Approximating PageRank from In-Degree

Santo Fortunato[1,2], Marián Boguñá[3],
Alessandro Flammini[1], and Filippo Menczer[1]

[1] School of Informatics, Indiana University
Bloomington, IN 47406, USA
[2] Complex Networks Lagrange Laboratory (CNLL),
ISI Foundation, Torino, Italy
[3] Departament de Física Fonamental, Universitat de Barcelona
08028 Barcelona, Spain

Abstract. PageRank is a key element in the success of search engines, allowing to rank the most important hits in the top screen of results. One key aspect that distinguishes PageRank from other prestige measures such as in-degree is its global nature. From the information provider perspective, this makes it difficult or impossible to predict how their pages will be ranked. Consequently a market has emerged for the optimization of search engine results. Here we study the accuracy with which PageRank can be approximated by in-degree, a local measure made freely available by search engines. Theoretical and empirical analyses lead to conclude that given the weak degree correlations in the Web link graph, the approximation can be relatively accurate, giving service and information providers an effective new marketing tool.

1 Introduction

PageRank has become a key element in the success of Web search engines, allowing to rank the most important hits in the top page of results. Certainly the introduction of PageRank as a factor in sorting results [1] has contributed considerably to Google's lasting dominance in the search engine market [2].

PageRank is not the only possible measure of importance or prestige among Web pages. The simplest possible way to measure the prestige of a page is to count the incoming links (in-links) to the page. There is a correlation between the number of in-links that a page receives from other pages (in-degree) and quality, especially when the in-degree is large. The in-degree of Web pages is very cheap to compute and maintain, so that a search engine can easily keep in-degree updated with the evolution of the Web.

However, in-degree is a local measure. All links to a page are considered equal, regardless of where they come from. Two pages with the same in-degree are considered equally important, even if one is cited by more prestigious sources than the other. To modulate the prestige of a page with that of the pages pointing to it means to move from the examination of an individual node in the link graph to that of the node together with its predecessor neighbors. PageRank represents

W. Aiello et al. (Eds.): WAW 2006, LNCS 4936, pp. 59–71, 2008.
© Springer-Verlag Berlin Heidelberg 2008

such a shift from the local measure given by in-degree toward a global measure where each Web page contributes to define the importance of every other page.

The use of PageRank in place of in-degree for applications such as ranking by Web search engines relies on two assumptions: (i) PageRank is a quantitatively different and better prestige measure compared to in-degree; and (ii) PageRank cannot be easily guessed or approximated by in-degree. To wit, Amento *et al.* [3] report a very high average correlation between in-degree and PageRank (Spearman $\rho = 0.93$, Kendall $\tau = 0.83$) based on five queries. Further, they report the same average precision at 10 (60%) based on relevance assessments by human subjects. In this paper we further quantitatively explore these assumptions answering the following questions: *What is the correlation between in-degree and PageRank across representative samples of the Web? How accurately can one approximate PageRank from local knowledge of in-degree?*

From the definition of PageRank, other things being equal, the PageRank of a page grows with the in-degree of the page. Beyond this zero-order approximation, the actual relation between PageRank and in-degree has not been thoroughly investigated in the past. It is known that the distributions of PageRank and in-degree follow an almost identical pattern [4,5], i.e., a curve ending with a broad tail that follows a power law with exponent $\gamma \simeq 2.1$. This fact may indicate a strong correlation between the two variables. Surprisingly there is no agreement in prior literature about the correlation between PageRank and in-degree. Pandurangan *et al.* [4] show very little correlation based on analysis of the `brown.edu` domain and the TREC WT10g collection. Donato *et al.* [5] report on a correlation coefficient which is basically zero based on analysis of a much larger sample ($2 \cdot 10^8$ pages) taken from the WebBase [6] collaboration. On the other hand, analysis of the University of Notre Dame domain by Nakamura [7] reveals a strong correlation.

In Section 2 we estimate PageRank for a generic directed network within a mean field approach. For a network without degree-degree correlations the average PageRank turns out to be simply proportional to the in-degree, modulo an additive constant. The prediction is validated empirically in Section 3, where we solve the equations numerically for four large samples of the Web graph; in each case the agreement between our theoretical estimate and the empirical data is excellent. We find that the Web graph is basically uncorrelated, so the average PageRank can be well approximated by a linear function of the in-degree. As an additional contribution we settle the issue of the correlation between PageRank and in-degree; the linear correlation coefficient is consistently large for all four samples we have examined, in agreement with Nakamura [7]. Finally, in Section 4, we present an application of our findings on the live Web.

2 Theoretical Analysis

The PageRank $p(i)$ of a page i is defined through the following expression [1]:

$$p(i) = \frac{q}{N} + (1-q) \sum_{j:j\to i} p(j)/k_{out}(j) \qquad i = 1, 2, \ldots, N \tag{1}$$

where N is the total number of pages, $j \to i$ indicates a hyperlink from j to i, $k_{out}(j)$ is the out-degree of page j and q is the so-called teleportation (or jumping) factor. The set of Equations 1 can be solved iteratively. From Eq. 1 it is clear that the PageRank of a page grows with the PageRank of the pages that point to it. However, the sum over predecessor neighbors implies that PageRank also increases with the in-degree of the page.

PageRank can be thought of as the stationary probability of a random walk process with additional random jumps. The physical description of the process is as follows: when a random walker is in a node of the network, at the next time step with probability q it jumps to a randomly chosen node and with probability $1 - q$ it moves to one of its successors with uniform probability. In the case of directed networks, a node may have no successors. In this case the walker jumps to a randomly chosen node of the network with probability one. The PageRank of a node i, $p(i)$, is then the probability to find the walker at node i when the process has reached the steady state, a condition that is always guaranteed by the teleportation probability q.

The probability to find the walker at node i at time step n follows a simple Markovian equation:

$$p_n(i) = \frac{q}{N} + (1-q) \sum_{j:k_{out}(j)\neq 0} \frac{a_{ji}}{k_{out}(j)} p_{n-1}(j) + \frac{1-q}{N} \sum_{j:k_{out}(j)=0} p_{n-1}(j), \quad (2)$$

where a_{ji} is the adjacency matrix with entry 1 if there is a direct connection between j and i and zero otherwise. The first term in Eq. 2 is the contribution of walkers jumping to a randomly chosen node, the second term is the random walk contribution, and the third term accounts for walkers that at the previous step were located in dangling nodes and now jump to random nodes. In the limit $n \to \infty$ this last contribution becomes a constant term affecting all the nodes in the same way, and thus it can be removed from Eq. 2 under the constraint that the final solution is properly normalized. Strictly speaking this would lead to an effective teleportation term, which we omit to keep the notation simple. Alternatively dangling nodes could be taken into account by a proper rescaling of the the second term [8]. Hereafter we intend all sums over nodes to exclude dangling ends, considering only nodes with $k_{out} > 0$. The PageRank of page i is the steady state solution of Eq. 2, $p(i) = \lim_{n\to\infty} p_n(i)$. Equation 2 cannot be analytical solved. We propose a mean field solution of Eq. 2 that, nevertheless, gives a very accurate description of the PageRank structure of the Web. The mean field approach is often used in statistical physics, and is reliable when each element of the system has many interaction partners,[1] as in this case the effect of the interactions can be taken into account in an average way, neglecting the variations among the elements.

Instead of analyzing the PageRank of single pages, we aggregate pages in classes according to their degree $\mathbf{k} \equiv (k_{in}, k_{out})$ and define the average PageRank of nodes of degree class $\mathbf{k}$ as

[1] On hypercubic lattices, the mean field limit for most spin models is reached in four dimensions, when each spin has eight neighbors.

$$\overline{p}_n(\mathbf{k}) \equiv \frac{1}{NP(\mathbf{k})} \sum_{i \in \mathbf{k}} p_n(i). \tag{3}$$

Note that now "degree class $\mathbf{k}$" means all the nodes with in-degree k_{in} and out-degree k_{out}; $P(\mathbf{k})$ is the probability that a node is in the degree class $\mathbf{k}$. Taking the average of Eq. 2 for all nodes of the degree class $\mathbf{k}$ we obtain

$$\frac{1}{NP(\mathbf{k})} \sum_{i \in \mathbf{k}} p_n(i) = \frac{q}{N} + \frac{(1-q)}{NP(\mathbf{k})} \sum_{i \in \mathbf{k}} \sum_{j : k_{out}(j) \neq 0} \frac{a_{ji}}{k_{out}(j)} p_{n-1}(j). \tag{4}$$

From Eq. 3 we see that the left-hand side of Eq. 4 is $\overline{p}_n(\mathbf{k})$. In the right-hand side we split the sum over j into two sums, one over all the degree classes $\mathbf{k}'$ and the other over all the nodes within each degree class $\mathbf{k}'$. We get

$$\overline{p}_n(\mathbf{k}) = \frac{q}{N} + \frac{(1-q)}{NP(\mathbf{k})} \sum_{\mathbf{k}'} \frac{1}{k'_{out}} \sum_{i \in \mathbf{k}} \sum_{j \in \mathbf{k}'} a_{ji} p_{n-1}(j). \tag{5}$$

At this point we perform our mean field approximation [9], which consists in substituting the PageRank of the predecessor neighbors of node i by its mean value, that is,

$$\begin{aligned} \sum_{i \in \mathbf{k}} \sum_{j \in \mathbf{k}'} a_{ji} p_{n-1}(j) &\simeq \overline{p}_{n-1}(\mathbf{k}') \sum_{i \in \mathbf{k}} \sum_{j \in \mathbf{k}'} a_{ji} \\ &= \overline{p}_{n-1}(\mathbf{k}') E_{\mathbf{k}' \to \mathbf{k}}, \end{aligned} \tag{6}$$

where $E_{\mathbf{k}' \to \mathbf{k}}$ is the total number of links pointing from nodes of degree $\mathbf{k}'$ to nodes of degree $\mathbf{k}$. This matrix can also be rewritten as

$$\begin{aligned} E_{\mathbf{k}' \to \mathbf{k}} &= k_{in} P(\mathbf{k}) N \frac{E_{\mathbf{k}' \to \mathbf{k}}}{k_{in} P(\mathbf{k}) N} \\ &= k_{in} P(\mathbf{k}) N P_{in}(\mathbf{k}'|\mathbf{k}), \end{aligned} \tag{7}$$

where $P_{in}(\mathbf{k}'|\mathbf{k})$ is the probability that a predecessor of a node belonging to degree class $\mathbf{k}$ belongs to degree class $\mathbf{k}'$. The conditional probability $P_{in}(\mathbf{k}'|\mathbf{k})$ incorporates the so-called *degree-degree correlation*, i.e., the correlation between the degree of a node and that of its neighbors (see [10] pp. 243–245). Using Equations 6 and 7 in Eq. 5 we finally obtain

$$\overline{p}_n(\mathbf{k}) = \frac{q}{N} + (1-q) k_{in} \sum_{\mathbf{k}'} \frac{P_{in}(\mathbf{k}'|\mathbf{k})}{k'_{out}} \overline{p}_{n-1}(\mathbf{k}'), \tag{8}$$

which is a closed set of equations for the average PageRank of pages in the same degree class. When the network has degree-degree correlations, the solution of this equation is non-trivial and the resulting PageRank can have a complex dependence on the degree. However, in the particular case of networks

without degree-degree correlations, the transition probability $P_{in}(\mathbf{k}'|\mathbf{k})$ becomes independent of $\mathbf{k}$ and takes the simpler form

$$P_{in}(\mathbf{k}'|\mathbf{k}) = \frac{k'_{out}P(\mathbf{k}')}{\langle k_{in}\rangle}, \tag{9}$$

where $\langle\cdot\rangle$ denotes the average value of the quantity in brackets. Using this expression in Eq. (8) and taking the limit $n \to \infty$, we obtain

$$\overline{p}(\mathbf{k}) = \frac{q}{N} + \frac{1-q}{N}\frac{k_{in}}{\langle k_{in}\rangle}, \tag{10}$$

that is, the average PageRank of nodes of degree class $\mathbf{k}$ is independent of k_{out} and proportional to k_{in}.

The same type of analysis allows to estimate the size of the fluctuations of PageRank for nodes in the same degree class $\mathbf{k}$. It turns out that, for uncorrelated networks, the standard deviation $\sigma(\mathbf{k})$ of the PageRank distribution about its mean value is

$$\sigma^2(\mathbf{k}) \simeq \frac{(1-q)^4}{N^2\langle k_{in}\rangle^3}\left\langle\frac{k_{in}^2}{k_{out}}\right\rangle k_{in}. \tag{11}$$

For large in-degrees, the coefficient of variation is

$$\frac{\sigma(\mathbf{k})}{\overline{p}(\mathbf{k})} \simeq (1-q)\left[\left\langle\frac{k_{in}^2}{k_{out}}\right\rangle\frac{1}{\langle k_{in}\rangle k_{in}}\right]^{1/2}. \tag{12}$$

The factor $\left\langle\frac{k_{in}^2}{k_{out}}\right\rangle$ in this expression can be very large when the network has a long-tailed degree distribution, which implies that the relative fluctuations are large for small in-degrees. Therefore the true PageRank of pages with small in-degree may differ significantly from its mean field approximation. However, for large in-degrees the relative fluctuations become less important — due to the factor k_{in} in the denominator — and the average PageRank from Eq. 10 gives a good approximation. Note that the expression in Eq. 12 relates to the relative fluctuations *within* a degree class, rather than across the entire graph. Since PageRank is distributed according to a power law with γ close to 2, the overall fluctuations diverge in the limit of infinite graph size. An analysis of the PageRank distribution and of the relative fluctuations within each degree class is omitted here for brevity, and will be included in an extended version of this paper.

3 Results

For an empirical validation of the theoretical predictions in the previous section, we analyzed four samples of the Web graph. Two of them were obtained by crawls performed in 2001 and 2003 by the WebBase collaboration [6]. The other two were collected by the WebGraph project [11]: the pages belong to two national

Table 1. Number of pages, links, and average degree ($\langle k \rangle = \langle k_{in} \rangle = \langle k_{out} \rangle$) for the four data sets we have analyzed

Data set	WB 2001	.uk 2002	WB 2003	.it 2004
# pages	8.1×10^7	1.9×10^7	4.9×10^7	4.1×10^7
# links	7.5×10^8	2.9×10^8	1.2×10^9	1.1×10^9
$\langle k \rangle$	9.34	15.78	24.05	27.50

Table 2. Exponents of the power law part of the PageRank distribution and linear correlation coefficients between PageRank and in-degree

Data set	WB 2001	.uk 2002	WB 2003	.it 2004
γ	2.2 ± 0.1	2.0 ± 0.1	2.0 ± 0.1	2.0 ± 0.1
ρ	0.538	0.554	0.483	0.733

domains, .uk (2002) and .it (2004), respectively. In Table 1 we list the number of vertices and edges and the average degree for each data set.

We calculated PageRank with the standard iterative procedure; the factor q was set to 0.15, as in the original paper by Brin and Page [1] and many successive studies. In Fig. 1 we show the cumulative distributions of PageRank, i.e. the function $R(p)$ representing the probability that PageRank exceeds the value p. Using the cumulative distribution allows to reduce the noise due to fluctuations at large PageRank values. In all four cases we obtained a pattern with a broad tail. The initial part of the distribution can be well fitted by a power law $p^{-\beta}$ with exponent β between 1.0 and 1.2. The exponents for the actual PageRank distribution are $\gamma = \beta + 1$, so they range from 2.0 to 2.2, in agreement with other studies [4,5]. The right-most part of each curve, corresponding to the pages with highest PageRank, decreases faster. For the WebBase sample of 2001 the tail of the curve up to the last point can be well fitted by a power law with exponent $\beta \approx 1.6$; in the other cases we see evidence of an exponential cutoff.

We also calculated the linear correlation coefficient between PageRank and in-degree. In Table 2 we list Pearson's ρ together with the slope of the power law portions of the PageRank distributions. The correlation between PageRank and in-degree is rather strong, in contrast to the findings of [4] and especially [5] but in agreement with [7] and consistently with the high correlation observed between in-degree and Kleinberg's authority score [12].

Let us now validate the expression derived from our mean field analysis for the average PageRank. We solved Eq. 8 with an analogous iterative procedure as the one we used to calculate PageRank. We now look for the vector $\bar{p}(\mathbf{k})$, defined for all pairs $\mathbf{k} \equiv (k_{in}, k_{out})$ which occur in the network. Since PageRank is a probability, it must be normalized so that its sum over all vertices of the network is one. So we initialized the vector with the constant $\bar{p}_0(\mathbf{k}) = 1/N$, and plugged it into the right-hand side of Eq. 8 to get the first approximation $\bar{p}_1(\mathbf{k})$. We then used $\bar{p}_1(\mathbf{k})$ as input to get $\bar{p}_2(\mathbf{k})$, and so on. We remark that the expression of the probability $P_{in}(\mathbf{k}'|\mathbf{k})$ is not a necessary ingredient of the

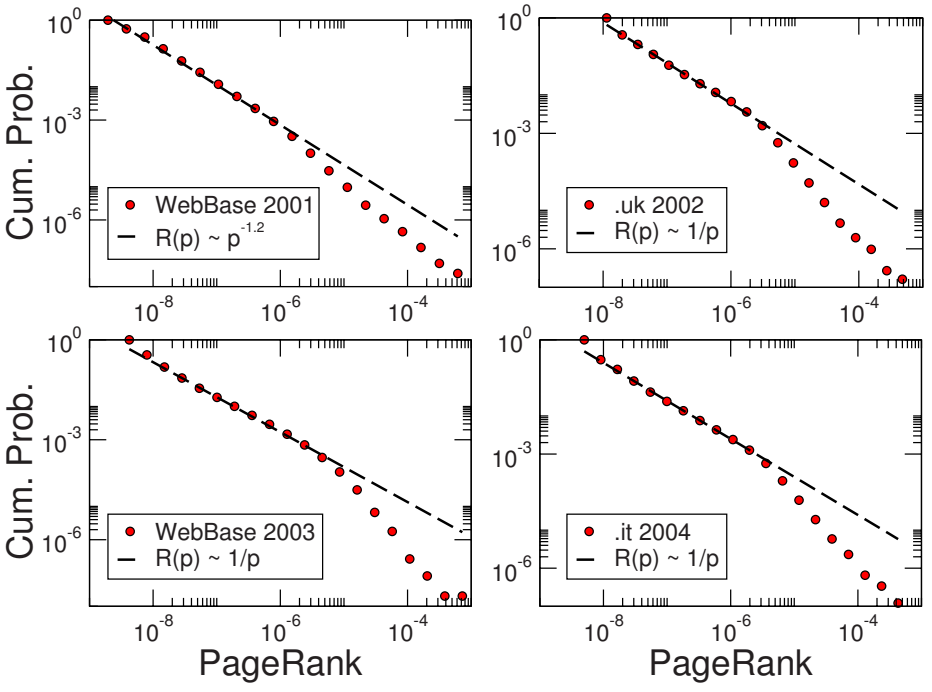

Fig. 1. Cumulative distributions of PageRank

calculation. In fact, the sum on the right-hand side of Eq. 8 is just the average value of $\bar{p}_{n-1}(\mathbf{k}')/k'_{out}$ among all predecessors of vertices with degree $\mathbf{k}$. The algorithm leads to convergence within a few iterations (we never needed more than 20). In Fig. 2 we compare the values of $\bar{p}(\mathbf{k})$ calculated from Eq. 8 with the corresponding empirical values. Here we averaged $\bar{p}(\mathbf{k})$ over out-degree, so it only depends on the in-degree k_{in}. The variation of $\bar{p}(\mathbf{k})$ with k_{out} (for fixed k_{in}) turns out to be very small. The scatter plots of Fig. 2 show that the mean field approximation gives excellent results: the points are very tightly concentrated about each frame bisector, drawn as a guide to the eye.

Next let us analyze explicitly the relation between PageRank and in-degree. To plot the function $\bar{p}(k_{in})$ directly is not very helpful because the wide fluctuations of PageRank within each degree class would mystify the pattern for large values of k_{in}. So we average PageRank within bins of in-degree, which is the standard procedure to derive trends from scatter plots (see [10] pp. 240–242). As both PageRank and in-degree are power-law distributed, we use logarithmic bins; the multiplicative factor for the bin size is 1.3. The resulting patterns for our four Web samples are presented in Fig. 3. The empirical curves are rather smooth, and show that the average PageRank (per degree class) is an increasing function of in-degree. The relation between the two variables is approximately linear

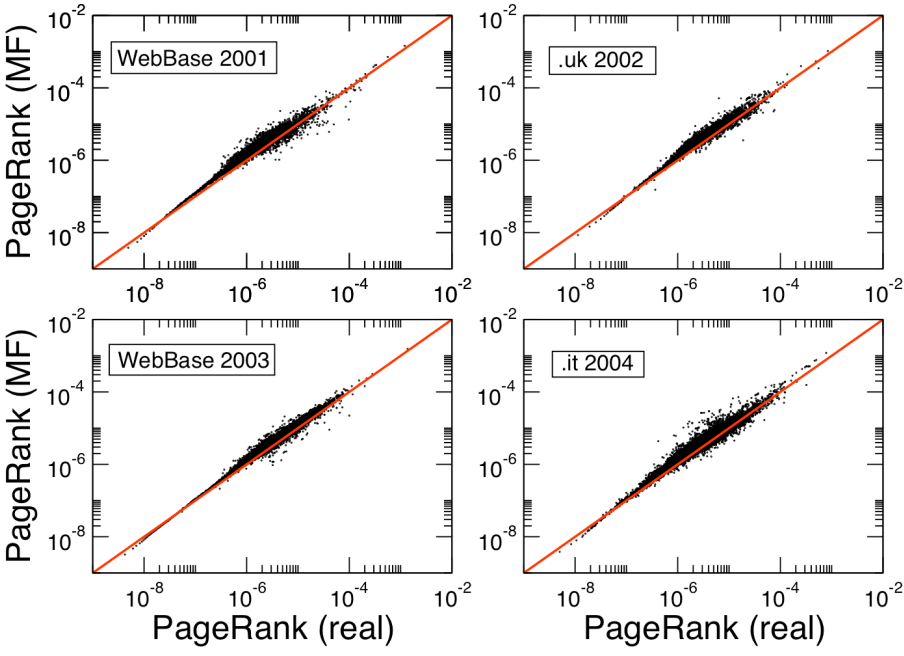

Fig. 2. Scatter plots of the empirical average PageRank per degree class versus our mean field (MF) estimate

for large in-degrees. This is exactly what we would expect if the degrees of pages were uncorrelated with those of their neighbors in the Web graph (cf. Section 2). In such a case the relation between PageRank and in-degree is given by Eq. 10. Indeed, the comparison of the empirical data with the curves of Eq. 10 in Fig. 3 is quite good for all data sets. We infer that the Web graph is an essentially uncorrelated graph; this is confirmed by direct measurements of degree-degree correlations in our four Web samples [13]. What is most important, the average PageRank of a page with in-degree k_{in} is well approximated by the simple expression of Eq. 10.

4 Applications to the Live Web

Knowing the relationship between PageRank and in-degree has potential applications for the Web graph. It is vital for many service and information providers to have good rankings by major search engines for relevant keywords, given that search engines are the primary way that Internet users find and visit Web sites [14,15]. Consequently a demand has emerged for companies that perform so-called *search engine optimization* or *search engine marketing* on behalf of business clients. The goal is to increase the rankings of their pages, thus directing traffic to their sites [16].

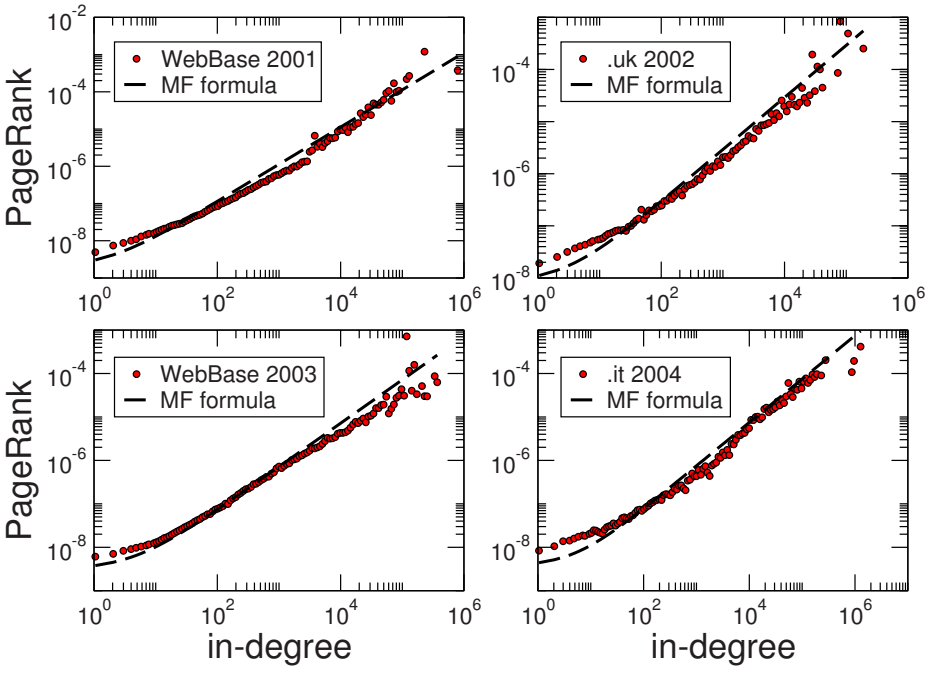

Fig. 3. PageRank versus in-degree; the dashed line is the approximation given by the closed formula of Eq. 10

In the previous section we have shown that the average PageRank of a page with in-degree k_{in} can be well approximated by the closed formula in Eq. 10. So Web authors may use local in-degree information as a proxy for estimating the global PageRank of their sites.

To use Eq. 10 for the Web we need to know the total number N of Web pages indexed by a search engine, say Google, and their average degree $\langle k_{in} \rangle$. The size of the Google index was published until recently; we use the last reported number, $N \simeq 8.1 \times 10^9$. The average degree is not known; the best we can do is extract it from samples of the Web graph. Our data sets do not deliver a unique value for $\langle k_{in} \rangle$, but they agree on the order of magnitude (see Table 1). Hereafter we use $\langle k_{in} \rangle = 10$.

Let us now consider whether Eq. 10 can be useful in the live Web. Ideally we should compare the PageRank values of a list of Web pages with the corresponding values derived through our formula. Unfortunately the real PageRank values calculated by a search engine such as Google are not accessible, so we need a different strategy. The simplest choice is to focus on rank rather than PageRank. We know that Google ranks Web pages according to their PageRank values as well as other features which do not depend on Web topology. The latter features are not disclosed; in the following we disregard them and assume for simplicity that the ranking of a Web page exclusively depends on its PageRank value.

There is a simple relation between the PageRank p of a Web page and the rank R of that page. The Zipf function $R(p)$ is simply proportional to the cumulative distribution of PageRank. Since the PageRank distribution is approximately a power law with exponent $\gamma \simeq 2.1$ (see Section 3), we find that

$$R(p) \simeq A\, p^{-\beta}, \tag{13}$$

where $\beta = \gamma - 1 \simeq 1.1$ and A is a proportionality constant. The rank R referred to above is the global rank of a page of PageRank p, i.e., its position in the list containing all pages of the Web in decreasing order of PageRank. More interesting for information providers and search engine marketers is the rank within hit lists returned for actual queries, where only a limited number of result pages appear. We need a criterion to pass from the global rank R to the rank r within a query's hit list. A page with global rank R could appear at any position $r = 1, 2, \ldots, n$ in a list with n hits. In our framework pages differ only by their PageRank values (or, equivalently, by their in-degrees), as we neglect lexical and other features. Therefore we can assume that each Web page has the same probability to appear in a hit list. This is a strong assumption, but even if it may fail to describe what happens at the level of an individual query, it is a fair approximation when one considers a large number of queries. Under this hypothesis the probability distribution of the possible positions is a Poissonian, and the expected local rank r of a page with global rank R is given by the mean value:

$$r = R\frac{n}{N}. \tag{14}$$

Now it is possible to test the applicability of Eq. 10 to the Web. We are able to estimate the rank of a Web page within a hit list if we know the number of in-links k_{in} of the page and the number n of hits in the list. The procedure consists of three simple steps:

1. from k_{in} we calculate the PageRank p of the page according to Eq. 10;
2. from p we determine the global rank R according to Eq. 13;
3. from R and n we derive the local rank r according to Eq. 14.

The combination of the three steps leads to the following expression of the local rank r as a function of k_{in} and n:

$$r = \frac{A\, n}{(\frac{q}{N} + \frac{1-q}{N\langle k_{in}\rangle} k_{in})^{1.1} N}. \tag{15}$$

We remark that A is a simple multiplicative constant, and its value has no effect on the dependence of the local rank r on the variables k_{in} and n. Therefore we decided to consider it as a free parameter, whose value is to be determined by the comparison with empirical data.

For our analysis we used a set of 65,207 actual queries from a September 2001 AltaVista log. We submitted each query to Google, and picked at random one of the pages of the corresponding hit list. For each selected page, we stored its actual rank r_{emp} within the hit list, as well as its number k_{in} of in-links,

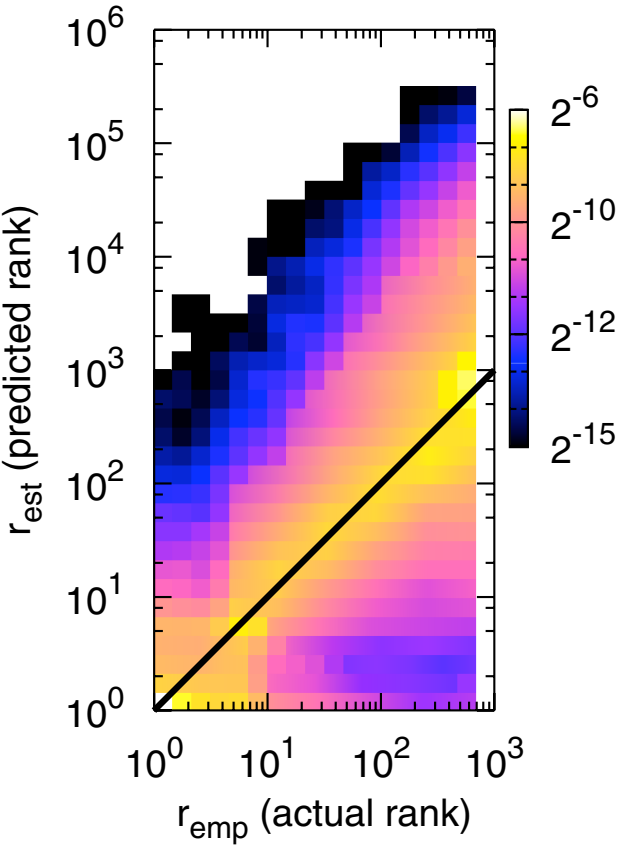

Fig. 4. Density map of the scatter plot between predicted rank r_{est} and actual rank r_{emp} for 65,207 queries. The fraction of points in each log-size bin is expressed by the color, also on a logarithmic scale. The diagonal guide to the eye is $r_{est} = r_{emp}$.

which was again determined through Google.[2] The number n of hits of the list was also stored. Google (like other search engines) never displays more than 1000 results per query, so we always have $r_{emp} \leq 1000$. From k_{in} and n we estimated the theoretical rank r_{est} by means of Eq. 15, and compared it with its empirical counterpart r_{emp}. The comparison can be seen in the scatter plot of Fig. 4. Given the large number of queries and the broad range of rank values, we visualize the density of points in logarithmic bins. The region with highest density is a stripe centered on the diagonal line $r_{est} = r_{emp}$ by a suitable choice of A ($A = 1.5 \times 10^{-4}$). We conclude that the rank derived through Eq. 15 is in

[2] The in-degree data provided by search engines is only an estimate of the true number. First, a search engine can only know of links from pages that it has crawled and indexed. Second, for performance reasons, the algorithms counting inlinks use various unpublished approximations based on sampling.

most cases close to the empirical one. We stress that this result is not trivial, because (i) Web pages are not ranked exclusively according to PageRank; (ii) we are neglecting PageRank fluctuations; and (iii) all pages do not have the same probability of being relevant with respect to a query.

5 Discussion

In this paper we have quantitatively explored two key assumptions around the current search status quo, namely that PageRank is very different from in-degree due to its global nature and that PageRank cannot be easily guessed or approximated without global knowledge of the Web graph. We have shown that due to the weak degree-degree correlations in the Web link graph, PageRank is strongly correlated with in-degree and thus the two measures provide very similar information, especially for the most popular pages. Further, we have introduced a general mean field approximation of PageRank that, in the specific case of the Web, allows to estimate PageRank from only local knowledge of in-degree. We have further quantified the fluctuations of this approximation, gauging the reliability of the estimate. Finally we have validated the approach with a simple procedure that predicts how actual Web pages are ranked by Google in response to actual queries, using only knowledge about in-degree and the number of query results.

Acknowledgments

We thank A. Vespignani and M. Serrano, for helpful discussions. We are grateful to Google for extensive use of its Web API, to the WebBase and WebGraph projects for their crawl data, and to AltaVista for use of their query logs. This work is funded in part by a Volkswagen Foundation grant to SF, by the Spanish government's DGES grant FIS2004-05923-CO2-02 and Generalitat de Catalunya grant SGR00889 to MB, by NSF Career award 0348940 to FM, and by the Indiana University School of Informatics.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks 30(1–7), 107–117 (1998)
2. Sullivan, D.: Nielsen//netratings search engine ratings (August (2005), `http://searchenginewatch.com/reports/article.php/2156451`
3. Amento, B., Terveen, L., Hill, W.: Does "authority" mean quality? Predicting expert quality ratings of Web documents. In: Proc. 23rd ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 296–303 (2000)
4. Pandurangan, G., Raghavan, P., Upfal, E.: Using pagerank to characterize Web structure. In: H. Ibarra, O., Zhang, L. (eds.) COCOON 2002. LNCS, vol. 2387, pp. 330–339. Springer, Heidelberg (2002)

5. Donato, D., Laura, L., Leonardi, S., Millozzi, S.: Large scale properties of the webgraph. European Physical Journal B 38, 239–243 (2004)
6. Garcia-Molina, H.: The Stanford WebBase Project (2005), http://www-diglib.stanford.edu/~testbed/doc2/WebBase/
7. Nakamura, I.: Large scale properties of the webgraph. Physical Review 68 (2003) 045104
8. Volkovich, Y., Litvak, N., Donato, D.: Determining factors behind the PageRank log-log plot. Technical Report 1823, Department of Applied Mathematics, University of Twente (2007)
9. Binney, J., Dowrick, N., Fisher, A., Newman, M.: The theory of critical phenomena. First edn. Oxford University Press, Oxford (1992)
10. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet. Cambridge University Press, Cambridge, UK (2004)
11. Laboratory for Web Algorithmics (LAW), University of Milan: WebGraph (2005), http://webgraph.dsi.unimi.it
12. Donato, D., Leonardi, S., Tsaparas, P.: Stability and similarity of link analysis ranking algorithms. In: Caires, L., Italiano, G.F., Monteiro, L., Palamidessi, C., Yung, M. (eds.) ICALP 2005. LNCS, vol. 3580, pp. 717–729. Springer, Heidelberg (2005)
13. Serrano, M., Maguitman, A., Boguñá, M., Fortunato, S., Vespignani, A.: Decoding the structure of the WWW: Facts versus bias. In: ACM Transactions on the Web (In press)
14. Websidestory: User navigation behavior to effect link popularity (May, Cited by Search Engine Round Table According to this source, Websidestory Vice President Jay McCarthy announced at the Search Engine Strategies Conference (Toronto 2005) that the number of page referrals from search engines has surpassed those from other pages (2005), http://www.seroundtable.com/archives/001901.html
15. Qiu, F., Liu, Z., Cho, J.: Analysis of user web traffic with a focus on search activities. In: Proc. International Workshop on the Web and Databases (WebDB). (2005)
16. Sullivan, D.: Intro to search engine optimization, http://searchenginewatch.com/webmasters/article.php/2167921