Characterization of Tail Dependence for In-Degree and PageRank*

Nelly Litvak[1,**], Werner Scheinhardt[1], Yana Volkovich[1], and Bert Zwart[2]

[1] University of Twente, Dept. of Applied Mathematics,
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{n.litvak,w.r.w.scheinhardt,y.volkovich}@ewi.utwente.nl
[2] CWI, Science Park Amsterdam, Kruislaan 413,
1098 SJ Amsterdam, The Netherlands
bert.zwart@cwi.nl

Abstract. The dependencies between power law parameters such as in-degree and PageRank, can be characterized by the so-called angular measure, a notion used in extreme value theory to describe the dependency between very large values of coordinates of a random vector. Basing on an analytical stochastic model, we argue that the angular measure for in-degree and personalized PageRank is concentrated in two points. This corresponds to the two main factors for high ranking: large in-degree and a high rank of one of the ancestors. Furthermore, we can formally establish the relative importance of these two factors.

Keywords: Power law graphs, PageRank, Regular variation, Multivariate extremes.

1 Introduction

Large self-organizing networks, such as Internet, the World Wide Web, social and biological networks, often exhibit power laws. In simple words, a random variable X has a power law distribution with exponent $\alpha > 0$ if its tail probability $\mathbb{P}(X > x)$ is roughly proportional to $x^{-\alpha}$, for large enough x. Power law distributions are *heavy-tailed* since the tail probability decreases much slower than negative exponential, and thus one can sometimes observe extremely large values of X. Statistical analysis of complex networks characterized by power laws has received a massive attention in recent literature, see e.g. [1,2,3] for excellent surveys. Nevertheless, we are still far from complete understanding of the structure of such networks. In particular, the question of measuring dependencies between network parameters remains an open and complex issue [1].

* 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.

** This author is supported by NWO Meervoud grant no. 632.002.401.

K. Avrachenkov, D. Donato, and N. Litvak (Eds.): WAW 2009, LNCS 5427, pp. 90–103, 2009.

A common example of two related power law characteristics is that of in-degree and PageRank of a Web page [4,5,6]. The (personalized) PageRank is defined in [7] as follows:

$$PR(k) = c \sum_{i=1}^{IN(k)} \frac{1}{OUT(k_i)} PR(k_i) + (1-c)PREF(k),\ k = 1, \ldots, n, \quad (1)$$

where $PR(k)$ is the PageRank of page k, n is the number of nodes in the network, $IN(k)$ in the in-degree of k, the sum is taken over all pages k_i that link to page k, $OUT(k_i)$ is the number of outgoing links of page k_i, $PREF(k)$ is the preference of the user for page k, with $\sum_{k=1}^{n} PREF(k) = n$, and $c \in (0,1)$ is a damping factor. If there is no outgoing link from a page then we say that the page is *dangling* and assume that it links to all nodes in the network. The PageRank in (1) is uniquely defined, and the PageRanks of all pages sum up to n. We note that in the literature the PageRank and the user preference vectors are often viewed as probability vectors, normalized to sum up to one.

Clearly, the PageRank is influenced largely by in-degree. However, there is still no agreement in the literature on the dependence between these two quantities. In particular, the values of the correlation coefficient vary considerably in different studies [4,8]. This only confirms that the correlation coefficient is an uninformative dependence measure in heavy-tailed (power law) data [9,1,10]. In fact, the correlation coefficient is a 'crude summary' of dependencies that is most informative for jointly normal random variables. It is a common and simple technique but it is not subtle enough to distinguish between the dependencies in large and in small values. This becomes a problem if we want to measure the dependence between two heavy tailed network parameters, because in that case we are mainly interested in the dependence between extremely large values.

We propose to solve the problem of evaluating the dependencies between network parameters, using the theory of multivariate extremes. This theory operates with the notion of *tail dependence* for a random vector (X, Y), that is, the dependence between extremely large values of X and Y. Such tail dependence is characterized by an *angular measure* on $[0,1]$ (see Section 4 for a formal definition). Informally, a concentration of the angular measure around 0 and/or 1 signals independence, while concentration around some other number $a \in (0,1)$ suggests that a certain fraction of large values of Y comes together with large values of X.

In [11,12] a first attempt was made to compute the angular measure between in-degree and PageRank, and completely different dependence structures were discovered in Wikipedia (independence), preferential attachment networks (complete dependence) and the Web (intermediate case). In this paper the goal is to compute the angular measure analytically, based on the stochastic model proposed in [13,6,14]. The resulting angular measure is concentrated in points 0 and $a \in (1/2, 1)$, and the mass distribution depends on the network parameters. Such angular measure is a formalization of the common understanding that there are two main sources for high ranking: high in-degree and a high rank of one of the ancestors. Furthermore, the fraction of the measure mass in 0 stands for the

proportion of highly ranked nodes that have a low in-degree. Thus, we obtain a description of the dependence structure, that is more informative and relates better to reality than the correlation coefficient.

In order to derive the tail dependencies, we employ the theory of regular variation, that provides a natural mathematical formalism for analyzing power laws [10]. By definition, the random variable X is *regularly varying* with index α, if $\mathbb{P}(X > u) = u^{-\alpha}L(u)$, $u > 0$, where $L(u)$ is a slowly varying function, that is, for $x > 0$, $L(ux)/L(u) \to 1$ as $u \to \infty$, for instance, $L(u)$ may be equal to a constant or $\log(u)$. In Section 2 we describe the model where power law network parameters are represented by regularly varying random variables. Basing on this model, the results on tail dependence are derived in Sections 3 and 4, while some of the proofs are deferred to the Appendix. In Section 5 we discuss the results and compare our findings to the graph data.

The derived two-point measure is only a first-order approximation of the complex angular measure observed on the data, since the realistic situation is way more complex than our simplified model. Further modifications of the model are needed in order to adequately describe the dependencies in real-life networks.

2 Model and Preliminaries

Choose a random node in the graph, let N and R denote its in-degree and PageRank, respectively, and let D_i denote the out-degree of its ith ancestor, where $i = 1, \dots, N$. As in [6,13,14] we assume that N and R are random variables that satisfy

$$R \stackrel{d}{=} c\sum_{i=1}^{N} \frac{1}{D_i} R_i + (1-c)T. \tag{2}$$

Here N, R_i's, D_i's and T are independent; R_i's are distributed as R with $\mathbb{E}R = 1$; $a \stackrel{d}{=} b$ means that a and b have the same probability distribution, and $c \in (0,1)$ is the damping factor. The equation above clearly corresponds to the definition of personalized PageRank (1). We note that compared to our previous work [6,13], here we account for personalization by setting T to be random. In this paper we neglect the presence of dangling nodes but they can be easily included in the model (see e.g. [6]).

For convenience we prefer to work with the following, slightly more general, representation of (2):

$$R \stackrel{d}{=} \sum_{i=1}^{N} A_i R_i + B, \tag{3}$$

where A_i's are independent and distributed as some random variable $A < 1$, and $B > 0$ is independent of the A_i's. Next, we define

$$\bar{F}_1(u) := \mathbb{P}(N > u) \qquad \text{and} \qquad \bar{F}_2(u) := \mathbb{P}(R > u), \qquad u > 0,$$

and assume that $\bar{F}_1(u)$ is regularly varying with index $\alpha > 1$. We also assume that B in (3) has a lighter tail than N, that is, $\mathbb{P}(B > u) = o(\mathbb{P}(N > u))$ as

$u \to \infty$. As a result, $\bar{F}_2(u)$ is also regularly varying. In fact, the next proposition was proved in [6,13]; a more general case is presented in [14]. For technical reasons, in [6,13,14] it was assumed that the index α is non-integer.

Proposition 1. *Under the assumptions above,*

$$\bar{F}_2(u) \sim K\bar{F}_1(u) \qquad as \qquad u \to \infty,$$

where $a \sim b$ means that $a/b \to 1$. The value of K depends on the precise assumptions on the A_i's and B; if $\mathbb{E}N = d$, $A = c/d$ and $B = 1 - c$ as in [13], we have

$$K = \frac{c^\alpha}{d^\alpha - dc^\alpha}. \tag{4}$$

In the sequel we will only use the specific form (4) in Corollary 1 and Section 5. We also note that within the same model (3), we could assume that the distribution of the R_i's is different from the one of R. In this case, if the tail of the R_i's is not heavier than the one of N, Proposition 1 still holds, only K will depend on the behavior of $\mathbb{P}(R > u)$ as $u \to \infty$ (see Lemma 3.7 in [15]).

We need to deal with a minor complication because $\bar{F}_1$ is not strictly decreasing, and we will in the sequel need to consider the behavior of its inverse function for small arguments. Instead of working with the generalized inverse $\bar{F}_1^{-1}(v) = \inf\{u > 0 : \bar{F}_1(u) \leq v\}$, which would make the proofs more involved, we prefer to simply work with some function that is strictly decreasing and asymptotically equivalent to $\bar{F}_1(u)$. Such a function can e.g. be defined as $f_1(u) := (1 + e^{-u})\bar{F}_1(u)$, for which the inverse function is well-defined. Thus, we arrive at the following:

$$\begin{aligned} \bar{F}_1(u) &:= \mathbb{P}(N > u) \sim f_1(u), \quad \text{as } u \to \infty \\ \bar{F}_2(u) &:= \mathbb{P}(R > u) \sim f_2(u), \quad \text{as } u \to \infty, \end{aligned} \tag{5}$$

where

$$f_1(u) = u^{-\alpha}L(u), \qquad f_2(u) = Ku^{-\alpha}L(u) = Kf_1(u),$$

for some slowly varying function $L(\cdot)$.

3 Tail Dependence

Let us introduce two functions that are defined on $\mathbb{R}_+^2$, namely the *stable tail dependence function* [9],

$$\ell(x, y) = \lim_{t \downarrow 0} t^{-1}\mathbb{P}(\bar{F}_1(N) \leq tx \text{ or } \bar{F}_2(R) \leq ty) \tag{6}$$

and the function

$$r(x, y) := \lim_{t \downarrow 0} t^{-1}\mathbb{P}(\bar{F}_1(N) \leq tx, \bar{F}_2(R) \leq ty).$$

Provided that the limit in (6) exists, these are closely related. In fact adding them gives

$$\ell(x,y) + r(x,y) = \lim_{t \downarrow 0} t^{-1}(\mathbb{P}(\bar{F}_1(N) \le tx) + \mathbb{P}(\bar{F}_2(R) \le ty)),$$

which would yield $x+y$ if $\bar{F}_1$ and $\bar{F}_2$ were stricly decreasing, because then $\bar{F}_1(N)$ and $\bar{F}_2(R)$ would be uniform random variables on $(0,1)$. The following lemma shows that this result holds anyway.

Lemma 1. *The functions ℓ and r satisfy $\ell(x,y) + r(x,y) = x + y$.*

Proof. We use the function f_1 to show that $\lim_{t \downarrow 0} t^{-1}(\mathbb{P}(\bar{F}_1(N) \le tx) = x$, as follows (the corresponding result for $\mathbb{P}(\bar{F}_2(R) \le ty)$ is proven analogously). Since $\bar{F}_1(u) \to 0$ and $|\bar{F}_1(u) - f_1(u)| = o(\bar{F}_1(u))$ as $u \to \infty$, then for any small $\varepsilon > 0$ we can choose t_1 small enough so that for any $t \le t_1$ and $u > 0$ that satisfy $\bar{F}_1(u) \le tx$ we also have $|\bar{F}_1(u) - f_1(u)| \le \varepsilon |\bar{F}_1(u)|$, and hence $|\bar{F}_1(u) - f_1(u)| \le \varepsilon tx$. If we now fix some small $\varepsilon > 0$, the above implies for any $t \le t_1$ that

$$\begin{aligned}
\mathbb{P}(\bar{F}_1(N) \le tx) &= \mathbb{P}(f_1(N) \le (f_1(N) - \bar{F}_1(N)) + tx) \\
&\le \mathbb{P}(f_1(N) \le (1+\varepsilon)tx) \;=\; \mathbb{P}\left(N \ge f_1^{-1}((1+\varepsilon)tx)\right) \\
&= \bar{F}_1(f_1^{-1}((1+\varepsilon)tx)) \;\sim\; f_1(f_1^{-1}((1+\varepsilon)tx)) \;=\; (1+\varepsilon)tx.
\end{aligned}$$

So we obtain

$$\begin{aligned}
&\limsup_{t \to 0} t^{-1}\ \mathbb{P}(\bar{F}_1(N) \le tx) \le (1+\varepsilon)x, \qquad \text{and similarly,} \\
&\liminf_{t \to 0} t^{-1}\ \mathbb{P}(\bar{F}_1(N) \le tx) \ge (1-\varepsilon)x.
\end{aligned}$$

The result now follows by letting ε go to 0.

The main result of this section gives the stable tail dependence function for N and R:

Theorem 1. *The function $r(x,y)$ for N and R is given by*

$$r(x,y) = \min\{x, y(\mathbb{E}A)^\alpha / K\}. \tag{7}$$

Consequently, $\ell(x,y) = \max\{y, x + y(1 - (\mathbb{E}A)^\alpha / K)\}$.

In the remainder of the paper we will mainly work with $r(x,y)$ rather than $\ell(x,y)$, since its derivation is more appealing.

To prove Theorem 1 we need to use the following lemma.

Lemma 2. *As $u \to \infty$, the following asymptotic relation holds for any constant $C > 0$,*

$$\mathbb{P}(N > u, R > Cu) \sim \min\{f_1(u), (\mathbb{E}A/C)^\alpha f_1(u)\}.$$

We refer to the Appendix for the proof of this lemma, but the intuition behind it is clear. It follows from (3) and the strong law of large numbers that when N is large, we have $R \approx \mathbb{E}A \cdot N$. Therefore, when $\mathbb{E}A > C$, the event $\{R > Cu\}$ is already 'implied' by $\{N > u\}$, so the joint probability behaves as $\mathbb{P}(N > u)$. When $\mathbb{E}A < C$, N needs to be larger for $R > Cu$ to hold, and the joint probability behaves like $\mathbb{P}(N > uC/\mathbb{E}A)$.

In order to understand Theorem 1 we fix $x, y > 0$ throughout this section and rewrite the joint probability in a form that enables application of Lemma 2. The schematic derivation is as follows, where the superscripts denote three issues to be resolved:

$$\begin{aligned}
&\mathbb{P}(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty) \\
&\quad \overset{1}{\sim} \mathbb{P}(f_1(N) \le tx, f_2(R) \le ty) \;=\; \mathbb{P}(N \ge f_1^{-1}(tx), R \ge f_2^{-1}(ty)) \\
&\quad \overset{2}{=} \mathbb{P}\left(N \ge f_1^{-1}(tx), R \ge \left(\frac{y}{Kx}\frac{L(f_1^{-1}(tx))}{L(f_2^{-1}(ty))}\right)^{-1/\alpha} f_1^{-1}(tx)\right) \\
&\quad \overset{3,1}{\sim} \mathbb{P}\left(N \ge f_1^{-1}(tx), R \ge \left(\frac{y}{Kx}\right)^{-1/\alpha} f_1^{-1}(tx)\right) \qquad (8)
\end{aligned}$$

The statement of Theorem 1 now follows from Lemma 2 since obviously $f_1(f_1^{-1}(tx)) = tx$, provided that each of the three steps indicated in (8) is justified. We resolve these issues as follows:

1. We deduce the asymptotic equivalence of the two probabilities from the asymptotic equivalence of the functions *inside* the probabilities. This step is intuitively clear but not mathematically rigorous. In the proof of Theorem 1 we will make the argument precise, see the Appendix.
2. This step is fairly straightforward. Indeed, $v = f_1(u) = u^{-\alpha}L(u)$ implies $u = (v/L(u))^{-1/\alpha}$, so $f_1^{-1}(v) = (v/L(f_1^{-1}(v)))^{-1/\alpha}$. Also, since $f_2(u) = Kf_1(u)$ we have $f_2^{-1}(v) = f_1^{-1}(v/K) = (v/KL(f_2^{-1}(v)))^{-1/\alpha}$. Hence,
$$\frac{f_2^{-1}(ty)}{f_1^{-1}(tx)} = \left(\frac{y}{Kx}\frac{L(f_1^{-1}(tx))}{L(f_2^{-1}(ty))}\right)^{-1/\alpha}. \qquad (9)$$
3. This is a consequence of the following statement (the proof of which can be found in the Appendix), combined with issue 1.

Lemma 3. *For all $x, y > 0$ we have $L(f_1^{-1}(tx)) \sim L(f_2^{-1}(ty))$ as $t \downarrow 0$.*

Now, in order to prove Theorem 1 we only need to resolve issue 1 twice in the derivation in (8). The formal proof of this can be found in the Appendix.

4 Angular Measure

In this section we find the angular measure that corresponds to the function $r(x, y)$ we found, but first we will give some preliminaries. In extreme value

theory (see [9]), it has been shown that a unique (nonnegative) measure $H(\cdot)$ exists on the set $\Xi = \{\omega \in \mathbb{R}_+^2 \mid ||\omega|| = 1\}$, such that the stable tail dependence function ℓ can be expressed as

$$\ell(x, y) = \int_{\Xi} \max(\omega_1 x, \omega_2 y) H(d\omega). \tag{10}$$

Here $||\cdot||$ is a norm that may be chosen freely, but for (10) to hold, the measure has to be normalized in such a way that

$$\int_{\Xi} \omega_1 H(d\omega) = \int_{\Xi} \omega_2 H(d\omega) = 1,$$

so that we have $\ell(x, 0) = x$ and $\ell(0, y) = y$, as should. In this work we choose the $||\cdot||_1$ norm, for which $||\omega||_1 = |\omega_1| + |\omega_2|$, since that is easiest to work with. Then (10) can be rewritten as

$$\ell(x, y) = \int_0^1 \max\{wx, (1-w)y\} H(dw),$$

and the normalization becomes

$$\int_0^1 wH(dw) = \int_0^1 (1-w)H(dw) = 1. \tag{11}$$

Here we let $w = \omega_1$, and we identify the measures on Ξ and $[0, 1]$. By (11) it follows that the function $r(x, y)$ can be written as

$$\begin{aligned} r(x, y) &= \int_0^1 wxH(dw) + \int_0^1 (1-w)yH(dw) - \int_0^1 \max\{wx, (1-w)y\} H(dw) \\ &= \int_0^1 \min\{wx, (1-w)y\} H(dw). \end{aligned} \tag{12}$$

We will now derive the function $r(x, y)$ in case when the angular measure has masses in 0 and a only, as we suspect to be the case for in-degree and PageRank. First of all, the normalization (11) boils down to $aH(a) = H(0) + (1-a)H(a) = 1$, which is easily solved to give

$$H(0) = 2 - 1/a \qquad \text{and} \qquad H(a) = 1/a. \tag{13}$$

Note that H has total measure 2 (as also follows for the general case by summing both integrals in (11)), and that $H(0) > 0$ implies $a > 1/2$. Combining (12) and (13), the function $r(x, y)$ can now be written as

$$r(x, y) = \min\{x, (1/a - 1)y\}.$$

This is a very similar form as we found earlier in (7), and it is not difficult to see that the expressions are equal for $a = K/(K + (\mathbb{E}A)^\alpha)$. Since the angular measure is uniquely determined by the stable tail dependence function ℓ, see [9], and hence by the function r, we showed that the angular measure of N and R is indeed a two-point measure. After using (13) we arrive at

Theorem 2. *The angular measure with respect to the* $||\cdot||_1$ *norm of* N *and* R *is a two-point measure, with masses*

$$H(0) = 1 - \frac{(\mathbb{E}A)^\alpha}{K} \qquad \text{in } 0,$$
$$H(a) = 1 + \frac{(\mathbb{E}A)^\alpha}{K} \qquad \text{in } a = \frac{K}{K + (\mathbb{E}A)^\alpha}.$$

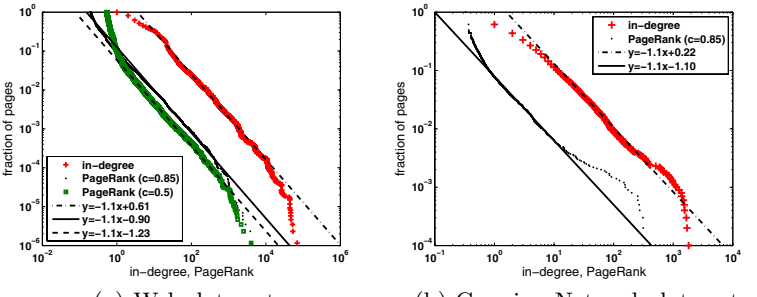
Corollary 1. *If, as in [13], K is given by (4) with* $\mathbb{E}N = d$ *and* $\mathbb{E}A = c/d$*, then the angular measure of N and R is a two-point measure, with masses*

$$\begin{aligned} H(0) &= c^\alpha d^{(1-\alpha)} && \text{in } 0, \\ H(a) &= 2 - c^\alpha d^{(1-\alpha)} && \text{in } a = \left(2 - c^\alpha d^{(1-\alpha)}\right)^{-1}. \end{aligned} \tag{14}$$

5 Examples and Discussion

We compare the above results to the measurements on two different network structures: Web and Growing Network data sets. For the Web sample we choose the EU-2005 data set with 862.664 nodes and 19.235.140 links. This set was collected by The Laboratory for Web Algorithmics (LAW) of the Università degli studi di Milano [16], and is available at `http://law.dsi.unimi.it/`. In this data set in-degree and PageRank exhibit well known power law behavior with exponent $\alpha = 1.1$. For the evaluation of the exponent we refer to [12]. In Figure 1(a) we present log-log plots for in-degree and PageRanks with $c = 0.85$ and $c = 0.5$ (the straight lines are fitted). We also simulate a Growing Network of 10.000 nodes with constant out-degree $d = 8$. We start with d initial nodes, and at each step we add a new node that links to already existing nodes. A new link points to a randomly chosen page with probability $q = 0.1$, and with probability $(1 - q)$ it follows the preferential attachment selection rule [17]. We present log-log plots for this Growing Network set in Figure 1(b).

Following [9, p.328] we define an estimator of the angular measure. We are interested in the dependencies between two regularly varying characteristics of a node, namely the in-degree N and the PageRank R. Let (N_j, R_j) be observations



(a) Web data set (b) Growing Network data set

Fig. 1. Cumulative log-log plots for in-degree and PageRanks

of (N, R) for the corresponding node j. Then we use the rank transformation of (N, R), leading to $\{(r_j^N, r_j^R),\ 1 \le j \le n\}$, where r_j^N is the descending rank of N_j in $(N_1, \ldots, N_n)$ and r_j^R is the descending rank of R_j in $(R_1, \ldots, R_n)$. Next we apply a coordinate transform $(r_j^N, r_j^R) \longrightarrow (r_j, \Theta_j)$, given by

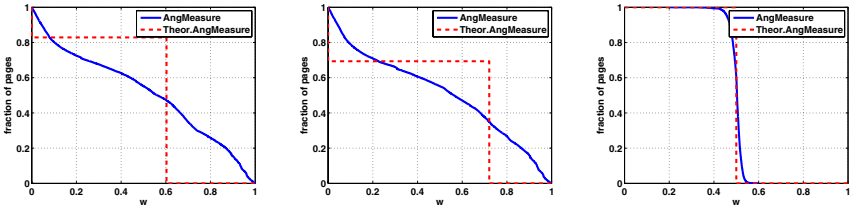
$$(r_j, \Theta_j) = \mathrm{TRANS}\left(\frac{1}{r_j^N}, \frac{1}{r_j^R}\right),$$

where we set $\mathrm{TRANS}(x, y) := (x + y, x/(x + y))$ since all results of this paper are proven for the $||\cdot||_1$ norm. Alternatively, we could use the polar coordinate transformation as in [11,12]: $\mathrm{TRANS}(x, y) := \left(\sqrt{x^2 + y^2}, \arctan(y/x)\right)$. However, in this case we need to transform the angular measure in Theorem 2 to the corresponding measure w.r.t. the $||\cdot||_2$ norm using formula (8.38) in [9]. Now we need to consider k points $\{\Theta_j : r_j \ge r_{(k)}\}$, where $r_{(k)}$ is the kth largest in $(r_1, \ldots, r_n)$, and make a plot for the cumulative distribution function of Θ, which gives the estimation of the probability measure $H(\cdot)/2$. The question how to choose the right k can be solved by employing the Starica plot (see [10,12]).

From (14) we can calculate the predicted angular measure concentrated in 0 and a. For the Web data sample with average in-degree $d = 22.2974$, taking $c = 0.5$ and $c = 0.85$, we obtain that $a_{0.5} = 0.6031$, $H(a_{0.5})/2 = 0.8290$, and $a_{0.85} = 0.7210$, $H(a_{0.85})/2 = 0.6934$, respectively. Recall that the values of $H(a)/2$ estimate the fraction of highly ranked pages whose large PageRank is explained by large in-degree. Observe that according to the model, this fraction becomes larger if c decreases.

In Figure 2 (a,b) we plot the theoretical angular measures together with the empirical ones. The comparison between the graphs shows that there is only a very rough similarity to be seen, in the sense that the value of $H(0)/2$ is a reasonable estimate for the fraction of pages with high PageRank and small in-degree (corresponding to the 'turn' around 0.8). However, the 'point mass' at a seems to be spread out in an almost uniform manner. To understand this, we should realize that the theoretical two-point measure we found is only a formalization of the idea that each large PageRank value has to be either due to a large in-degree, or due to a large contributing PageRank. In the data (representing 'reality'), such a strict division is not reasonable; for instance there will surely be pages with high PageRank due to a high in-degree *and* a high contributing PageRank, or due to more than one high contributing PageRanks. Thus we see that although our model roughly captures the idea of different causes for large PageRank values, it is not subtle enough to properly represent the angular measure as found from a realistic data set. In particular, the assumption of the branching structure of the Web in (2) 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 in (2), or between the R_i on the one hand and N on the other.

Finally, we perform experiments on the Growing Network. It was proved in [18] that the PageRank in such models follows a power law with the same exponent as the in-degree. However, in our model based on stochastic equation (2) we



(a) Web data set: c=0.5, k=100.000 (b) Web data set: c=0.85, k=100.000 (c) Growing Network data set: c=0.85, k=100.000

Fig. 2. Angular measure and theoretically predicted angular measure

cannot assume anymore that R is distributed as the R_i's since R_i's are the ranks of 'younger' nodes, and presumably, the R_i will have lighter tails than R itself. Assuming that $\mathbb{P}(R_i > u) = o(\mathbb{P}(N > u))$ as $u \to \infty$, from Lemma 3.7 in [15] we obtain that for this simple model the value of K is just $K = (c/d)^\alpha$. Substituting this into (14) gives us $a = 1/2$, $H(a) = 2$, and $H(0) = 0$, i.e. the measure is concentrated in one point $a = 1/2$. In Figure 2 (c) we again plot the empirical and theoretical measures, which match perfectly. We see that in synthetic graphs constructed by the preferential attachment rule, large PageRank is always due to large in-degree, and this can be easily captured by our stochastic model.

In further research, it will be interesting to consider other graph models of the Web, for instance, a configuration model, where the degree of each node is chosen independently according to a pre-defined power law distribution [19,20]. The configuration model is not as centered as the preferential attachment network, and it is known to be close to the tree structure. Thus, one may expect that equation (2) provides an accurate description of the dependencies between in-degree and PageRank in such a model.

Finally, we would like to note that by measuring and comparison of tail dependencies in synthetic graphs and experimental data one can easily reveal whether a specific graph model adequately reflects the dependence structure observed in the experiments. From this point of view, the analysis of tail dependencies contributes towards better modelling and understanding of real-life networks.

References

1. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Comput. Surv. 38(1), 2 (2006)
2. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet Math. 1(2), 226–251 (2004)
3. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. 45(2), 167–256 (2003)
4. Donato, D., Laura, L., Leonardi, S., Millozi, S.: Large scale properties of the Webgraph. Eur. Phys. J. 38, 239–243 (2004)
5. Pandurangan, G., Raghavan, P., Upfal, E.: Using PageRank to characterize Web structure. In: Ibarra, O.H., Zhang, L. (eds.) COCOON 2002. LNCS, vol. 2387, p. 330. Springer, Heidelberg (2002)

6. Volkovich, Y., Litvak, N., Donato, D.: Determining factors behind the PageRank log-log plot. In: Bonato, A., Chung, F.R.K. (eds.) WAW 2007. LNCS, vol. 4863, pp. 108–123. Springer, Heidelberg (2007)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Comput. Networks 33, 107–117 (1998)
8. Fortunato, S., Boguñá, M., Flammini, A., Menczer, F.: Approximating PageRank from in-degree. In: Aiello, W., Broder, A., Janssen, J., Milios, E.E. (eds.) WAW 2006. LNCS, vol. 4936, pp. 59–71. Springer, Heidelberg (2008)
9. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes: Theory and Applications. Wiley, Chichester (2004)
10. Resnick, S.I.: Heavy-tail Phenomena. Springer, New York (2007)
11. Volkovich, Y., Litvak, N., Zwart, B.: Measuring extremal dependencies in Web graphs. In: WWW 2008: Proceedings of the 17th international conference on World Wide Web, pp. 1113–1114. ACM Press, New York (2008)
12. Volkovich, Y., Litvak, N., Zwart, B.: A framework for evaluating statistical dependencies and rank correlations in power law graphs. Memorandum 1868, Enschede (2008)
13. Litvak, N., Scheinhardt, W.R.W., Volkovich, Y.: Probabilistic relation between in-degree and PageRank. In: Aiello, W., Broder, A., Janssen, J., Milios, E.E. (eds.) WAW 2006. LNCS, vol. 4936, pp. 72–83. Springer, Heidelberg (2008)
14. Volkovich, Y., Litvak, N.: Asymptotic analysis for personalized Web search. Memorandum 1884, Enschede (2008)
15. Jessen, A.H., Mikosch, T.: Regularly varying functions. Publications de l'institut mathematique, Nouvelle série 79(93) (2006)
16. Boldi, P., Vigna, S.: The WebGraph framework I: Compression techniques. In: WWW 2004: Proceedings of the 13th International Conference on World Wide Web, pp. 595–601. ACM Press, New York (2004)
17. Albert, R., Barabási, A.L.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
18. Avrachenkov, K., Lebedev, D.: PageRank of scale-free growing networks. Internet Math. 3(2), 207–231 (2006)
19. van der Hofstad, R., Hooghiemstra, G., van Mieghem, P.: Distances in random graphs with finite variance degrees. Random Structures Algorithms 27(1), 76–123 (2005)
20. van der Hofstad, R., Hooghiemstra, G., Znamenski, D.: Distances in random graphs with finite mean and infinite variance degrees. Electron. J. Probab. 12(25), 703–766 (2007)

A Proofs

Proof (of Lemma 2). The proof is based on the strong law of large numbers. Informally, we use the fact that if N is large, then (3) implies $R \approx \mathbb{E}A \cdot N$.

Assume first that $C < \mathbb{E}A$. Then we write

$$\mathbb{P}(N > u, R > Cu) = \mathbb{P}(N > u)\mathbb{P}(R > Cu | N > u), \tag{15}$$

and we further obtain

$$\mathbb{P}(R > Cu | N > u) \geq \mathbb{P}\left(\sum_{i=1}^{\lfloor u \rfloor} A_i R_i + B > Cu\right) \geq \mathbb{P}\left(\sum_{i=1}^{\lfloor u \rfloor} A_i R_i > Cu\right)$$
$$= \mathbb{P}\left(C^{-1}u^{-1}\sum_{i=1}^{\lfloor u \rfloor} A_i R_i > 1\right) \to 1 \text{ as } u \to \infty,$$

where the convergence holds by the strong law of large numbers for any $C < \mathbb{E}A$. Hence when $C < \mathbb{E}A$ the result follows directly from (5) and (15).

Now assume that $C > \mathbb{E}A$. We would like to show that

$$\lim_{u\to\infty} \frac{\mathbb{P}(N > u, R > Cu)}{f_1([C/\mathbb{E}A]u)} \to 1. \tag{16}$$

Then the result of the lemma will follow since $L(u) \sim L([C/\mathbb{E}A]u)$ as $u \to \infty$. For the proof, we choose a sufficiently small δ so that we can break the joint probability into three terms:

$$\begin{aligned}\mathbb{P}(N > u, R > Cu) = {} & \mathbb{P}(N > [C/\mathbb{E}A + \delta]u, R > Cu) \\ & + \mathbb{P}([C/\mathbb{E}A - \delta]u < N \leq [C/\mathbb{E}A + \delta]u, R > Cu) \\ & + \mathbb{P}(u < N \leq [C/\mathbb{E}A - \delta]u, R > Cu).\end{aligned} \tag{17}$$

Exactly as in case $C < \mathbb{E}A$, using (5), we have

$$\lim_{u\to\infty} \frac{\mathbb{P}(N > [C/\mathbb{E}A + \delta]u, R > Cu)}{f_1([C/\mathbb{E}A]u)} = \lim_{u\to\infty} \frac{\mathbb{P}(N > [C/\mathbb{E}A + \delta]u)}{f_1([C/\mathbb{E}A]u)} = 1 + O(\delta). \tag{18}$$

Moreover, applying the argument as in the case when $C < \mathbb{E}A$, from the law of large numbers we obtain that

$$\mathbb{P}(R > Cu | u < N \leq [C/\mathbb{E}A - \delta]u) \to 0 \text{ as } u \to \infty,$$

and thus

$$\begin{aligned}0 &\leq \lim_{u\to\infty} \frac{\mathbb{P}(u < N \leq [C/\mathbb{E}A - \delta]u, R > Cu)}{f_1([C/\mathbb{E}A]u)} \\ &\leq \lim_{u\to\infty} \frac{\mathbb{P}(N > u)\mathbb{P}(R > Cu | u < N \leq [C/\mathbb{E}A - \delta]u)}{f_1([C/\mathbb{E}A]u)} = 0.\end{aligned} \tag{19}$$

Finally, we get

$$\begin{aligned} 0 &\le \lim_{u\to\infty} \frac{\mathbb{P}([C/\mathbb{E}A-\delta]u < N \le [C/\mathbb{E}A+\delta]u, R > Cu)}{\mathbb{P}(N > [C/\mathbb{E}A]u)} \\ &\le \lim_{u\to\infty} \frac{\mathbb{P}([C/\mathbb{E}A-\delta]u < N \le [C/\mathbb{E}A+\delta]u)}{\mathbb{P}(N > [C/\mathbb{E}A]u)} \\ &= \lim_{u\to\infty} \frac{f_1([C/\mathbb{E}A-\delta]u) - f_1([C/\mathbb{E}A+\delta]u)}{f_1([C/\mathbb{E}A]u)} = O(\delta). \end{aligned} \tag{20}$$

The result (16) now follows from (17)–(20) by letting $\delta \downarrow 0$.

In the case $C = \mathbb{E}A$ the argument is similar, only we write

$$\begin{aligned} \mathbb{P}(N > u, R > \mathbb{E}Au) = \mathbb{P}(N > [C/\mathbb{E}A+\delta]u, R > Cu) \\ + \mathbb{P}(u < N \le [C/\mathbb{E}A+\delta]u, R > Cu). \end{aligned}$$

This completes the proof of the lemma.

Proof (of Lemma 3). It will be convenient to use the functions

$$g_1(t) := f_1^{-1}(tx) \qquad \text{and} \qquad g_2(t) := f_2^{-1}(ty) = f_1^{-1}(ty/K) = g_1(ty/Kx), \tag{21}$$

which, for fixed $x, y > 0$, are well-defined for all $t > 0$, due to the monotonicity of f_1, and hence also f_2. Applying the Potter bounds, see Resnick [10, p.32], we obtain that for all $A > 1, \delta > 0$ one can choose t sufficiently small such that

$$A^{-1}\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{-\delta} \le \frac{L(g_1(t))}{L(g_2(t))} \le A\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{\delta}$$

which by (9) is the same as

$$A^{-1}\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{-\delta} \le \frac{Kx}{y}\left(\frac{g_1(t)}{g_2(t)}\right)^{\alpha} \le A\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{\delta}.$$

From the first inequality above we get

$$\liminf_{t\downarrow 0} A^{1/\alpha}\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{\delta/\alpha}\left(\frac{Kx}{y}\right)^{1/\alpha}\frac{g_1(t)}{g_2(t)} \ge 1$$

for all $A > 1$, $\delta > 0$. Taking $A \to 1$ and $\delta \downarrow 0$ we obtain that

$$\liminf_{t\downarrow 0}\left(\frac{Kx}{y}\right)^{1/\alpha}\frac{g_1(t)}{g_2(t)} \ge 1.$$

Analogously, we can show that

$$\limsup_{t\downarrow 0}\left(\frac{Kx}{y}\right)^{1/\alpha}\frac{g_1(t)}{g_2(t)} \le 1.$$

so that the limit of the left-hand side is 1. This implies the result, again by (9).

Proof (of Theorem 1). Since $\bar{F}_i(u) \to 0$ and $|\bar{F}_i(u) - f_i(u)| = o(\bar{F}_i(u))$, $i = 1, 2$, as $u \to \infty$, then for any small $\varepsilon > 0$ we can choose t_1 small enough so that for any $t \le t_1$ and $u > 0$ that satisfy $\bar{F}_1(u) \le tx$ we also have $|\bar{F}_1(u) - f_1(u)| \le \varepsilon|\bar{F}_1(u)|$, and hence $|\bar{F}_1(u) - f_1(u)| \le \varepsilon tx$. Moreover, we can choose $t_2 \le t_1$ small enough such that $\bar{F}_2(u) \le ty$ implies $|\bar{F}_2(u) - f_2(u)| \le \varepsilon ty$ for all $t \le t_2$. Also, for any small $\delta > 0$ it follows from Lemma 3 that there exists a positive number $t_3 \le t_2$ such that for all $t \le t_3$,

$$1 - \delta \le \frac{L(f_1^{-1}((1+\varepsilon)tx))}{L(f_2^{-1}((1+\varepsilon)ty))} \le 1 + \delta.$$

If we now fix some small $\varepsilon > 0$ and $\delta > 0$, the above implies for any $t \le t_3$ that

$$\begin{aligned}
&\mathbb{P}(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty) \\
&\quad = \mathbb{P}(f_1(N) \le (f_1(N) - \bar{F}_1(N)) + tx, f_2(R) \le (f_2(R) - \bar{F}_2(R)) + ty) \\
&\quad \le \mathbb{P}(f_1(N) \le (1+\varepsilon)tx, f_2(R) \le (1+\varepsilon)ty) \\
&\quad = \mathbb{P}\left(N \ge f_1^{-1}((1+\varepsilon)tx), R \ge f_2^{-1}((1+\varepsilon)ty)\right) \\
&\quad = \mathbb{P}\left(N \ge f_1^{-1}((1+\varepsilon)tx), R \ge \left(\frac{y}{Kx}\frac{L(f_1^{-1}((1+\varepsilon)tx))}{L(f_2^{-1}((1+\varepsilon)ty))}\right)^{-1/\alpha} f_1^{-1}((1+\varepsilon)tx)\right) \\
&\quad \le \mathbb{P}\left(N \ge f_1^{-1}((1+\varepsilon)tx), R \ge \left(\frac{y}{Kx}(1+\delta)\right)^{-1/\alpha} f_1^{-1}((1+\varepsilon)tx)\right).
\end{aligned}$$

Note that the above closely follows the derivation in (8), with $\sim$ signs replaced by inequalities; in particular the 5th line follow immediately from (9) upon replacing t by $(1+\varepsilon)t$. Noting that $f_1(f_1^{-1}((1+\varepsilon)tx)) = (1+\varepsilon)tx$, we can now apply Lemma 2 to the above and then let $t \to 0$, to obtain

$$\limsup_{t\to 0} t^{-1}\ \mathbb{P}(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty)\ \le\ (1+\varepsilon)\min\{x, (1+\delta)y(\mathbb{E}A)^\alpha/K\}.$$

Similarly we can obtain

$$\liminf_{t\to 0} t^{-1}\ \mathbb{P}(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty)\ \ge\ (1-\varepsilon)\min\{x, (1-\delta)y(\mathbb{E}A)^\alpha/K\},$$

so that the statement of the theorem follows by letting ε and δ go to 0.