The Distribution of PageRank Follows a Power-Law only for Particular Values of the Damping Factor

Luca Becchetti
Università di Roma
"La Sapienza"
Rome, Italy
becchett@dis.uniroma1.it

Carlos Castillo
Università di Roma
"La Sapienza"
Rome, Italy
castillo@dis.uniroma1.it

ABSTRACT

We show that the empirical distribution of the PageRank values in a large set of Web pages does not follow a power-law except for some particular choices of the damping factor. We argue that for a graph with an in-degree distribution following a power-law with exponent between 2.1 and 2.2, choosing a damping factor around 0.85 for PageRank yields a power-law distribution of its values. We suggest that power-law distributions of PageRank in Web graphs have been observed because the typical damping factor used in practice is between 0.85 and 0.90.

Categories and Subject Descriptors: H.4.m [Information Systems]: Miscellaneous

General Terms: Measurement

Keywords: PageRank distribution, Web graph

1. DISTRIBUTION OF PAGERANK VALUES

PageRank [5] is a link-based ranking function with a recursive definition: a page with a high PageRank is cited by many pages with high PageRank. More precisely, let $\mathbf{A}_{n\times n}$ be the link matrix of a Web graph of n pages, such that $a_{i,j} = 1$ iff there is a hyperlink between pages i and j in the Web graph. Let $\mathbf{P}$ be the row-normalized version of this matrix, such that $\sum_{j=1..n} p_{i,j} = 1 \quad \forall i$. The PageRank vector $\mathbf{r}(\alpha)$ is defined as the stationary distribution of the matrix $\alpha\mathbf{P} + \frac{1-\alpha}{n}\mathbf{1}_{n\times n}$.

We calculated the PageRank over a graph with 1 million nodes and 22 million edges; this graph was obtained from a crawl restricted to the .GR domain [1], and corresponds only to the main strongly connected component of that graph. The reason why we are using only the strongly connected component is that we want to compute the PageRank distribution also for high values of α, up to 0.99.

In Figure 1 we show the PageRank distribution for varying values of the damping factor, using a complementary cumulative distribution function (1 - C.D.F.) plot. The tail, starting around the top 5%-10% of the nodes, always follows a power-law with the same exponent no matter which damping factor we choose. The distribution for the remaining 90%-95% of the data varies with the damping factor.

We tried fitting a power-law distribution to the distribution of PageRank using the Hill estimator [2]; in this data set with $\alpha \approx 0.7 - 0.8$ there seems to be a power-law distribution over the entire range, but in other cases the power-law only fits the tail.

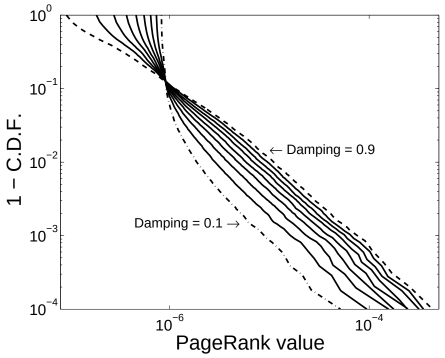

Figure 1: Distribution of PageRank for varying values of the damping factor (Damping=$0.1, 0.2, \ldots, 0.9$).

Next we consider a Double Pareto distribution[4]. The Double Pareto distribution looks like a bi-lineal in a log-log plot, and can be seen as similar to a log-normal. If we assume this model, we can fit a power-law separately to the body and the tail of the distribution, as shown in Figure 2. We confirmed that there is a varying slope in the body of the distribution and a fixed slope in the tail. The tail corresponds to highly-ranked pages and begins roughly at the intersection point in Figure 1, which corresponds to PageRank= $1/n$.

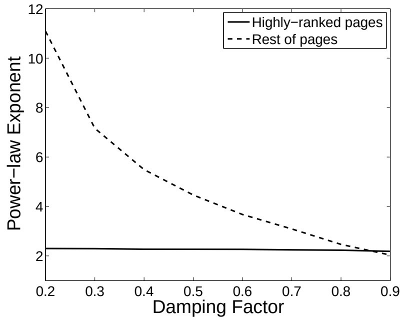

Figure 2: Exponent of two fits using power-laws to different parts of the distribution of PageRank, under varying damping factors.

Copyright is held by the author/owner.
WWW 2006, May 22–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

The tail of the distribution appears to follow a constant exponent, no matter which damping factor is used, while for most of the pages the distribution depends on the damping factor. For some particular values of the damping factor, both exponents are close to each other and the distribution looks like a power-law over the entire range. The exponent in the tail is about 2.2, which is the same exponent as the power-law for the in-degree in this dataset. This coincidence between the distribution of in-degree and the tail of PageRank was also observed in [6].

2. A DAMPING FACTOR TO GET A SINGLE POWER-LAW

In this section we assume the Double pareto model with intersection around $1/n$ and show that this hypothesis provides a reasonable explanation to the observation of a power-law distribution for the PageRank in many cases. Observe first that at every iteration of the computation, every page is given a baseline probability $x_{\min}$ given by

$$x_{\min} = \frac{1-\alpha}{n} \ .$$

If for all points greater than x_{min} a distribution follows a power-law with exponent θ, its probability density function $p(x)$ and cumulative distribution function $F(x)$ are [4]

$$p(x) = \frac{\theta - 1}{x_{min}^{-\theta+1} - 1} x^{-\theta} \qquad F(x) = \frac{x^{-\theta+1} - x_{min}^{-\theta+1}}{1 - x_{min}^{-\theta+1}}$$

where we have imposed the constraint that this particular distribution has a maximum possible value of 1. Following figure 1, we now compute $1 - F(1/n)$. Straightforward computations show that, if $n \gg 1$, as is in all cases of practical interest, $1 - F(1/n) \simeq (1-\alpha)^{\theta-1}$. The meaning of this result is that, if we assume that a single Pareto distribution approximates the PageRank distribution over the entire range of values, then $1 - F(1/n)$ does not depend on n. If we adopt the hypothesis of the Double Pareto distribution, this is what should happen for the particular value of α such that the distributions that approximate the body and the tail have the same exponent.

In our case, if we accept the hypothesis in [6], that is, the power-law exponent for the distribution of the tail of the PageRank values is the same as for the in-degree of pages (as has been observed experimentally in most Web characterization studies), then in our collection of 1 million nodes with $\theta = 2.2$ and $\alpha = 0.85$, the value predicted is $1 - F(1/n) \simeq 0.10$ and the observed 0.12. In the WebBase collection of 130 million documents [3] with $\theta = 2.07$ and $\alpha = 0.85$ the predicted is $1 - F(1/n) \simeq 0.13$ and the observed is 0.16. The value of $1 - F(1/n)$ in fact, does not seem to depend on n. Finally, also the `*.brown.edu` example considered in [6] shows a change of behaviour around $1/n$ (slightly more than 10^5 in their case), although they provide the raw histogram and not the cumulative distribution.

Having a single power-law over the entire range could be useful if we want to combine the PageRank values with other scoring functions, as in that case log(PageRank) has a uniform distribution.

3. LOGNORMAL PLUS BASELINE MODEL

Next we consider the power-law distribution as the sum of two random variables; let X be a random variable distributed according to a log-normal with parameters μ and σ. Let $F(x)$ be the cumulative density function (CDF) of X. We now consider

$$Y = \alpha X + \frac{1-\alpha}{n}$$

as a random variable which should have a distribution similar to the one of PageRank. The CDF of this distribution is:

$$P(Y \leq y) = F\left(\frac{y - \frac{1-\alpha}{n}}{\alpha}\right) \ .$$

For obtaining the μ and σ parameters, we fitted this distribution to the distribution of PageRank with $\alpha = 0.99$ using least-squares method. The result is shown in Figure 3. The obtained parameters are $\mu = -16.90$, $\sigma = 2.45$. We have observed that with the same μ and σ, our model fits PageRank also for all the other values of $\alpha = 0.1, 0.2, \ldots, 0.9$ with an average relative error of less than 1.5% over the entire range of PageRank values.

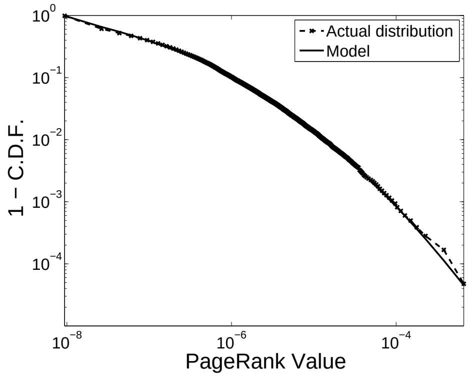

Figure 3: Fit of a log-normal plus a baseline distribution to the PageRank with $\alpha = 0.99$.

4. REFERENCES

[1] E. Efthimiadis and C. Castillo. Charting the Greek Web. In *Proc. of ASIST*, Providence, Rhode Island, USA, Nov. 2004.

[2] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1174, 1975.

[3] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. Webbase: a repository of web pages. *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):277–293, 2000.

[4] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, December 2005.

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

[6] G. Pandurangan, P. Raghavan, and E. Upfal. Using Pagerank to characterize Web structure. In *Proc. of COCOON*, vol. 2387 of *Lecture Notes in Computer Science*, pages 330–390, Singapore, Aug. 2002.