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

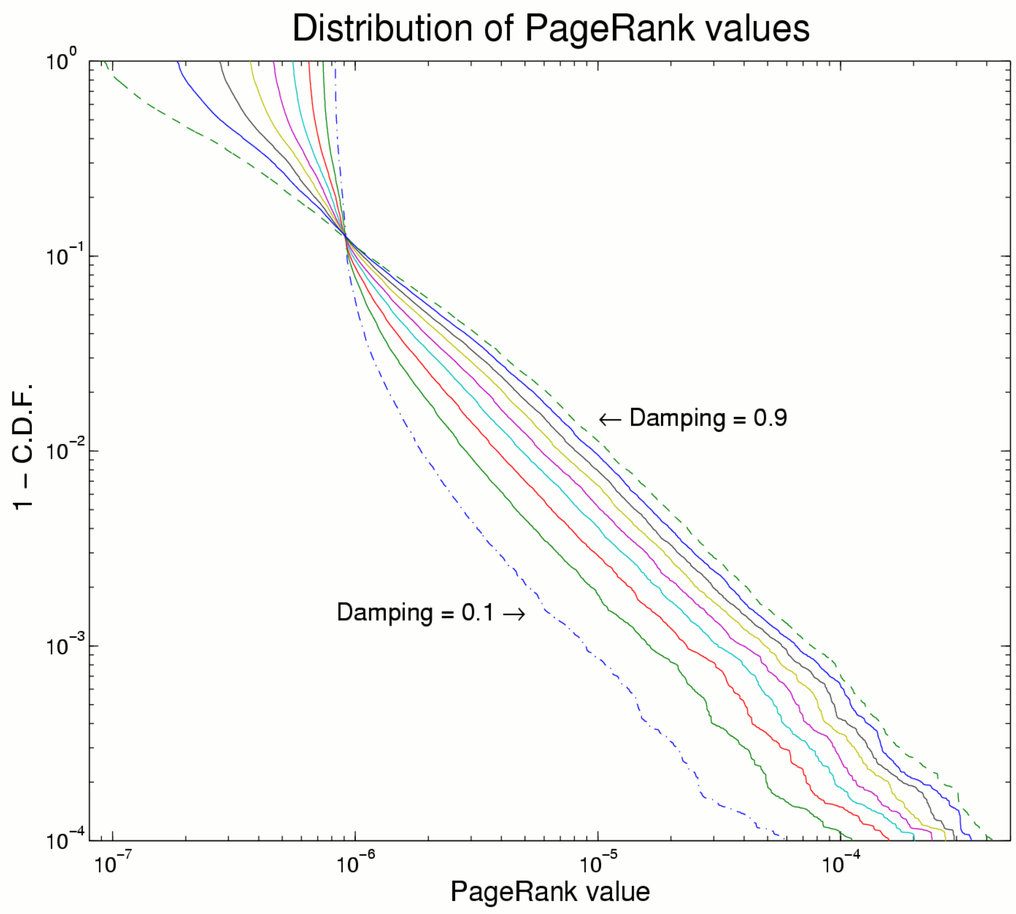
Luca Becchetti
Università di Roma "La Sapienza"
Rome, Italy
becchett@dis.uniroma1.it

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

The empirical distribution of PageRank in a large sample of Web pages does not follow a power-law except for **particular choices** of the damping factor.

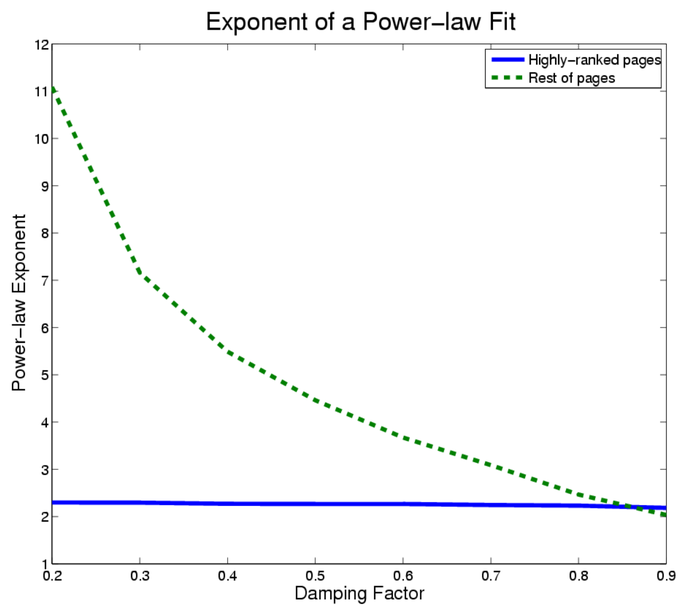
The tail, comprising 5%-10% of the nodes, always follows a power law, but the distribution for the remaining 90%-95% of pages varies.

This was observed in several Web samples having from 1 to 50 million pages with damping factors from 0.1 to 0.9 and the resulting behavior was very similar, specially if we restrict the sample to the main strongly connected component.



Double-Pareto Model

As in [Mitzenmacher 2003], we can fit the power-law to the body and the tail of the distribution separately, as in the figure on the right.



Given that the minimum value of PageRank is the baseline probability $(1\text{-}\alpha / N)$ if we assume the following:

(i) The exponent of the body is equal to the exponent of the tail
(ii)The distribution is normalized to 1, as in the case of PageRank
(iii) $N >> 1$

... we can show that the intersection point of the first figure, $1\text{-}F(1/N)$ does not depend on N:

$$1\text{-}F(1/N) \simeq (1\text{-}\alpha)^{\theta\text{-}1}$$

If we accept the hypothesis in [Pandurangan et al. 2002], 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, then:

- In our collection of 1 million nodes with θ=2.2 and α=0.85, the value predicted for 1-F(1/N) is 0.10 and the observed 0.12
- In the WebBase collection of 130 million documents with θ=2.07 and α=0.85 the predicted value is 0.13, and the 0.16

Conjecture

In a graph with power-law exponent θ for the indegree, calculating PageRank with:

$$\alpha = \frac{1}{\theta - 1}$$

Yields a power-law over the entire range of values

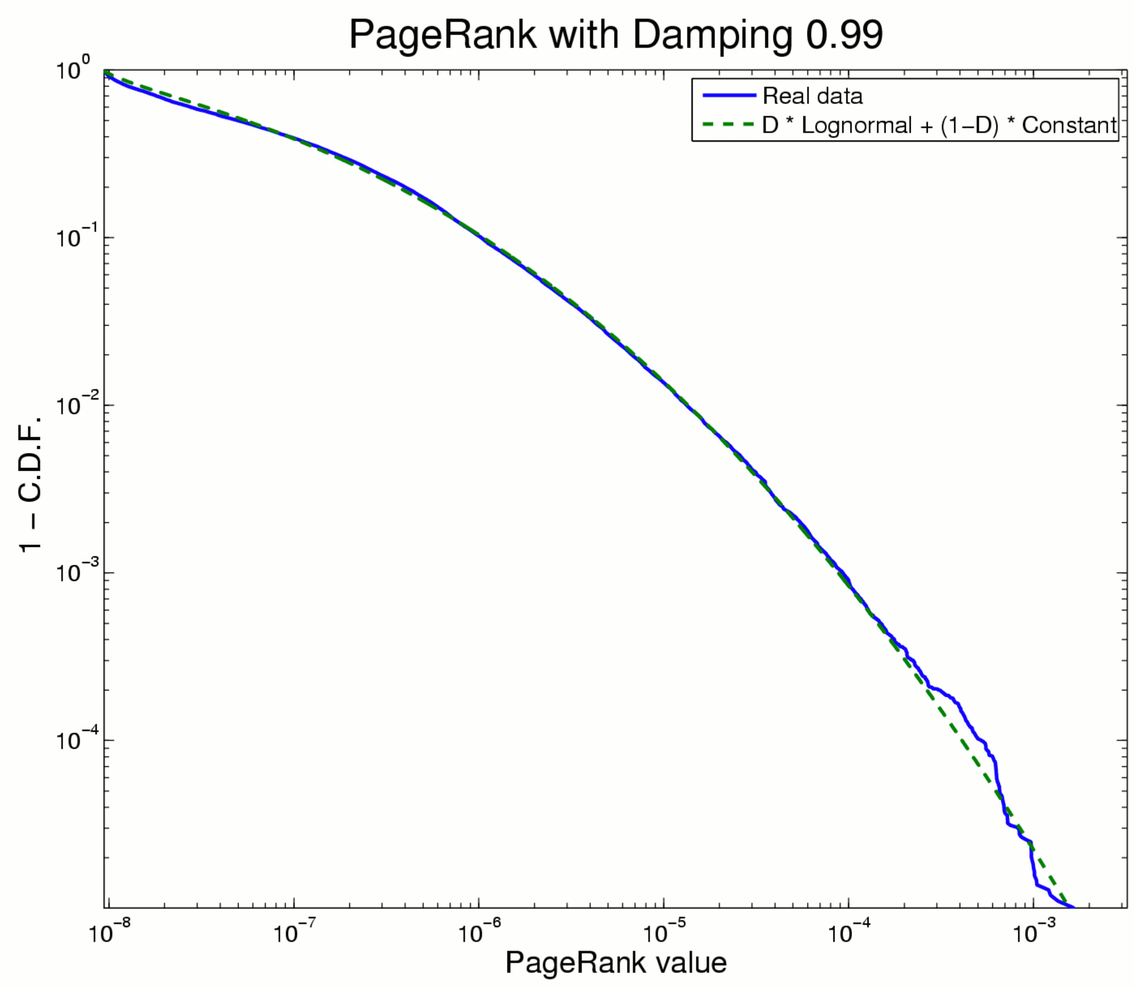
Examples: θ=2.1, then α=0.90 yields a power-law
θ=2.2, then α=0.83 yields a power-law

Lognormal + Baseline Model

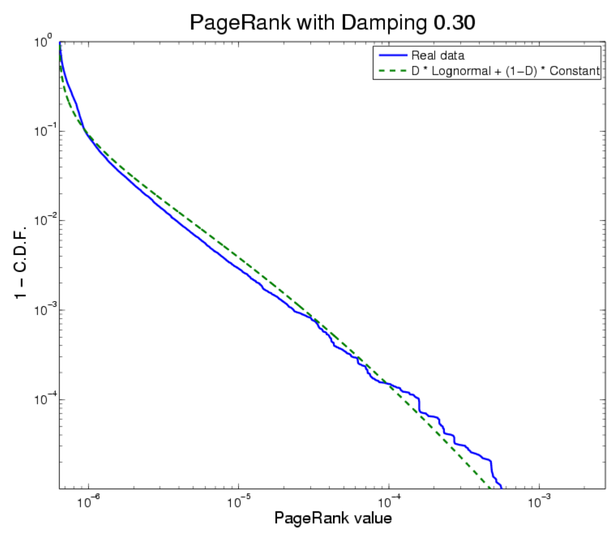
Let X be a random variable distributed according to a **lognormal distribution**. We propose the following model for the PageRank distribution:

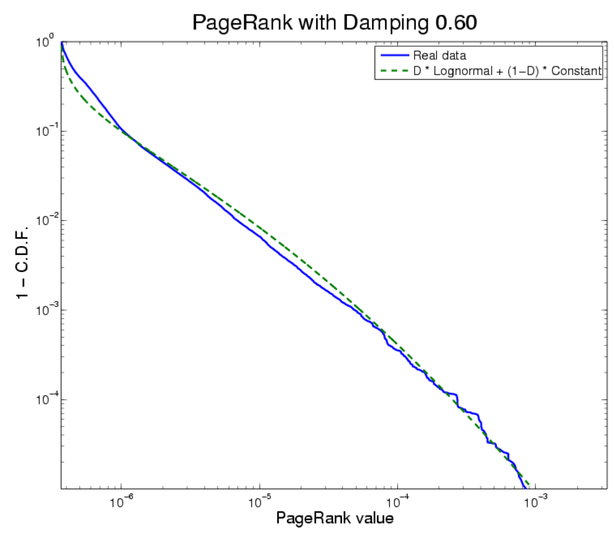
$$X + (1-\alpha)/N$$

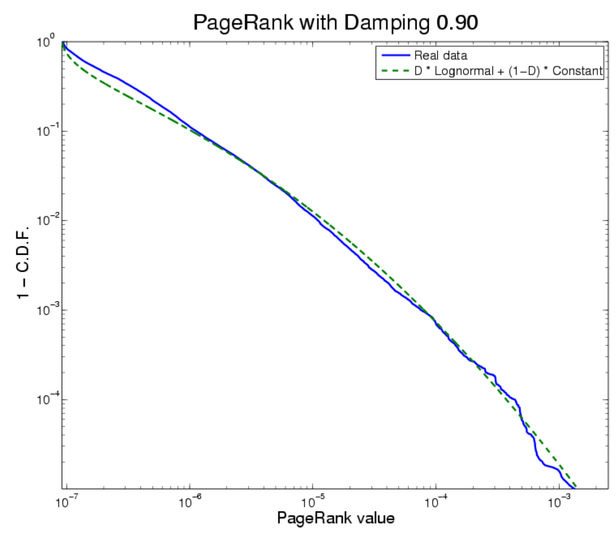
To obtain the lognormal parameters, we fit this to the distribution of PageRank with $\alpha=0.99$. Then **using the same parameters**, we can fit the PageRank distribution obtained with other damping factors with high precision.



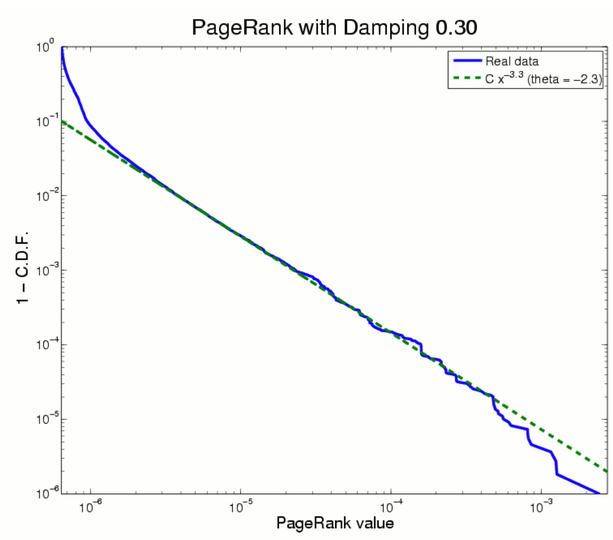
Fit with our model:

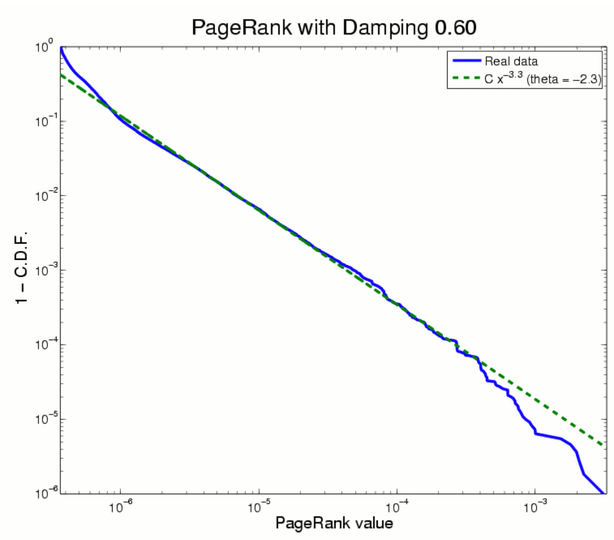


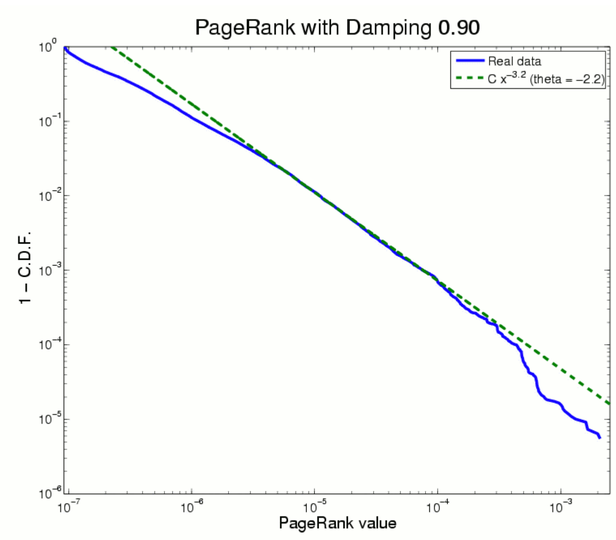




Fit with a power-law:







Conclusions

The PageRank distribution has been reported as a power-law because of the damping factor typically used (0.85-0.90)

Tuning the damping factor carefully to have 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.

A lognormal distribution plus a baseline probability, given by the random jumps, is a good approximation of the distribution of PageRank in the general case.

References:

M. Mitzenmacher: Dynamic models for file sizes and double Pareto distributions. *Internet Mathematics*, 1(3): 305-333, 2003.

G. Pandurangan, P. Raghavan and E. Upfal: Using PageRank to characterize Web structure.
In *Proc. of 6th COCOON*, vol. 2387 of *Lecture Notes in Computer Science*, pp. 330-390, Singapore, August 2002. Springer.