WEB GRAPH PARAMETERS AND THE PAGERANK DISTRIBUTION

YANA VOLKOVICH, NELLY LITVAK, AND DEBORA DONATO

Originally created for Web ranking, *PageRank* has become a major method for evaluating popularity of nodes in information networks. Besides its primary application in search engines, PageRank is successfully used for solving other important problems such as graph partitioning [3], spam detection [8], and finding gems in scientific citations [6], just to name a few. The PageRank [5] is defined as a stationary distribution of a random walk on the Web graph. At each step, with probability c, the random walk follows a randomly chosen outgoing link, and with probability $1-c$, the walk starts afresh from a page chosen at random according to some distribution f. Such random jump also occurs if a page is *dangling*, i.e. it does not have outgoing links. In the original definition, the teleportation distribution f is uniform over all Web pages. Then the PageRank values satisfy the equation

$$PR(i) = c\sum_{j\to i}\frac{1}{d_j}PR(j) + \frac{c}{n}\sum_{j\in\mathcal{D}}PR(j) + \frac{1-c}{n},\ i=1,\dots,n, \tag{1}$$

where $PR(i)$ is the PageRank of page i, d_j is the number of outgoing links of page j, the sum is taken over all pages j that link to page i, $\mathcal{D}$ is a set of dangling nodes, n is the number of pages in the Web, and c is the damping factor, which is a constant between 0 and 1.

Most experimental studies of the Web agree that *in-degree*, the number of incoming links of a page, and PageRank follow similar power laws with exponent $\alpha = 1.1$. It is clear from the definition (1) that the PageRank of a page depends on the popularity and the number of pages that link to it. Thus it could be expected that the distribution of PageRank should be related to the distribution of in-degree. It is also clear that PageRank is a *global* characteristic of the Web, which should depend on *out-degrees*, correlations, and other characteristics of the underlying graph. We study the influence of in-degrees, out-degrees and dangling nodes on the PageRank distribution. We model the relation between these variables through a stochastic equation inspired by the definition of PageRank (1). To this end, we view the PageRank of a random page as a random variable R with $\mathbb{E}(R) = 1$. We formally describe the concept of power law in terms of regular varying random variables [4]. Thus, we take a non-negative, integer and regularly varying random variable N for the in-degree of a random page. We consider a random variable D (*effective out-degree*), which represents the out-degree of a page that links to a particular randomly chosen page. We note that D is not the same random variable as the out-degree of a random page. Further, we assume that the fraction of the total PageRank mass concentrated in dangling nodes, equals the fraction of dangling nodes p_0. Then the PageRank R is a solution of the following stochastic equation:

$$R \stackrel{d}{=} c\sum_{j=1}^{N}\frac{1}{D_j}R_j + [1-c(1-p_0)]. \tag{2}$$

Here N, the R_j's and D_j's are independent; the R_j's are distributed as R, the D_j's are distributed as D. As before, $c \in (0,1)$ is the damping factor.

We also provide a recurrent stochastic model for the power iteration algorithm commonly used in PageRank computations. We start with initial distribution $R^{(0)}$, satisfying $\mathbb{E}\left(R^{(0)}\right) = 1$, and for every $k \geq 1$, we define the result of the kth iteration through the distributional identity

$$R^{(k)} \stackrel{d}{=} c\sum_{j=1}^{N}\frac{1}{D_j}R_j^{(k-1)} + [1-c(1-p_0)],$$

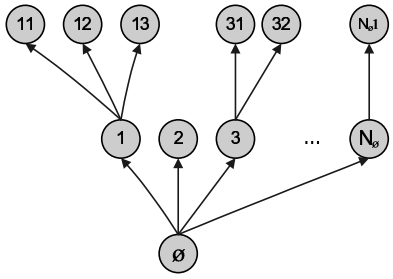


FIGURE 1. An example of Galton-Watson tree

where N, $R_j^{(k-1)}$ and D_j, $j \geq 1$, are independent.

These iterations lead to the distribution of PageRank, that we can explicitly quantify. To this end, we introduce the following notation. Let $\left\{\left(N_u, \frac{1}{D_{u_1}}, \frac{1}{D_{u_2}}, \dots\right)\right\}_u$ be a family of independent copies of $\left(N, \frac{1}{D_1}, \frac{1}{D_2}, \dots\right)$ indexed by all finite sequences $u = u_1 \dots u_n,\ u_i \in \{1, 2, \dots\}$. 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 $i \in \{1, 2, \dots\}$, then concatenation $ui \in \mathbb{T}$ if and only if $1 \leq i \leq N_u$. In other words, we indexed the nodes of the tree with root $\emptyset$ and the first level nodes $1, 2, ..N_\emptyset$, and at every subsequent level, the ith offspring of u is named ui (see Figure 1).

Equation (2) has a unique non-trivial solution with mean 1 given by

$$R = R^{(\infty)} = \lim_{k\to\infty} R^{(k)} = [1 - c(1 - p_0)] \sum_{n=0}^{\infty} c^n Y^{(n)}, \tag{3}$$

where

$$Y^{(n)} = \sum_{u=u_1\dots u_n \in \mathbb{T}} \frac{1}{D_{u_1}} \cdots \frac{1}{D_{u_1\dots u_n}},\ n \geq 1.$$

The random variable $Y^{(n)}$ represents the sum of the weights of the nth level of the Galton-Watson tree, where the root has weight 1, each edge has a random weight distributed as $1/D$, and the weight of a node is a product of weights of the edges, which are on the way from the root to this node.

We use recent results on regular variation [9] to obtain PageRank asymptotics.

Theorem 1. *If* $\mathbb{P}\left(R^{(0)} > x\right) = o(\mathbb{P}(N > x))$, *then for all* $k \geq 1$,

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

where $C_k = \left(\frac{c(1-p_0)}{d}\right)^\alpha \sum_{j=0}^{k-1} c^{j\alpha} b^j$, *and* $b = d\mathbb{E}\,(1/D^\alpha)$.

The form of the coefficient C_k arises from the proof [12]. For large enough k, C_k can be approximated by

$$C = \lim_{k\to\infty} C_k = \frac{c^\alpha (1 - p_0)^\alpha}{d^\alpha (1 - c^\alpha b)}.$$

It follows that $b \geq (1 - p_0)^\alpha d^{1-\alpha}$, and hence,

$$C \geq \frac{c^\alpha (1 - p_0)^\alpha}{d^\alpha (1 - c^\alpha (1 - p_0)^\alpha d^{1-\alpha})}. \tag{4}$$

The last expression is the value of C if out-degree of all non-dangling nodes is a constant [10]. Note that if $\alpha \approx 1.1$, then the difference between the left- and the right-hand sides of (4) is really small for any reasonable out-degree distribution.

Thus, we clearly show that the power law of in-degree remains a major factor shaping the PageRank distribution. The difference between the power laws is in the multiplicative constant C_k, which depends mainly on the fraction of dangling nodes, the average in-degree, the power law exponent, and the damping factor.

Our theoretical predictions also show a good agreement with experimental data on the different Web samples. We performed experiments on Indochina-2004 and EU-2005 Web samples collected by The Laboratory for Web Algorithmics (LAW), Dipartimento di Scienze dell'Informazione (DSI) of the Università degli studi di Milano [1]. We also used a Stanford-2002 Web sample [2]. In Figures 2–4 below we present cumulative log-log plots for in-degree/PageRank. The y-axis corresponds to the fraction of pages with in-degree/PageRank greater than the value on the x-axis. For in-degree, the power law exponent in evaluated using the maximum likelihood estimator from [11], and the straight line is fitted accordingly. For the PageRank, we plot the *theoretically predicted* straight lines obtained from Theorem 1.

FIGURE 2. Indochina data set: cumulative log-log plots for in-degree/PageRank. The straight lines for the PageRank plots are predicted by the model.

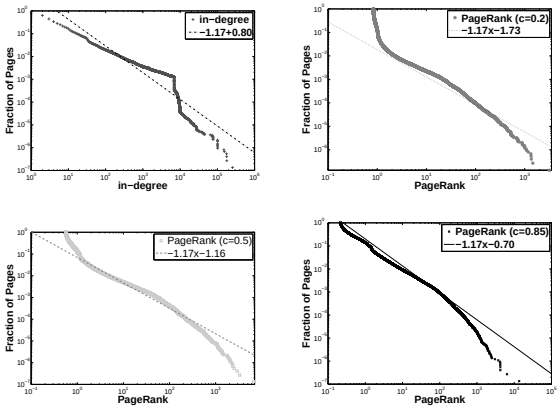
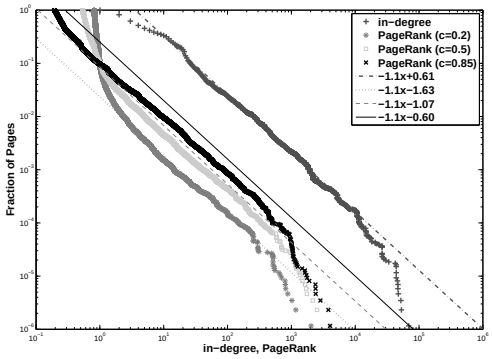
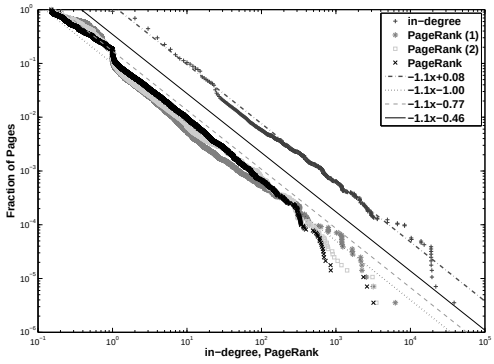


FIGURE 3. EU-2005 data set: cumulative log-log plots for in-degree/PageRank. The straight lines for the PageRank plots are predicted by the model.



We clearly see that our analytical stochastic model helps to predict the shape of the PageRank log-log plot on basis of in-degree distribution, the damping factor, and the fraction of dangling nodes. Experiments show that our theoretical model matches the Web data with a good accuracy. This result explains and quantifies the influence of various Web parameters on the PageRank distribution. This can be used, for instance, for refining the relation between global rank and PageRank [7]. However, to make our mathematical model analytically tractable, we had to allow for several simplifying assumptions, such as independence of certain parameters and uniform teleportation. This leads to inaccuracy of our predictions on the real data. In further research, we will gradually improve our model including dependencies, personalization, and other important factors relevant for the contemporary Web search.

FIGURE 4. Stanford data set: cumulative log-log plots for in-degree/PageRank. The straight lines for the PageRank plots are predicted by the model for the 1st, the 2nd, and the last power iterations



REFERENCES

[1] http://law.dsi.unimi.it/. Accessed in January 2007.
[2] http://www.stanford.edu/∼sdkamvar/research.html. Accessed in March 2006.
[3] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.
[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, 1989.
[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks ISDN Systems*, 33:107–117, 1998.
[6] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google. Technical Report 0604130, arxiv/physics/, 2006.
[7] S. Fortunato, M. Boguna, A. Flammini, and F.Menczer. How to make the top ten: Approximating PageRank from in-degree, 2005. arXiv.org/cs/cs.IR/0511016.
[8] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *30th International Conference on Very Large Data Bases*, page 576587, 2004.
[9] A. H. Jessen and T. Mikosch. Regularly varying functions. *Publications de L'Institut Mathematique, Nouvelle Série*, 79(93), 2006.
[10] N. Litvak, W. R. W. Scheinhardt, and Y. Volkovich. In-degree and PageRank: Why do they follow similar power laws? To appear in *Internet Math.*
[11] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2005.
[12] Y. Volkovich, N. Litvak, and D. Donato. Determining factors behind the PageRank log-log plot. Memorandum 1823, Enschede, February 2007.

(Y. Volkovich) FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE, UNIVERSITY OF TWENTE, 7500 AE ENSCHEDE, THE NETHERLANDS
E-mail address: Y.Volkovich@ewi.utwente.nl

(N. Litvak) FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE, UNIVERSITY OF TWENTE, 7500 AE ENSCHEDE, THE NETHERLANDS
E-mail address: N.Litvak@ewi.utwente.nl

(D. Donato) YAHOO! RESEARCH, BARCELONA OCATA 1, 1ST FLOOR, 08003 BARCELONA CATALUNYA, SPAIN
E-mail address: debora@yahoo-inc.com