i$.

Proof.

(i) We will use induction. For $k = 1$ we apply Lemma A.1 (i) and (iv) to obtain

$$\mathbb{P}\left(R^{(1)} > x\right) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} > x\right)$$
$$\sim (\mathbb{E}(A))^{\alpha_N}\mathbb{P}(N > x) \text{ as } x \to \infty,$$

since $\mathbb{E}(N) < \infty$, $\mathbb{E}\left(A_1 R_1^{(0)}\right) = \mathbb{E}(A) \; < \infty$, and $\mathbb{P}\left(A_1 R_1^{(0)} > x\right) = o(\mathbb{P}(N > x))$. Now, assume that the result has been shown for the $(k-1)$th iteration, $k \geq 2$, then Lemma A.1 (iii) yields

$$\mathbb{P}\left(A_1 R_1^{(k-1)} > x\right) \sim C_N^{(k-1)} \mathbb{E}\left(A^{\alpha_N}\right) \mathbb{P}(N > x), \tag{13}$$

Because of (13) and $\mathbb{E}\left(A_1 R_1^{(k-1)}\right) = \mathbb{E}(A) < \infty$, we can apply Lemma A.1 (i), and (vi) to obtain

$$\mathbb{P}\left(R^{(k)} > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right)$$
$$\sim \left[C_N^{(k-1)} \mathbb{E}(A^{\alpha_N}) \mathbb{E}(N) + (\mathbb{E}(A))^{\alpha_N}\right] \mathbb{P}(N > x) = C_N^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty.$$

(ii) From Lemma A.1 (i) we have that

$$\mathbb{P}\left(R^{(1)} > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} + B > x\right) \sim \mathbb{P}(B > x) \text{ as } x \to \infty.$$

Assume that the statement holds for $(k-1)$, where $k \geq 2$. Then, from Lemma A.1 (iii) we obtain

$$\mathbb{P}\left(A_1 R_1^{(k-1)} > x\right) \sim C_B^{(k-1)} \mathbb{E}\left(A^{\alpha_B}\right) \mathbb{P}(B > x).$$

Because $\mathbb{E}(N) < \infty$, we apply Lemma A.1 (ii) and (v) to obtain

$$\mathbb{P}\left(R^{(k)} > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right)$$
$$\sim \left[\mathbb{E}(N) C_B^{(k-1)} \mathbb{E}(A^{\alpha_B}) + 1\right] \mathbb{P}(B > x) = C_B^{(k)} \mathbb{P}(B > x) \text{ as } x \to \infty.$$

(iii) We start the induction with $k = 1$ as follows

$$\mathbb{P}\left(R^{(1)} > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} + B > x\right) \sim (\mathbb{E}(A))^{\alpha_N} \mathbb{P}(N > x)$$
$$+ \mathbb{P}(B > x) \sim [(\mathbb{E}(A))^{\alpha_N} + C_{BN}] \mathbb{P}(N > x) \text{ as } x \to \infty,$$

where we use Lemma A.1 (ii) and (iv). Next, from (13), $\mathbb{E}\left(A_1 R_1^{(k-1)}\right) = \mathbb{E}(A) < \infty$, and using of Lemma A.1 (ii) and (vi) we obtain that for any $k \geq 2$:

$$\mathbb{P}\left(R^{(k)} > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right)$$
$$\sim \left[\mathbb{E}(N) C^{(k-1)} \mathbb{E}(A^{\alpha_N}) + (\mathbb{E}(A))^{\alpha_N} + C_{BN}\right] \mathbb{P}(N > x)$$
$$= C^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty.$$

□

In short, Theorem 2 states that the tail behavior of $R^{(k)}$ is determined by the asymptotics of the random variable with the heaviest tail among N and B. Moreover, if the tails of N and B are equally heavy, then in fact we get the sum of two asymptotic expressions.

With $R^{(k)}$ for $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)wT$, the random variable $R^{(k)}$ serves as a stochastic model for the result of the kth matrix iteration [17] in the PageRank computation. Since the PageRank vector is always a result of a finite number of iterations, we can conclude that the distribution of PageRank should follow power law with exponent $\alpha = \min(\alpha_N, \alpha_B)$. However, if the initial distribution $R^{(0)}$ has one of the heaviest tails, then the following results hold.

Remark 1. *Let $R^{(0)}$ be a regularly varying random variable with index $\alpha_R > 0$. Then the following statements hold.*

(i) If $\mathbb{P}(N > x) = o(\mathbb{P}(R^{(0)} > x))$ and $\mathbb{P}(B > x) = o(\mathbb{P}(R^0 > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_R^{(k)} \mathbb{P}(R^{(0)} > x) \text{ as } x \to \infty,$$

where $C_R^{(k)} = \prod_{i=0}^{k} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_R})\right]^i$.

(ii) If $\mathbb{P}(R^0 > x) \sim C_{RN}\mathbb{P}(N > x)$, and $\mathbb{P}(B > x) = o(\mathbb{P}(R^{(0)} > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_{RN}^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty,$$

where $C_{RN}^{(k)} = [\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^k C_{RN} + [\mathbb{E}(A)]^{\alpha_N} \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i$.

(iii) If $\mathbb{P}(N > x) = o(\mathbb{P}(R^{(0)} > x))$, $\mathbb{P}(R^{(0)} > x) \sim C_{RB}\mathbb{P}(B > x)$, and $\mathbb{P}(R^{(0)} > x, B > x) = o(\mathbb{P}(B > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_{RB}^{(k)} \mathbb{P}(B > x) \text{ as } x \to \infty,$$

where $C_{RB}^{(k)} = [\mathbb{E}(N)\mathbb{E}(A^{\alpha_B})]^k C_{RB} + \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_B})\right]^i$.

(iv) If $\mathbb{P}(R^0 > x) \sim C_{RN}\mathbb{P}(N > x)$, $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$, $\mathbb{P}(R^{(0)} > x, N > x) = o(\mathbb{P}(N > x))$, and $\mathbb{P}(B > x, N > x) = o(\mathbb{P}(N > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_{RBN}^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty,$$

$$C_{RBN}^{(k)} = [\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^k C_{RN} + [C_{BN} + [\mathbb{E}(A)]^{\alpha_N}] \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i.$$

Proof. We again use induction. We start with $k = 1$ for which all statements are valid. Next, we assume that result has been shown for $(k-1)$th iteration, where $k > 2$. Then we consider every case respectively.

(i) We apply Lemma A.1 (i), (iii) and (v) to obtain the following:

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$\sim \mathbb{E}(N)\mathbb{E}(A^{\alpha_R})\mathbb{P}(R^{(k-1)} > 0) = C_R^{(k)}\mathbb{P}(R^{(0)} > 0).$$

(ii) In this case we have

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$\sim \left[\mathbb{E}(A^{\alpha_N})\mathbb{E}(N)C_{RN}^{(k-1)} + (\mathbb{E}(A))^{\alpha_N}\right]\mathbb{P}(N > x) = C_{RN}^{(k)}\mathbb{P}(N > x),$$

where we use Lemma A.1 (i), (iii) and (vi).

(iii) From Lemma A.1 (ii), (iii) and (v) we obtain the statement:

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$+ \mathbb{P}(B > x) \sim \left[\mathbb{E}(A^{\alpha_B})\mathbb{E}(N)C_{RB}^{(k-1)} + 1\right]\mathbb{P}(B > x) = C_{RB}^{(k)}\mathbb{P}(B > x).$$

(iv) Here we use Lemma A.1 (ii), (iii) and (vi) and get the following result:

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$+ \mathbb{P}(B > x) \sim \left[\mathbb{E}(A^{\alpha_N})\mathbb{E}(N)C_{RBN}^{(k-1)} + (\mathbb{E}(A))^{\alpha_N} + C_{BN}\right]\mathbb{P}(N > x)$$
$$= C_{RBN}^{(k)}\mathbb{P}(N > x).$$

□

3.3 Asymptotics: from $R^{(k)}$ to $R^{(\infty)}$

Combining results from Theorem 1 and 2, we can presume the following asymptotic similarities for $R^{(\infty)}$, the unique non-trivial solution of (8):

(i) If $\mathbb{P}(B > x) = o(\mathbb{P}(N > x))$, then $\mathbb{P}(R^{(\infty)} > x) \sim C_N\mathbb{P}(N > x)$ as $x \to \infty$, where $C_N = \lim_{k\to\infty} C_N^{(k)} = (\mathbb{E}(A))^{\alpha_N}\left[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^{-1}$.

(ii) If $\mathbb{P}(N > x) = o(\mathbb{P}(B > x))$, then $\mathbb{P}(R^{(\infty)} > x) \sim C_B\mathbb{P}(B > x)$ as $x \to \infty$, where $C_B = \lim_{k\to\infty} C_B^{(k)} = \left[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_B})\right]^{-1}$.

(iii) If $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$ for some constant C_{BN}, and $\mathbb{P}(N > x, B > x) = o(\mathbb{P}(N > x))$, then $\mathbb{P}(R^{(\infty)} > x) \sim C\,\mathbb{P}(N > x)$ as $x \to \infty$, where $C = \lim_{k\to\infty} C^{(k)} = \left[C_{BN} + (\mathbb{E}(A))^{\alpha_N}\right]\left[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^{-1}$.

In the next section we prove the above similarities using the Laplace-Stieltjes transforms analysis. Note that probabilistic analysis is not working in this case because $\mathbb{P}(R^{(\infty)} > x) \sim \lim_{k\to\infty} \mathbb{P}(R^{(k)} > x)$ is not true in general. Indeed, from Remark 1 we know that the asymptotics of $R^{(k)}$ can be defined by the asymptotics of $R^{(0)}$, whereas representation (12) clarifies that $R^{(\infty)}$ does not depend on the distribution of $R^{(0)}$.

4 Laplace-Stieltjes transforms' analysis

As in our previous work [18], we follow technique from [21]. We start with equation for Laplace-Stieltjes transforms of N, B, and R. The idea is to use this equation and Tauberian theorem (Theorem A.1) to classify the asymptotic behavior of R. To this end, we first show that conditions of Theorem A.1 are satisfied. Particularly, in Lemma 2 and 3 we justify that the existence of the kth moments of N and B implies the existence of the kth moment of R, and vice versa. Then, we define the necessary equivalences for the Laplace-Stieltjes transforms of N, B, and R in Corollary 1; and obtain the main result in Theorem 3.

In this section we need to assume that $A < 1$, and $\alpha = \min(\alpha_N, \alpha_B) > 1$ is non-integer. Moreover, we model in-degree N as the number of Poisson(1) events on $[0, X]$, where X is a regular varying random variable with index α_N. Asymptotic behavior of $N(X)$ is given by (5).

4.1 Equation for Laplace-Stieltjes transforms

Define $f(s)$ and $\phi(s)$ to be the Laplace-Stieltjes transforms of X and $N = N(X)$ respectively, where X is regularly varying with index α_N and $N(x)$ is the number of Poisson arrivals on the time interval $[0, x]$, as before. Then we can write the following expression:

$$\phi(s) = E\left(e^{-sN}\right) = f(1 - e^{-s}). \tag{14}$$

Moreover, since the corresponding moments of X and N always exist together [18], we use only moments of X, and we denote them by $\xi_0 = 1,\ \xi_1 = \mathbb{E}(N), \xi_2, \ldots, \xi_n$. Then, provided that ξ_n is finite, we define

$$f_n(s) = (-1)^{n+1}\left(f(s) - \sum_{i=0}^{n} \frac{\xi_i}{i!}(-s)^i\right). \tag{15}$$

Next, we denote the first m moments of B by $\beta_1, \beta_2, \ldots, \beta_m$, and $\beta_0 = 1$. Then, provided that β_m is finite, we define

$$b_m(s) = (-1)^{m+1}\left(b(s) - \sum_{i=0}^{m} \frac{\beta_i}{i!}(-s)^i\right), \tag{16}$$

where $b(s)$ is the Laplace-Stieltjes transform of B.

We also introduce the following function:

$$G(t,s) = \mathbb{E}\left(e^{-tX}e^{-sB}\right), \tag{17}$$

where it is easy to see that $G(t,0) = f(t)$ and $G(0,s) = b(s)$. Moreover, if X and B are independent, implying that N and B are independent, then we have

$$G(t,s) = f(t)b(s).$$

Let $r(s)$ be the Laplace-Stieltjes transform of R. Then, by (8) and (14) the following holds:

$$\begin{aligned} r(s) = \mathbb{E}\left(e^{-sR}\right) &= \mathbb{E}\left[\exp\left(-s\sum_{j=1}^{N} A_j R_j\right) e^{-sB}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left(\exp\left(-s\sum_{j=1}^{N} A_j R_j\right) e^{-sB}\middle| N, B\right)\right] = G\left[1 - \mathbb{E}\left(r\left(As\right)\right), s\right]. \end{aligned}$$

Thus, derive the next equation:

$$r(s) = G\left[1 - \mathbb{E}\left(r\left(As\right)\right), s\right]. \tag{18}$$

Denote

$$t(s) = 1 - \mathbb{E}\left(r\left(As\right)\right), \tag{19}$$

and write (18) as

$$r(s) = G(t(s), s). \tag{20}$$

4.2 Auxiliary results

We define $\rho_1, \ldots, \rho_k$ to be the first k moments of R. If $\rho_k < \infty$, we have the following:

$$r_k(s) = (-1)^{k+1}\left(r(s) - \sum_{i=0}^{k} \frac{\rho_i}{i!}\left(-s\right)^i\right), \tag{21}$$

as in Lemma A.2.

We denote $k = \min(m,n)$, where m and n are integer, and such that $\beta_m = \mathbb{E}(B^m) < \infty$ and $\xi_n = \mathbb{E}(X^n) < \infty$. Next, we assume that $\mathbb{E}(X^j B^{k+1-j}) < \infty$ for all $0 < j < k+1$. We note that this assumption is always true in the case of the independent N and B. Then we can prove the following lemma.

Lemma 2. *If $\xi_n < \infty$ and $\beta_m < \infty$ for some integer $m, n \geq 1$, and $\mathbb{E}(X^j B^{k+1-j}) < \infty$ for all $0 < j < k+1$, where $k = \min(m,n)$; then $\rho_k < \infty$.*

Proof. We use induction, starting from $k = 1$ for which the statement is valid. Assume that for $i = 1, 2, \ldots, k-1$, lemma has been proved, so we can use the following extension:

$$r(s) = 1 - s + \sum_{i=2}^{k-1} \frac{\rho_i}{i!}(-s)^i + o(s^{k-1}),$$

to present $t(s)$ as a sum

$$t(s) = -\mathbb{E}\left(\sum_{i=1}^{k-1} \frac{\rho_i}{i!} A^i (-s)^i + o(s^{k-1})\right) = -\sum_{i=1}^{k-1} \frac{\rho_i}{i!} \mathbb{E}(A^i)(-s)^i + o(s^{k-1}).$$

As the result of this, we can actually obtain $t^i(s)$:

$$t^i(s) = \sum_{j=i}^{k+i-2} \zeta_{i,j} s^j + o(s^{k+i-2}), \tag{22}$$

for $i \geq 1$ and some appropriate constants $\zeta_{i,j},\ j = i, \ldots, k+i-2$.

Now, we consider the Taylor expansion of $G(t(s), s)$:

$$\begin{aligned} G(t(s), s) = &\left[\sum_{i=0}^{k} \frac{\xi_i}{i!}(-t(s))^i + (-1)^{k+1} f_k\left(t(s)\right)\right] \\ &+ \left[\sum_{i=0}^{k} \frac{\beta_i}{i!}(-s)^i + (-1)^{k+1} b_k(s)\right] \\ &- 1 + \sum_{i=0}^{k+1} \frac{(-1)^i}{i!} \sum_{j=1}^{i-1} \binom{i}{j} \mathbb{E}\left(X^j B^{i-j}\right) t^j(s) s^{i-j} + o(s^{k+1}), \end{aligned} \tag{23}$$

where $t(s) \sim \mathbb{E}(A)s$. Here we use that $G'_{t^j s^{i-j}}(0,0) = (-1)^i \mathbb{E}\left(X^j B^{i-j}\right) < \infty$ for all $0 \leq i \leq k+1$ and $0 < j < k+1$. Then, from (19), (20), and (23), we obtain the following:

$$\begin{aligned} r(s) = &1 - \mathbb{E}(N)t(s) + \left[\sum_{i=2}^{k} \frac{\xi_i}{i!}(-t(s))^i + (-1)^{k+1} f_k\left(t(s)\right)\right] + \left[\sum_{i=0}^{k} \frac{\beta_i}{i!}(-s)^i \right. \\ &\left. + (-1)^{k+1} b_k(s)\right] - 1 + \sum_{i=0}^{k+1} \frac{(-1)^i}{i!} \sum_{j=1}^{i-1} \binom{i}{j} \mathbb{E}\left(X^j B^{i-j}\right) t^j(s) s^{i-j} + o(s^{k+1}) \\ = &1 - \mathbb{E}(N)\left[1 - \mathbb{E}(r(As))\right] + \sum_{i=1}^{k} \eta_i s^i + o(s^k), \end{aligned}$$

where we use (22), $f_k(t(s)) = o(s^k)$, and $b_k(s) = o(s^k)$ to find appropriate constants $\eta_1, \ldots, \eta_k$. Next, we rewrite the last equation

$$r(s) - \mathbb{E}(N)\mathbb{E}(r(As)) = 1 - \mathbb{E}(N) + \sum_{i=1}^{k} \eta_i s^i + o(s^k),$$

and apply (21) to obtain the following:

$$r_{k-1}(s) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(As\right)\right) + (-1)^k \sum_{i=0}^{k-1} \frac{\rho_i}{i!}(1-\mathbb{E}(A^i))(-s)^i = 1 - \mathbb{E}(N)$$
$$+ \sum_{i=0}^{k} \eta_i s^i + o(s^k).$$

Because $r_{k-1}(s) = o(s^{k-1})$, $\mathbb{E}(r_{k-1}(As)) = o(s^{k-1})$ and the uniqueness of the series expansion, we can remove all powers up to k:

$$r_{k-1}(s) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(As\right)\right) = \eta_k s^k + o(s^k). \tag{24}$$

Now, we let $A_1, A_2 \ldots$ be independent and distributed as A. We consider the following partial sums

$$\sum_{j=0}^{M} (\mathbb{E}(N))^j \left[\mathbb{E}\left(r_{k-1}(A_1 \ldots A_j s)\right) - \mathbb{E}(N)\, \mathbb{E}\left(r_{k-1}(A_1 \ldots A_{j+1} s)\right)\right]$$
$$= r_{k-1}(s) - (\mathbb{E}(N))^{M+1} \mathbb{E}(r_{k-1}(A_1 \ldots A_{M+1} s))$$

We claim that the second term converges to 0 as $M \to \infty$. From induction hypothesis and the definition of $o(s^{k-1})$, for all $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that $|r_{k-1}(s)| < \varepsilon s^{k-1}$ whenever $0 < s \leq \delta$. Fix some ε and take $\delta = \delta(\varepsilon)$ then the following holds:

$$\mathbb{E}\left|r_{k-1}\left(A_1 \ldots A_{M+1} s\right)\right| < \varepsilon s^{k-1} \mathbb{E}\left(A_1^{k-1} \ldots A_{M+1}^{k-1}\right) = \varepsilon s^{k-1} \left(\mathbb{E}(A^{k-1})\right)^{M+1},$$

where final equation holds since the independence of the A's. Taking the limit as $M \to \infty$, since $\mathbb{E}(B) < 1$, $A < 1$, $\mathbb{E}(A) = (1-\mathbb{E}(B))/\mathbb{E}(N)$ and $\mathbb{E}(A^{n-1}) \leq \mathbb{E}(A)$ we have $\lim_{M\to\infty} \mathbb{E}(N)^{M+1} \mathbb{E}(r_{k-1}(A_1 \ldots A_{M+1} s)) = 0$. It follows that we can express $r_{k-1}(s)$ as an infinite sum:

$$r_{k-1}(s) = \sum_{j=0}^{\infty} (\mathbb{E}(N))^j \left[\mathbb{E}\left(r_{k-1}(A_1 \ldots A_j s)\right) - \mathbb{E}(N)\, \mathbb{E}\left(r_{k-1}(A_1 \ldots A_{j+1} s)\right)\right], \tag{25}$$

where we can apply (24) to each of the terms. Form the definition of $o(s^k)$, for every $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that

$$\left|r_{k-1}(s) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(As\right)\right) - \eta_k s^k\right| < \varepsilon s^k$$

whenever $0 < s \leq \delta$. Moreover, for this ε and $0 < s \leq \delta$, we also have

$$\left|\mathbb{E}\left(r_{k-1}\left(A_1 \ldots A_j s\right)\right) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(A_1 \ldots A_{j+1} s\right)\right) - \eta_k s^k \mathbb{E}\left(A_1^k \ldots A_j^k\right)\right|$$
$$\leq \mathbb{E}\left|\mathbb{E}\left[r_{k-1}\left(A_1 \ldots A_j s\right) - \mathbb{E}(N) r_{k-1}\left(A_1 \ldots A_{j+1} s\right) - \eta_k s^k A_1^k \ldots A_j^k\right.\right.$$
$$\left.\left. \mid A_1, \ldots, A_j\right]\right| < \varepsilon s^k (\mathbb{E}(A^k))^j,$$

for every $j \geq 0$ and $A_1, \ldots, A_{j+1}$, which are independent and distributed as A. Here the last inequality holds because $A < 1$, and then $0 < A_1 \ldots A_{j+1} s \leq s < \delta$ for every $j \geq 0$. Using the representation of $r_{k-1}(s)$ as an infinite sum, (25), we obtain

$$\left| r_{k-1}(s) - \eta_k \sum_{j=0}^{\infty} (\mathbb{E}(N))^j \mathbb{E}\left(A_1^k \ldots A_j^k\right) s^k \right|$$

$$= \left| \sum_{j=0}^{\infty} (\mathbb{E}(N))^j \left[\mathbb{E}\left(r_{k-1}(A_1 \ldots A_j s)\right) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}(A_1 \ldots A_{j+1} s)\right)\right] \right.$$

$$\left. -\eta_k \sum_{j=0}^{\infty} (\mathbb{E}(N))^j \mathbb{E}\left(A_1^k \ldots A_j^k\right) s^k \right| \leq \varepsilon s^k \sum_{j=1}^{\infty} (\mathbb{E}(N)\mathbb{E}(A^k))^j$$

$$= \varepsilon [1 - \mathbb{E}(N)\mathbb{E}(A^k)]^{-1} s^k.$$

Thus, we have shown that $r_{k-1}(s) - \eta_k [1 - \mathbb{E}(N)\mathbb{E}(A^k)]^{-1} s^k = o(s^k)$. Taking $\rho_k = -\eta_k [1 - \mathbb{E}(N)\mathbb{E}(A^k)]^{-1}$, from Lemma A.2 and the last equation we conclude that ρ_k is the kth moment of R and it is finite. □

We can also proof the conversed lemma.

Lemma 3. *If $\rho_k < \infty$, $k \geq 1$, then $\xi_k < \infty$ and $\beta_k < \infty$.*

Proof. Let R be non-negative random variable, that satisfies (8) and has finite kth moment. Equation (8) implies that R is stochastically greater than B, and thus R is also stochastically greater than $B(AN(X) + 1)$. Hence, the existence of the kth moment of R ensures the existence of the kth moment of B and $N(X)$, which in turn ensures the existence of the kth moment of X. □

The next Corollary follows from the proof of Lemma 2.

Corollary 1. *It follows from Lemma 2 that*

(i) if $n < m$, then $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n(As)\right) = f_n(t(s)) + O(s^{n+1})$.

(ii) if $n > m$, then $r_m(s) - \mathbb{E}(N)\mathbb{E}\left(r_m(As)\right) = b_m(s) + O(s^{m+1})$.

(iii) if $n = m$, then $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n(As)\right) = f_n(t(s)) + b_n(s) + O(s^{n+1})$.

Proof. Recall k to be $\min(m, n)$. Because $r_k(s) = o(s^k)$ we can consider the following expansion of (22):

$$t^i(s) = \sum_{j=l}^{k+i-1} \zeta_{i,j} s^j + o(s^{k+i-1}), \tag{26}$$

for $i \geq 1$ and appropriate constants $\zeta_{i,j}$, $j = i, \ldots, k + i - 1$.

From (20), (23), (26), the definitions of $r_k(s)$, $b_k(t)$, $t(s)$, Lemma 2, it follows that

$$
\begin{aligned}
&(-1)^{k+1}r_k(s) + \sum_{i=0}^{k}\frac{\rho_i}{i!}(-s)^i = \Bigg[(-1)^{k+1}f_k\left(t(s)\right) + \sum_{i=2}^{k}\frac{\xi_i}{i!}(-t(s))^i + 1 \\
&-\mathbb{E}(N)\left[1-\mathbb{E}\left((-1)^{k+1}r_k(As) + \sum_{i=0}^{k}\frac{\rho_i}{i!}(-As)^i\right)\right]\Bigg] - 1 + \Bigg[\sum_{i=0}^{k}\frac{\beta_i}{i!}(-s)^i \\
&+(-1)^{k+1}b_k(s)\Bigg] + \sum_{i=0}^{k+1}\frac{(-1)^i}{i!}\sum_{j=1}^{i-1}\binom{i}{j}\mathbb{E}\left(X^jB^{i-j}\right)t^j(s)s^{i-j} + o(s^{k+1}) \\
&= (-1)^{k+1}[b_k(s) + f_k(t) + \mathbb{E}(N)\mathbb{E}(r_k(As))] + \sum_{i=0}^{k+1}\varsigma_i s^i + o(s^{k+1}),
\end{aligned}
$$

where $\varsigma_0, \ldots, \varsigma_{k+1}$ are appropriate constants. Due to the uniqueness of the series expansion, we can reduce the above formula to

$$r_k(s) = b_k(s) + f_k(t) + \mathbb{E}(N)\mathbb{E}(r_k(As)) + (-1)^{k+1}\varsigma_{k+1}s^{k+1} + o(s^{k+1}).$$

Then, since $t(s) \sim \mathbb{E}(A)s$ as $s \to 0$, we get

(i) if $n < m$, then $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n\left(As\right)\right) = f_n(t) + O(t^{n+1})$;

(ii) if $n > m$, then $r_m(s) - \mathbb{E}(N)\mathbb{E}\left(r_m\left(As\right)\right) = b_m(s) + O(t^{m+1})$;

(iii) if $n = m$, then $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n\left(As\right)\right) = f_n(t(s)) + b_n(s) + O(t^{n+1})$.

□

Now we are ready to prove our main result.

4.3 Main theorem

In the next theorem we obtain our main result that establishes the tail behavior of the PageRank distribution under various assumptions on the distribution of the in-degree and the teleportation.

Theorem 3. *(i) if* $\mathbb{P}(B > x) = o(\mathbb{P}(N > x))$, *then the following are equivalent:*

- *(i.1)* $\mathbb{P}(N > x) \sim x^{-\alpha_N}L_N(x)$ *as* $x \to \infty$,
- *(i.2)* $\mathbb{P}(R > x) \sim C_N x^{-\alpha_N}L_N(x)$ *as* $x \to \infty$, *where* $C_N = (E(A))^{\alpha_N}[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^{-1}$;

(ii) if $\mathbb{P}(N > x) = o(\mathbb{P}(B > x))$, *then the following are equivalent:*

- *(ii.1)* $\mathbb{P}(B > x) \sim x^{-\alpha_B}L_B(x)$ *as* $x \to \infty$,
- *(ii.2)* $\mathbb{P}(R > x) \sim C_B x^{-\alpha_B}L_B(x)$ *as* $x \to \infty$, *where* $C_B = [1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_B})]^{-1}$;

(iii) if $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$, *then the following are equivalent:*

(iii.1) $\mathbb{P}(N > x) \sim x^{-\alpha_N} L_N(x)$, *and* $\mathbb{P}(B > x) \sim x^{-\alpha_N} L_B(x) \sim C_{BN} x^{-\alpha_N} L_N(x)$ *as* $x \to \infty$,

(iii.2) $\mathbb{P}(R > x) \sim C_N x^{-\alpha_N} L_N(x)$ *as* $x \to \infty$, *where* $C = [C_{BN} + (\mathbb{E}(A))^{\alpha_N}] \times [1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^{-1}$.

The results of Theorem 3 describe the tail behavior of R under various assumptions on the distribution of the Web parameters. First of all, we observe that the power law exponent is defined by the random variable with the heaviest tail among N and B, representing the in-degree and the user preference, respectively. Next, we see that the obtained multiplicative constants agree with the results of Section 3.3. When B has a lighter tail than N, we observe that the distribution of B has no influence on the asymptotics of the PageRank. In the next case we find that C_B only depends on the mean value of the in-degree $\mathbb{E}(N)$, and in the case of the similar tails of N and B we have the effects from both of them. We also note that if A is defined as c/D, then $\mathbb{E}(A) = c(1 - p_0)/E(N)$. So, the obtained constants also depend on the damping factor c and the fraction of the dangling nodes p_0. The distribution of the effective out-degree D has a negligible effect.

Proof of Theorem 3.

$(i, ii, iii.1) \Rightarrow (i, ii, iii.2)$ It follows from $(i, ii, iii.1)$ and Theorem A.1 that

(*i*) $f_n(t) \sim (-1)^n \Gamma(1 - \alpha_N) t^{\alpha_N} L_N\left(\frac{1}{t}\right)$ as $t \to 0$;

(*ii*) $b_m(s) \sim (-1)^m \Gamma(1 - \alpha_B) s^{\alpha_B} L_B\left(\frac{1}{s}\right)$ as $s \to 0$;

(*iii*) both previous equivalences,

where m and n are the largest integer values not exceeding α_B and α_N, respectively.

Recall that $t(s) \sim \mathbb{E}(A)s$ as $s \to 0$, because of (19) and $r(s) = 1 - s + o(s)$. Then, by applying Corollary 1 we can obtain as $s \to 0$:

(*i*) $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n(As)\right) \sim (-1)^n \Gamma(1 - \alpha_N)(\mathbb{E}(A))^{\alpha_N} L_N\left(\frac{1}{s}\right) s^{\alpha_N}$

(*ii*) $r_m(s) - \mathbb{E}(N)\mathbb{E}\left(r_m(As)\right) \sim (-1)^m \Gamma(1 - \alpha_B) L_B\left(\frac{1}{s}\right) s^{\alpha_B}$

(*iii*) $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n(As)\right) \sim (-1)^n \Gamma(1 - \alpha_N)\left[(\mathbb{E}(A))^{\alpha_N} L_N\left(\frac{1}{s}\right) + L_B\left(\frac{1}{s}\right)\right] s^{\alpha_N}$.

Let V_N and V_B be constants that are defined as follows:
(*i*) $V_N = (\mathbb{E}(A))^{\alpha}$ and $V_B = 0$;
(*ii*) $V_N = 0$ and $V_B = 1$;
(*iii*) $V_N = (\mathbb{E}(A))^{\alpha}$ and $V_B = 1$.
Next, we denote

$$Z(s) = r_k(s) - \mathbb{E}(N)\mathbb{E}\left(r_k(As)\right),$$
$$Y(s) = (-1)^k \Gamma(1 - \alpha)\left[V_N L_N\left(\frac{1}{s}\right) + V_B L_B\left(\frac{1}{s}\right)\right] s^{\alpha},$$

where $\alpha = \min(\alpha_N, \alpha_B)$, and $k = \min(n, m)$. We note that $Y(s) \geq 0$ for every $s > 0$.

We prove the statement of the theorem in two steps. First, we use the representation (25) for $r_k(s)$, and show that the following asymptotic similarity holds:

$$\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Z\left(A_1 \dots A_i s\right)\right) \sim \sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Y\left(A_1 \dots A_i s\right)\right), \tag{27}$$

as $s \to 0$. Second, we demonstrate that the right-hand side of (27) has the desired asymptotics.

As we saw above, $Z(s) \sim Y(s)$ as $s \to 0$. Then, for every $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that $|Z(s)/Y(s) - 1| < \varepsilon$ whenever $0 < s \leq \delta$. We fix some ε and take $\delta = \delta(\varepsilon)$. Now again let $A_1, A_2, \dots$ be independent random variables, which are distributed as A. Because $A < 1$, and then $0 < A_1 \dots A_i s \leq s \leq \delta$, for every $i \geq 0$ we have

$$\left| \frac{Z(A_1 \dots A_i s)}{Y(A_1 \dots A_i s)} - 1 \right| < \varepsilon. \tag{28}$$

From (28) we obtain the following:

$$\begin{aligned}
&\left| \frac{\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Z\left(A_1 \dots A_i s\right)\right)}{\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Y\left(A_1 \dots A_i s\right)\right)} - 1 \right| \\
&\leq \frac{\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \left|\mathbb{E}\left[Z\left(A_1 \dots A_i s\right) - Y\left(A_1 \dots A_i s\right)\right]\right|}{\left|\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Y\left(A_1 \dots A_i s\right)\right)\right|} \\
&\leq \frac{\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left[\left|\frac{Z(A_1 \dots A_i s)}{Y(A_1 \dots A_i s)} - 1\right| Y\left(A_1 \dots A_i s\right)\right]}{\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Y\left(A_1 \dots A_i s\right)\right)} \\
&< \frac{\varepsilon \sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Y\left(A_1 \dots A_i s\right)\right)}{\sum_{i=0}^{\infty} (\mathbb{E}(N))^i \mathbb{E}\left(Y\left(A_1 \dots A_i s\right)\right)} = \varepsilon,
\end{aligned}$$

which implies (27).

Next, we use Lemma A.3, and then for every $\vartheta > 1$ and $\delta > 0$ we can find finite constants s_B and s_N such that for all $i > 0$ and $0 < s < \min(s_B, s_N)$,

$$\begin{aligned}
\vartheta^{-1} \left(A_1 \dots A_i\right)^{\delta} &\leq \frac{L_B\left(\frac{1}{A_1 \dots A_i s}\right)}{L_B\left(\frac{1}{s}\right)} \leq \vartheta \left(A_1 \dots A_i\right)^{-\delta}, \text{ and} \\
\vartheta^{-1} \left(A_1 \dots A_i\right)^{\delta} &\leq \frac{L_N\left(\frac{1}{A_1 \dots A_i s}\right)}{L_N\left(\frac{1}{s}\right)} \leq \vartheta \left(A_1 \dots A_i\right)^{-\delta}.
\end{aligned} \tag{29}$$

We divide the right-hand side of (27) by $L_B\left(\frac{1}{s}\right) L_N\left(\frac{1}{s}\right)$, and apply (29) to

$Y(A_1 \dots A_i s)/L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)$ to obtain the following:

$$\begin{aligned}
&\vartheta^{-1}(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^\alpha\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left((A_1\dots A_i)^{\alpha+\delta}\right)\\
&\quad\le\frac{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_i s\right)\right)}{L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)}\\
&\quad\le\vartheta(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^\alpha\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left((A_1\dots A_i)^{\alpha-\delta}\right).
\end{aligned}$$

Because $A_1, A_2 \dots$ are independent and identically distributed as A we can conclude the following:

$$\begin{aligned}
&\vartheta^{-1}(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^\alpha\frac{1}{1-\mathbb{E}(N)\mathbb{E}\left(A^{\alpha+\delta}\right)}\\
&\quad\le\frac{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_i s\right)\right)}{L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)}\\
&\quad\le\vartheta(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^\alpha\frac{1}{1-\mathbb{E}(N)\mathbb{E}\left(A^{\alpha-\delta}\right)}.
\end{aligned}$$

Taking $\vartheta \to 1$ and $\delta \to 0$ by the dominated convergence we obtain

$$\begin{aligned}
&\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_i s\right)\right)\sim(-1)^k\Gamma(1-\alpha)[1-\mathbb{E}(N)\mathbb{E}(A^\alpha)]^{-1}\\
&\times\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)s^\alpha \text{ as } s\to 0.
\end{aligned}$$

Combining the last equivalence, (27), and the infinite-sum representation (25) for $r_k(s)$:

$$r_k(s)=\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\left[\mathbb{E}\left(r_k\left(A_1\dots A_i s\right)\right)-\mathbb{E}(N)\mathbb{E}\left(r_k\left(A_1\dots A_i s\right)\right)\right], \tag{30}$$

we then obtain

$$r_k(s)\sim(-1)^k\Gamma(1-\alpha)\left[V_N L_N\left(\frac{1}{s}\right)+V_B L_B\left(\frac{1}{s}\right)\right][1-\mathbb{E}(N)\mathbb{E}(A^\alpha)]^{-1}s^\alpha \tag{31}$$

as $s \to 0$. Now, we again apply Theorem A.1 that leads to the statement of the theorem.

$(i, ii, iii.1) \Leftarrow (i, ii, iii.2)$ We denote V_N and V_B, $k = \min(n, m)$, and $\alpha \in (k, k+1)$, as before. Then, from $(i, ii, iii.2)$, and Theorem A.1 we can obtain (31), that leads to the asymptotic equivalence:

$$r_k(s)-\mathbb{E}(N)\mathbb{E}(r_k(As))\sim(-1)^k\Gamma(1-\alpha)L\left(\frac{1}{s}\right)[1-\mathbb{E}(N)\mathbb{E}(A^\alpha)]^{-1}s^\alpha, \tag{32}$$

as $s \to 0$, where we denote

$$L\left(\frac{1}{s}\right) = V_N\left[L_N\left(\frac{1}{s}\right) - \mathbb{E}(N)\mathbb{E}\left(A^\alpha L_N\left(\frac{1}{As}\right)\right)\right]$$
$$+ V_B\left[L_B\left(\frac{1}{s}\right) - \mathbb{E}(N)\mathbb{E}\left(A^\alpha L_B\left(\frac{1}{As}\right)\right)\right]$$

Next, we again use bounds (29) to obtain

$$\left[\frac{V_N}{L_B\left(\frac{1}{s}\right)} + \frac{V_B}{L_N\left(\frac{1}{s}\right)}\right]\left[1 - \vartheta^{-1}\mathbb{E}(N)\mathbb{E}(A^{\alpha+\delta})\right] \le \frac{L\left(\frac{1}{s}\right)}{L_N\left(\frac{1}{s}\right)L_B\left(\frac{1}{s}\right)}$$
$$\le \left[\frac{V_N}{L_B\left(\frac{1}{s}\right)} + \frac{V_B}{L_N\left(\frac{1}{s}\right)}\right]\left[1 - \vartheta\mathbb{E}(N)\mathbb{E}(A^{\alpha-\delta})\right]$$

Thus, by the dominated convergence for $\vartheta \to 1$ and $\delta \to 0$ we have

$$L\left(\frac{1}{s}\right) \sim \left[1 - \mathbb{E}(N)\mathbb{E}(A^\alpha)\right]\left[C_N L_N\left(\frac{1}{s}\right) + C_B L_B\left(\frac{1}{s}\right)\right].$$

From last similarity and (32) we obtain

$$r_k(s) - \mathbb{E}(N)\mathbb{E}(r(As)) \sim (-1)^k\Gamma(1-\alpha)\left[V_N L_N\left(\frac{1}{s}\right) + V_B L_B\left(\frac{1}{s}\right)\right]s^\alpha,$$

as $s \to 0$, from where by applying Corollary 1 we show $(i, ii, iii.1)$. □

5 Numerical results

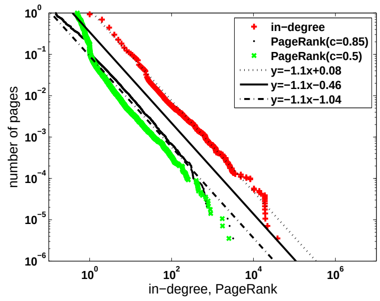
In order to illustrate the results of Theorem 3 we perform a number of small scale experiments. More numerical results can be found in [27], where we considered a simpler model of the standard PageRank with uniform teleportation. Here we use Stanford data set* with $w = 281.903$ pages and 2.312.497 links. It is a relatively small Web sample, however, it is known to possess basic properties of the Web. In particular, in this data set, in-degree shows typical power law behavior with exponent $\alpha_N = 1.1$.

We create the teleportation distribution by using inverse transformation method. First, we generate random numbers $u_1, \ldots, u_w$ from the standard uniform distribution, and then we set $t_i = (1-u_i)^{-1/\alpha_B}$, where $i = 1, \ldots, w$. These t_i's are random numbers that are Pareto distributed with exponent α_B. We choose $\alpha_B = 0.5$, 1.1 and 3.0. Second, we denote $\bar{t}$ as the mean value of $t_1, \ldots, t_w$, and define the teleportation probability of a jump to page i as $T(i) = t_i/(w\bar{t})$. Next, we use formula (3) to obtain personalized PageRanks. We also compute PageRank with uniform teleportation jumps. The calculation

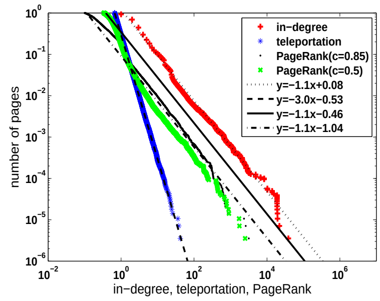
*www.stanford.edu/∼sdkamvar/research.html; (Accessed in March 2006).

of PageRank is done by applying the matrix power iteration method (see [17] for more details).

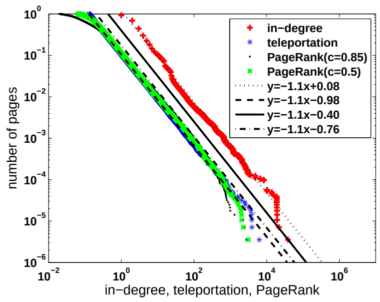
In Figure 3(a)-(d) we present cumulative log-log plots for in-degree, teleportation and PageRanks for damping factors $c = 0.5$ and $c = 0.85$. Here we consider a scale-free teleportation, so we plot complementary cumulative distribution function $\mathbb{P}(wT > x) = (\bar{t}x)^{-\alpha_B}$. Then, $y = -\alpha_B x - \alpha_B * \log_{10}(\bar{t})$ is the straight line that corresponds to the teleportation log-log plot. We also fit in-degree plot with the straight line $y = -1.1x + 0.08$.



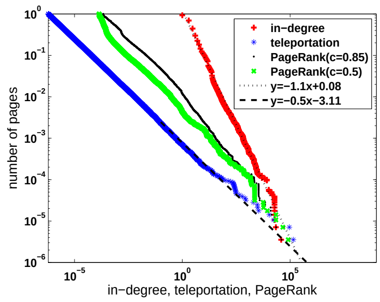
(a) uniform



(b) $\alpha_B = 3.0$



(c) $\alpha_B = 1.1$



(d) $\alpha_B = 0.5$

Figure 3: Number of pages with in-degree/teleportation/PageRank greater than x versus x in log-log scale.

First, we consider the log-log plots of the standard PageRank with uniform teleportation (see Figure 3(a)). In this case we use Theorem 3(i) for $A \stackrel{d}{=} c/D$ to obtain the distance between in-degree and PageRank log-log plots as

$$\log_{10}(C_N) = \log_{10}\left[\frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}(1 - c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N}))}\right], \tag{33}$$

where, as before, N is the in-degree, and D is the effective out-degree. From $\mathbb{E}(N) = 8.2032$, $p_0 = 0.006$ and $\mathbb{E}(1/D^{1.1}) = 0.1043$, we predict the PageRank log-log plots: $y = -1.1x - 0.46$ for $c = 0.85$, and $y = -1.1x - 1.04$ for $c = 0.5$. In the plot we show these theoretically predicted lines and experimental PageRank log-log plots. We see that both lines perfectly match the slopes of the PageRanks, and they trace the direction of changes in the PageRank distribution in respect with changes of the damping factor. Indeed, the plot of the PageRank with $c = 0.5$ is further from the in-degree log-log plot, then the plot of the PageRank with $c = 0.85$. We note that we underestimate the predicted distance in the case of $c = 0.85$, that can be caused by some assumptions of our model. We refer to Section 6 for discussion.

We again use the results of Theorem 3(i) for the case of the PageRank with teleportation that follows power law with exponent $\alpha_B = 3.0$. Then we end up with the same constant as in (33), and therefore we get the same predicted lines for the PageRank log-log plots: $y = -1.1x - 0.46$ for $c = 0.85$, and $y = -1.1x - 1.04$ for $c = 0.5$. In Figure 3(b) we plot the distributions of the teleportation and the PageRanks along with the predicted straight lines. The results are similar to the previous case. Thus, we can see that the distribution of the teleportation has no influence on the tail behavior of the PageRank in case when the teleportation has a lighter tail than the in-degree.

Next, we consider the $T(i)$'s with $\alpha_B = 1.1$ and define $B(i) = cp_0 + (1 - c)wT(i)$, where $i = 1, \ldots, w$. Then, $\mathbb{P}(B > x) \sim (1-c)^{\alpha_B}\mathbb{P}(wT > x) \sim C_{NB}\mathbb{P}(N > x)$ as $x \to \infty$. Because $y = -1.1x + 0.08$ and $y = -1.1x - 0.98$ are the fitted lines for log-log plots for in-degree and teleportation, respectively, we can find that $C_{NB} = 0.0108$ for $c = 0.85$, and $C_{NB} = 0.4063$ for $c = 0.5$. So, in the case when the in-degree and the teleportation are regular varying with the same index $\alpha_N = \alpha_B = 1.1$, we can define the distance in the following way:

$$\log_{10}(C) = \log_{10}\left[\frac{(\mathbb{E}(N))^{\alpha_N} C_{NB} + c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}(1 - c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N}))}\right]. \tag{34}$$

We apply these constants in the above formula to obtain $y = -1.1x - 0.41$ and $y = -1.1x - 0.76$ for PageRank plots for $c = 0.85$ and $c = 0.5$, respectively. We plot these lines in Figure 3(c). Compared to Figures 3(a) and (b), here the teleportation distribution smooths the log-log plots of the PageRanks. Thus, we can hardly see the difference between the plots for $c = 0.5$ and $c = 0.85$. The slopes of the experimental PageRanks again correspond to the predicted power law exponent 1.1. The differences between the log-log plots of the in-degree and the PageRanks agree better than in the previous cases.

Finally, we present results for the teleportation with power law exponent $\alpha_B = 0.5$ in Figure 3(d). Note that we can not find the distance in this case, because the first moment of B does not exist. However, we can clearly see that the PageRank tends to follow a power law with the same exponent as the teleportation distribution.

Note that the constant in (33) is the same as the predicted constant from [27], where we assume that the out-degree is random and the teleportation is uniform.

Furthermore, from the Jensen's inequality $\mathbb{E}(1/D^{\alpha_N}) \geq (\mathbb{E}(1/D))^{\alpha_N} = [(1 - p_0)/\mathbb{E}(N)]^{\alpha_N}$, it follows that

$$C_N \geq \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}[1 - c^{\alpha_N}(1-p_0)^{\alpha_N}(\mathbb{E}(N))^{1-\alpha_N}]}. \tag{35}$$

The last expression is the value of C_N in case when the out-degree of all non-dangling nodes is a constant $\mathbb{E}(N)/(1-p_0)$ as in [18]. If $\alpha_N = 1.1$, then the difference between the left- and the right-hand sides of (35) is really small for any reasonable out-degree distribution. We can also ignore the term $c^{\alpha_N}(1 - p_0)^{\alpha_N}(\mathbb{E}(N))^{1-\alpha_N}$ in (33), then C_N can be approximated from above as follows

$$C_N \geq \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}} = c^{\alpha_N}\left[\mathbb{E}\left(\frac{1}{D}\right)\right]^{\alpha_N} = C'_N.$$

Note that the asymptotic equivalence $\mathbb{P}(R > x) \sim C'_N \mathbb{P}(N > x)$ as $x \to \infty$ holds if we assume that the values of the PageRank R can be approximated by $cN\mathbb{E}(1/D)$ as proposed in [8]. Furthermore, we can repeat a similar reasoning for (34) to obtain

$$C \geq \frac{(\mathbb{E}(N))^{\alpha_N} C_{NB} + c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}[1 - c^{\alpha_N}(1-p_0)^{\alpha_N}(\mathbb{E}(N))^{1-\alpha_N}]} \geq C_{NB} + c^{\alpha_N}\left[\mathbb{E}\left(\frac{1}{D}\right)\right]^{\alpha_N}.$$

6 Conclusions

This work has proposed a generalized stochastic model that characterizes the distribution of the personalized PageRank scores. Under various assumptions on the distribution of the Web parameters and teleportation, the model captures essential features of the PageRank tail behavior, and reveals which properties of the Web graph influence this behavior the most. In particular, the results show that the in-degree and, sometimes, the teleportation play an important role while the influence of the out-degree distribution is minimal. The results have been obtained by means of analyzing the asymptotic properties of the solution of a stochastic equation that is related to branching processes and, to the best of our knowledge, has not been studied to this extent before.

Our results are in a good agreement with the Web data. The differences between the model and the data depend on many factors, in particular, on the choice of a data set, as we observed in [27]. Furthermore, the assumption of the branching structure of the Web implicitly made in (7) is probably not justified. Future work could try to investigate how to improve the model in that respect, mainly by studying the dependencies amongst the R_i's in (7), or between the R_i's on the one hand and N on the other.

Acknowledgements

We would like to thank Bert Zwart for useful discussions. This work is supported by NWO Meervoud grant no. 632.002.401. Part of this research has been funded by the Dutch BSIK/BRICKS project. This article is also the result of joint research in the 3TU Centre of Competence NIRICT (Netherlands Institute for Research on ICT) within the Federation of Three Universities of Technology in The Netherlands.

References

[1] Albert, R. and Barabási, A. L. (1999). Emergence of scaling in random networks. *Science* **286,** 509–512.

[2] Andersen, R., Chung, F. and Lang, K. (2006). Local graph partitioning using PageRank vectors. In *Proceedings of FOCS2006.* pp. 475–486.

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

[4] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1989). *Regular Variation.* Cambridge University Press.

[5] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Statac, R., Tomkins, A. and Wiener, J. (2000). Graph structure in the Web. *Comput. Networks* **33,** 309–320.

[6] Chen, P., Xie, H., Maslov, S. and Redner, S. (2007). Finding scientific gems with Googles PageRank algorithm. *J.Informet.* **1,** 8–15.

[7] Constantine, P. and Gleich, D. (2007). Using polynomial chaos to compute the influence of multiple random surfers in the PageRank model. In *Proceeding of WAW2007.* vol. 4863 of *LNCS.* pp. 82–95.

[8] Fortunato, S., Boguñá, M., Flammini, A. and Menczer, F. (2006). Approximating PageRank from in-degree. In *Proceeding of WAW2007.* vol. 4936 of *LNCS.* pp. 59–71.

[9] Fortunato, S. and Flammini, A. (2006). Random walks on directed networks: the case of PageRank. *Technical report* 0604203. arXiv/physics.

[10] Gyöngyi, Z., Garcia-Molina, H. and Pedersen, J. (2004). Combating Web spam with TrustRank. In *Proceeding of VLDB2004.* pp. 576–587.

[11] Haveliwala, T. H. (2003). Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering* **15,** 784–796.

[12] Haveliwala, T. H., Kamvar, S. and Jeh, G. (2003). An analytical comparison of approaches to personalizing PageRank. *Technical report.* Stanford University.

[13] Jeh, G. and Widom, J. (2003). Scaling personalized Web search. In *Proceeding of WWW2003.* pp. 271–279.

[14] Jessen, A. H. and Mikosch, T. (2006). Regularly varying functions. *Publications de l'institut mathematique, Nouvelle série* **79(93),**.

[15] Kamvar, S. D., Haveliwala, T. H., Manning, C. D. and Golub, G. H. (2003). Exploiting the block structure of the web for computing. *Technical report.* Stanford University.

[16] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *JACM* **46,** 604–632.

[17] Langville, A. N. and Meyer, C. D. (2006). *Google's PageRank and beyond: the science of search engine rankings.* Princeton University Press, Princeton, NJ.

[18] Litvak, N., Scheinhardt, W. R. W. and Volkovich, Y. In-degree and PageRank: Why do they follow similar power laws? *To appear in Internet Math.*.

[19] Liu, Q. (1998). Fixed points of a generalized smoothing transformation and applications to the branching random walk. *Adv. in Appl. Probab* **30,** 85–112.

[20] Liu, Q. (2001). Asymptotic properties and absolute continuity of laws stable by random weighted mean. *Stochastic Processes and their Applications* **95,** 83–107.

[21] Meyer, A. D. and Teugels, J. (1980). On the asymptotic behaviour of the distributions of the busy period and service time in M/G/1. *J. App. Probab.* **17,** 802–813.

[22] Micarelli, A., Gasparetti, F., Sciarrone, F. and Gauch, S. (2007). Personalized search on the World Wide Web. *LNCS* **4321,** 195–230.

[23] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. *Technical report.* Stanford Digital Library Technologies Project.

[24] Pandurangan, G., Raghavan, P. and Upfal, E. (2002). Using PageRank to characterize Web structure. In *Proceeding of COCOON2002*.

[25] Richardson, M. and Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in PageRank. *Adv. NIPS* **14,** 1441–1448.

[26] Ross, S. (2003). The inspection paradox. *Probability in the Engineering and Informational Sciences* **17,** 47–51.

[27] Volkovich, Y., Litvak, N. and Donato, D. (2007). Determining factors behind the pagerank log-log plot. In *Proceeding of WAW2007.* vol. 4863 of *LNCS.* pp. 108–123.

[28] Volkovich, Y., Litvak, N. and Zwart, B. (2008). A framework for evaluating statistical dependencies and rank correlations in power law graphs. Memorandum 1868. University of Twente Enschede.

[29] Zwart, A. (2001). Queueing systems with heavy tails. *PhD thesis.* Eindhoven University of Technology.

A Regular variation preliminaries

The theory of regular variation is a natural mathematical formalism for analyzing power laws. In this section we provide main definitions and some facts that we will use throughout this paper. For more details, we refer to the classic book by Bingham et al. [4], and to the recent review by Jessen and Mikosch [14].

The next lemma describes the asymptotic behavior of product, sum and random sums of regularly varying random variables. We use these results for defining asymptotic properties of PageRank, when the PageRank is a result of the finite number of the iteration steps (see Section 3). In the lemma, relation (*iii*) is known as Breiman's theorem (see e.g. Lemma 4.2.(1) in [14]). Properties (*iv*), (*v*), and (*vi*) are statements (2), (1) and (5) of Lemma 3.7 in [14], respectively. The similarity for sums (*i*) and (*ii*) follows from Lemma 3.12 and 3.1 in [14], respectively.

Lemma A.1. (*i*) *Assume that X_1 is non-negative regularly varying random variable with index $\alpha \geq 0$. If random variable $X_2 > 0$ is such that $\mathbb{P}(X_2 > x) = o(\mathbb{P}(X_1 > x))$, then*

$$\mathbb{P}(X_1 + X_2 > x) \sim \mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(*ii*) *Assume that X_1 is non-negative regularly varying random variable with index $\alpha \geq 0$. If random variable $X_2 > 0$ satisfies $\mathbb{P}(X_2 > x) \sim C\mathbb{P}(X_1 > x)$ for some $C > 0$, and $\mathbb{P}(X_1 > x, X_2 > x) = o(\mathbb{P}(X_1 > x))$, then*

$$\mathbb{P}(X_1 + X_2 > x) \sim (1 + C)\mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(*iii*) *Assume that X_1 and X_2 are two independent non-negative random variables such that X_1 is regularly varying with index α and that $\mathbb{E}(X_2^{\alpha+\epsilon}) < \infty$ for some $\epsilon > 0$. Then*

$$\mathbb{P}(X_1 X_2 > x) \sim \mathbb{E}(X_2^{\alpha})\mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(*iv*) *Assume that N is regularly varying with index $\alpha \geq 0$; if $\alpha = 1$, then assume that $\mathbb{E}(N) < \infty$. Moreover, let (X_i) be i.i.d. sequence such that $\mathbb{E}(X_1) < \infty$ and $\mathbb{P}(X_1 > x) = o(\mathbb{P}(N > x))$. Then as $x \to \infty$,*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i > x\right) \sim (\mathbb{E}(X_1))^{\alpha}\mathbb{P}(N > x) \text{ as } x \to \infty.$$

(*v*) *Assume* (X_i) *is i.i.d. sequence of regular varying random variables with index* $\alpha > 0$, $\mathbb{E}(N) < \infty$, *and* $\mathbb{P}(N > x) = o(\mathbb{P}(X_1 > x))$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i > x\right) \sim \mathbb{E}(N)\mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(*vi*) *Assume that* $\mathbb{P}(X_1 > x) \sim C\ \mathbb{P}(N > x)$ *for some* $C > 0$, *that* X_1 *is regularly varying with index* $\alpha \geq 1$, *and* $\mathbb{E}(X_1) < \infty$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i > x\right) \sim (C\ \mathbb{E}(N) + (\mathbb{E}(X_1))^{\alpha})\mathbb{P}(N > x) \text{ as } x \to \infty.$$

In this paper we present PageRank as a solution of stochastic equation. In order to define its asymptotics, we need to use *the Laplace-Stieltjes transforms* analysis (see Section 4). We denote by $f(s) = \mathbb{E}e^{-sX}$, $s > 0$, the Laplace-Stieltjes transform of X, and let $\xi_i = \int_0^\infty x^i dF_X(x)$ be the ith moment of X, where F_X is distribution function of X. The successive moments of X can be obtained by expanding $f(s)$ in a series at $s = 0$. More precisely, we write the following.

Lemma A.2. *The nth moment of* X *is finite if and only if there exist finite numbers* $\xi_0 = 1$ *and* $\xi_1, ..., \xi_n$, *such that*

$$f_n(s) = (-1)^{n+1}\left(f(s) - \sum_{i=0}^{n} \frac{\xi_i}{i!}(-s)^i\right) = o(s^n) \text{ as } s \to 0.$$

In that case, ξ_i *is the ith moment of* X.

The following theorem establishes the relation between the asymptotic behavior of a regularly varying distribution and its Laplace-Stieltjes transform. We use this result in the proof of Theorem 3.

Theorem A.1. *(Tauberian Theorem) If* $n \in \mathbb{N}$, $\xi_n < \infty$, $\alpha \in (n, n+1)$, *then the following are equivalent*

(*i*) $f_n(s) \sim (-1)^n \Gamma(1-\alpha) s^{\alpha} L(\frac{1}{s})$ *as* $s \to 0$,

(*ii*) $\mathbb{P}(X > x) \sim x^{-\alpha} L(x)$ *as* $x \to \infty$.

The next lemma provides a useful bound for slowly varying functions.

Lemma A.3. *(Potter bounds) Let* L *be a slowly varying function. Then, for any fixed* $\vartheta > 1, \delta > 0$ *there exists a finite constant* $s_0 < 1$ *such that for all* $s_1, s_2 < s_0$,

$$\frac{L\left(\frac{1}{s_1}\right)}{L\left(\frac{1}{s_2}\right)} \leq \vartheta \max\left\{\left(\frac{s_1}{s_2}\right)^{\delta}, \left(\frac{s_1}{s_2}\right)^{-\delta}\right\}.$$